TEL-AVIV UNIVERSITY
RAYMOND AND BEVERLY SACKLER
FACULTY OF EXACT SCIENCES

Multiple-Ancestor Localization for Recently Admixed Individuals

Thesis submitted in partial fulfillment of the requirements for the
M. Sc. degree
at the Blavatnik School of Computer Sciences, Tel-Aviv University
by

Yaron Margalit

The research work for this thesis has been carried out at Tel-Aviv
University under the supervision of

Prof. Eran Halperin

October 2014

Abstract

Inference of ancestry from genetic data is a fundamental problem in computational genetics that has wide applications in human genetics, population genetics, and ecology. The treatment of ancestry as a continuum instead of a dichotomous trait has been recently advocated in the literature. Particularly, it has been shown that a European individual's ancestral origin can be determined up to a few hundred kilometers of error using spatial ancestry inference methods such as Principal Component Analysis.

Current methods for the inference of spatial ancestry focus on individuals for which all ancestors originated from the same location. In this work we develop a spatial ancestry inference method that aims at the inference of the ancestral locations of admixed individuals, *i.e.* individuals for which admixture occurred multiple generations back in time.

Our approach is based on a two-layered HMM, modeling both short-range correlations due to linkage disequilibrium, and long-range correlations due to recombination events. We evaluate the method on both simulated and real data from Europe and on a Jewish population from across the world, and demonstrate that it achieves accurate results in up to two generations of admixture.

Acknowledgements

I would like to thank my thesis advisor, Prof. Eran Halperin for helping me focus on the right problems, providing me with ideas and helpful insights when I needed and his guidance throughout my work.

My thanks to the current and past colleagues at Eran Halperin's lab, for always willing to answer any question, work-related or not.

Also, I would like to thank to the love of my life and best friend Reut for making me laugh every day and being there for me whenever I needed.

Contents

Chapter 1

Introduction

In this chapter, we will outline the basis for this research: In section 1.1 we provide a briefly related background on population genetics in general and ancestry inference in particular. In sections 1.2, 1.3 we present the ancestry inference problem and the recent tools that were developed in that domain. Finally, in section 1.4 we present an overview of the multiple-ancestor spatial localization methods that we developed and provide a brief outline of the other chapters in this thesis.

1.1 Population genetics

Each one of us is truly exceptional: the imprint of genetic ancestry and population structure carried in our genome, is unique. The patterns of human genetic variations around the globe may arise by forming local communities (*populations*). These populations are separated by geographic, cultural and social barriers which cause over time genetic variations between populations. These genetic variations are influenced by the evolutionary processes: natural selection and generic drift ([1, 2]). Understanding genome-wide patterns of population structure, evolutionary history and migration patterns have been a major focus in the study of human population genetics.

A common type of genetic variation between genomes occurs at single nucleotide locations and called single-nucleotide variations (SNVs). Multiple sequence alternatives (alleles) exist at each SNV, but observing more than two is rare. In this

work, we will assume only two alleles exist at each SNV. The more common alternative is usually referred to as the *major allele* and the less common alternative is usually referred to as *minor allele*. SNVs whose minor allele occurs relatively commonly in the population (usually defined as in at least 1% of the population), are known as *single-nucleotide polymorphisms* (SNP). A *haplotype* (haploid genotype) consists of a set of associated SNPs on a single chromosome.

The minor *allele frequency* is highly variable across human subpopulations, gender and disease status. For example, some disease causing variants are known to differ substantially in frequency across populations such as Autoimmune diseases [3], Lactase Persistence [4] and Parkinson's disease [5]. For these reasons, SNPs are invaluable and prominent in the field of genetics for the last decade; they present us with significant amounts of information and new opportunities for studies of population structure, diseases [6].

Studying populations can be done by using the allele frequencies: it is possible to make geographic maps representing these proportions for a particular SNP. It is well known that the allele frequency varies considerably from place to place, but usually there is little difference between neighboring populations and high variation is observed at large distances.

High throughput genotyping and sequencing technologies increase the availability of large quantities of human genomic data. Genomic data include now up to millions of SNPs, spanning the whole genome. The availability of increasingly larger population genomic datasets and dense SNP maps provide exciting opportunities for improved understanding of genetic diversity. The dramatic acceleration in availability of genomic data also brings new challenges. It is becoming clear that the bottleneck in population genetics analysis is no longer the collection of data, but instead, the availability of efficient, highly accurate and unbiased population genetics methods that will make full use of the increasingly larger population genetics datasets.

Sequence and SNP data generally take the form of unphased genotypes, which does not directly observe which of the haplotypes, the two parental chromosomes, a particular allele falls in. The related problem of haplotype reconstruction from unphased genotypes called *haplotype phasing*. Computational methods were developed to infer the phasing of whole genotype sequence by considering sets of common haplotypes that can explain the observed genotypes data [7]. When

haplotypes are inherited from a common ancestor, they are *identical-by-descent* (IBD). The inference of IBD consists of detecting genomic regions in which a pair of individuals share a haplotype descended from a recent common ancestor. IBD is informative for phase estimation. Also an accurate IBD map provides important insights for the reconstruction of family relations which can use to discover association of genetics and diseases [8].

In population genetics, *ancestry inference* is required to infer an individual's population of origin. The use of SNPs can facilitate an accurate and reliable ancestry inference by exploiting allele frequency differences across populations. Ancestry inference is a standard part of genetic analysis in a wide range of fields: it is critical for disease studies, in which it is a major confounder that needs to be accurately accounted for [9], as well as a factor that can be cleverly used to increase power in certain scenarios such as disease mapping in admixed populations [10, 11]. It is often needed to infer an individual's population of origin even in the case of self-reported ancestry. For example, it provides a potential clinical utility, as the fraction of the genome traced to native American ancestry which strongly correlate with relapse of acute lymphoblastic leukaemia, even when children identify themselves as Caucasians [12]. In addition, ancestry inference has multiple applications for population genetics studies [13–17].

An *admixed individual* is one whose genome is a mosaic of genomes from multiple origins mixing for several generations. Due to *recombination events* each chromosome of an admixed individual is a combination of chromosomal regions originating from different ancestral populations. In recent admixed individuals (*e.g.* African Americans) where mixture event occurred for relatively a small number of generations, only a few recombination events occurred in the ancestral chromosomes since the mixing event resulting in long-range genomic that may share the same origin.

One of the challenges in the field of ancestry inference arises from the *short-range* and the *long-range* correlations among SNPs. These correlations exist due to *linkage disequilibrium* (LD), non-random (correlation) between genetic variants. A range of demographic, molecular and evolutionary processes has a significant effect on the patterns of LD. Genetic drift can cause big losses of genetic variation for small populations which tend to increase LD, as haplotypes are lost from the population. Recombination might have the most evident impact on LD [18]. Recombination events may break part of these associations and decrease them

proportionally with the increase in the number of generations. Recombination rates are known to vary across the genome, hence, the extent of LD is expected to vary in an inverse relation to the local recombination rate. Also there are other contributors to the extent and distribution of disequilibrium such as admixture, migration (gene flow), natural selection, population bottleneck, and population growth may also influence LD [18, 19].

LD-correlations may be a source of biased results and can effectively avoid spurious conclusions on the power of statistical tests [20]. Understanding the population structure and linkage disequilibrium may improve the accuracy in association mapping [21]. Intuitively, some LD correlations may be reduced by eliminating linked SNPs in the data but this process could result in the loss of relevant SNPs and more importantly, it does not exploit the potentially of a dense genome-wide data.

1.2 The ancestry inference problem

The ability to identify the geographic origin of an individual based on his DNA sequence is one of the basic building blocks of population genetics. The problem of ancestry inference is commonly viewed at one of two levels:

1. at the global scale, inferring an individual's single origin out of several possible homogeneous ancestries, or determining an individual's ancestral genomic composition in the case of admixed individual.

2. at the local scale, labeling the different ancestries along the chromosomes of an admixed individual

Ancestry inference methods can be generally divided into *dichotomous* and *spatial* methods. In dichotomous methods, the world is divided into populations typically based on continents or countries, and the goal is to predict the percentage of the genome of an individual originating from each of the populations (e.g., 50% African and 50% European). In contrast, in spatial analysis, each individual is assigned to a point on a 2-dimensional map, corresponding to its inferred ancestry.

In dichotomous approaches, the arbitrary partition of the world to populations could bias the results and reveal only a crude low-resolution picture of the population structure. Another drawback is the number of populations that must be

provided to the algorithms. Heuristic approaches may be introduced to select the number of populations, for example, based on comparing the penalized likelihood of the different runs. Still, it may be the case that the penalized value will be high for the case of additional unrelated population or the removal of population. Also, typically more than 10 populations may cause difficulties in running time and the presence of distinct local optimum classification [22, 23].

Spatial analysis provides a considerable opportunity for a leap in our ability to predict an individual's ancestry. The high-resolution ancestry inference provided by spatial analysis will substantially increase the resolution of population genetics analysis. It can be further leveraged to study spatial genetic variation, i.e., allele distribution as a function of space and time, and as a function of evolutionary forces such as selection, recombination, and mutations.

1.3 Related Work

1.3.1 Dichotomous ancestry inference methods

Several tools were suggested to treat ancestry as a discrete entity, *i.e.* to classify individuals to their ancestry, or to multiple ancestries in the case of individuals of admixed ancestry [24–30].

Methods such as STRUCTURE [24] and ADMIXTURE [25] provide the global ancestry inference by clustering individuals into their appropriate populations, on the basis of observed genotypes at multiple SNPs. STRUCTURE is based on a Bayesian approach and ADMIXTURE uses a maximum likelihood approach.

In the case of local ancestry inference, several methods are based on *Hidden Markov models* (HMM), where the hidden states correspond to ancestral populations and generate the observed genotypes [28]. Other methods explored window-based approaches in which at most a single admixture event exist within a window; a popular such method is LAMP [26]. Later extensions such as LAMP-LD [30] were created to explicit modelling LD within populations. LAMP-LD improved accuracy, for example, in Latinos Mexican dataset the accuracy results of 10% diploid error (the percentage of incorrect ancestry error of all SNPs) [30].

1.3.2 Spatial ancestry inference by dimensionality reduction techniques

At the same time, non-parametric methods, typically dimensionality reduction techniques such as PCA (Principal Component Analysis)[31] or MDS (multidimensional scaling)[32] were heuristically used to visualize the variation in the genetic samples on a continuous space. PCA can effectively extract the fundamental structure of a dataset without any need for modelling of the data. While PCA is a common generic method that does not directly model the generic samples, in some instances, particularly in Europe, PCA maps fit well to the geographic map [33, 34]: presumably because the isolation-by-distance assumption holds relatively well. In such cases PCA maps can be used for localizing individuals in the geographic space based on their genetic information.

1.3.3 Spatial ancestry inference by explicit probabilistic models

During the last couple of years, a few spatial localization methods which explicitly model the relation between geography and genetics, have been suggested. These methods describe the allele frequency of each SNP using a continuous function of the geographic coordinates, thereby specifying a link between location and sequence.

SPA (Spatial Ancestry Analysis, [35]), defines a probabilistic model in which the allele frequency of each single nucleotide polymorphism (SNP) is modeled as a function of (x, y) coordinates of an individual on a map. In this framework, we assign a slope function for each SNP, defining a model by which the probability of observing a specific allele depends on the geographic location of an individual. More formally, for each SNP j, we define a slope function:

$$f_j(x, y) = \text{logit}^{-1}(a_j x + b_j y + c_j) \tag{1.1}$$

where $\text{logit}^{-1}(z) = \frac{1}{1+e^{-z}}$. The model assumes that the probability of observing the minor allele at position (x,y) is given by $f_j(x, y)$. With the use of a logistic function, we are able to capture some of the genetic variations, but not all of them.

An example of variation that cannot be captured is when there are multiple peaks in the allele frequency surface.

SPA uses a maximum likelihood approach to estimate simultaneously the slope function for each SNP and the spatial positioning of each individual. In that case, SPA can functional as an unsupervised method. Assuming that the individuals are independently sample from the population, the observed genotype samples can be calculated from the log-likelihood:

$$L(G; X, A, B) \propto \sum_i \sum_j g_{i,j} \ln f_{i,j} + (2 - g_{i,j}) \ln(1 - f_{i,j}) \tag{1.2}$$

where $g_{i,j}$ is the observed number of minor alleles (0-2) at SNP j of individual i and $f_{i,j}$ is the slope function f_j with position individual i (x_i, y_i). Each row in A, B contain the coefficients (a_j, b_j) of the allele frequency logistic function.

This model provides a probabilistic interpretation of spatial analysis, allowing more flexibility in spatial analysis. Particularly, it was shown that SPA can be used for the detection of selection signature, and that it provides better accuracy than principal component analysis in geographic localization [35].

PCA and SPA assume that each SNP is an independent evidence of a sample's origin. Linkage disequilibrium (LD), non-random correlations among close SNPs, invalidates the independency assumption. For example, a pair of two perfectly linked SNPs should only be counted as one. Such unaccounted correlations could bias the results [36]. LOCO-LD [37] (Localization corrected for LD) is a developed method that leverages this information and provides an improved ability to localize samples. LOCO-LD is a window-based probabilistic modeling of linkage disequilibrium where each short-window describes the allele frequencies and the linkage patterns by multivariate normal distribution (we detailed the haplotype-version of LOCO in section 2.1). By modelling LD, LOCO-LD achieved better results than SPA, a localization error of 206 km in European data compared to 226 km in SPA [37].

1.3.4 Spatial ancestry inference for recent admixed individuals

All of the tools listed for spatial ancestry inference are designed only for a local ancestry of homogeneous ancestry, whereas for the admixed case, individuals may come from two or more different ancestries. Specifically, if we go back k generations, we may expect that each of the 2^k ancestors of an individual can originate from a different (x,y) location. To account for mixed ancestry, we will treat the genome of each individual as a mixture of genomes, centered at points $(x_1, y_1), \ldots, (x_{2^k}, y_{2^k})$: our objective will be to infer the location of these centers.

In the last few years, many studies focused not only on homogenous populations, but also on admixed populations (e.g., Latino Americans, African-Americans and Hispanics-Americans). The development of accurate ancestry inference for admixed individuals is required in a wide range of applications [38–40]. While several methods for the analysis of admixed individuals were suggested in the dichotomous ancestry inference field and are now being used in recent studies [41–43], not much progress was achieved in the spatial ancestry inference field. High-resolution maps and the use of a large number of populations (moving from continental to subcontinental) in studies are required. Moreover, in the last few years, many datasets include recent admixed individuals. For example, in the POPRES dataset we observe that out of 1906 European individuals, 470 are recent admixed (up to 2 generations) [44, 45]. The increase in immigration rates [46] and the rapid and easy way people can move from one country to another, as a result of globalization, suggest that more recent admixed individuals will be included in new datasets.

Dimension reduction methods like PCA would localize an admixed individual in a single location in between the mixing locations [47]. Furthermore, it is not clear how three-way or multi-way mixture would infer in the PC-plot pattern, and how the admixture proportions can be inferred.

The advantage of explicit model methods is that their models can be naturally extended to describe individuals of admixed origin, thus allowing for inference of multiple origins per individual. SPA [35] provides limited progress in this direction; SPA's accuracy on mixed individuals is severely limited and it assumes that the mother and father are not of mixed ancestry.

Current spatial analysis methods only address limited situations and low quality results in which an individual has a mixed ancestry. The goal of this thesis is to improve the current results of spatial ancestry inference for recent admixed individuals. We show that our new method provides substantial improvement for accuracy of up to 2 generations admixture.

1.4 Overview and Outline

In this thesis, we provide a method called SLAM (Spatial Localization of AdMixed individuals), for estimating the multiple geographic origins of individuals of recent admixed ancestry. Our approach is based on modeling the long- and short-range correlations that exist in the genetic sequence of admixed individuals. The long-range correlations result from the recent mixing: As only a few recombination events occurred in the ancestral chromosomes since the mixing event, a long-range of SNPs will share the same origin.

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), similarly to the approach in [37]. 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.

In chapter 2, we begin by introducing a model for individuals of 1-generation admixture. Next, in chapter 3, we extend this model to localize individuals of g-generations admixture (for instance, a case of two generations with four different ancestries). In chapters 4, 5, we evaluate SLAM on simulated data. In the simulations we use individuals of known geographic origins to generate the DNA sequence of individuals who exhibit different mixing patterns. SLAM attains high accuracy in localizing the two parental locations of individuals whose father and mother originate from different geographic regions, with a median error of 374 km.

In chapter 6, we use individuals of admixed ancestry from Europe whose multiple origins are known to validate the method on real data, and observe that SLAM's localization accuracy remains high and similar to the simulation results. Finally, in chapter 7 we use individuals from a dataset of Jewish individuals sampled across

Europe, the Mediterranean and Africa. This dataset present a challenge to ancestry inference methods due to the small sample sizes of the Jewish populations, as well as the small effective population size of the Jewish people [48]. Small population size results in long regions of identity by descent (IBD) that may interfere with ancestry inference. Interestingly, we find that SLAM's accuracy, when applied to the Jewish individuals, is comparable to the accuracy when applied to the general European population.

Chapter 2

Spatial model for individuals of 1-generation admixture

We introduce a model for individuals of 1-generation admixture, *i.e.* whose parents came from different locations but are themselves of homogeneous ancestry.

We define an individual to be of *homogeneous* ancestry if all of their ancestors originated from the same geographic location. 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.

We assume that phased haplotype reference panel is available. This panel includes $2n$ haplotypes, $H = (h_1, ..., h_{2n})$ over a common set of SNPs, each pair belonging to an individual of homogeneous ancestry. In addition, we assume that the corresponding locations of origins $X = (x_1, ..., x_{2n})$ are known. We note that each location is a vector which contains longitude and latitude coordinates, and that $x_{2i-1} = x_{2i}\ \forall i \in \{1 \dots n\}$ since the individuals are homogeneous. Given new individuals of recent admixed ancestry, our goal is to predict their geographical origin by utilizing the information in the reference panel.

We are given haplotypes h_1, h_2 whose origins x_1, x_2 are unknown. If phasing was error-free, our problem would be reduced to finding location per haplotype. To do that, we can use known methods that infer the location of haplotypes with homogeneous ancestry [35, 37]. Unfortunately, perfect phasing is typically not feasible, and our model therefore needs to account for phasing errors. Specifically,

each haplotype can be viewed as a mosaic of segments from paternal and maternal DNA which result from phasing errors.

2.1 Estimation of window-specific spatial parameters from training data

We split the haplotypes into L non-overlapping contiguous windows of l SNPs per window, and simplify our model by assuming that there are no phasing errors within a window. This assumption will be violated in some windows, and the inference in such windows will be less accurate. However, if we choose l to be small enough, in most windows the assumption will hold.

Denote by h_{ij} the part of haplotype h_i confined to window j. Similarly to [37], our model assumes that 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. The score function that we obtain for haplotype h in window j is therefore

$$L(h_{ij} \mid \beta_j, \Sigma_j, x_i) = MVN(h_{ij}; \beta_j x_i, \Sigma_j) = \frac{1}{(2\pi)^{\frac{l}{2}} \mid \Sigma_j \mid^{\frac{1}{2}}} e^{-\frac{1}{2}(\beta_j x_i - h_{ij})^T \Sigma_j^{-1} (\beta_j x_i - h_{ij})} \tag{2.1}$$

Note that β_j is a $l \times d$ matrix and Σ_j is a $l \times l$ matrix, where d is the dimension of the spatial representation.

In the training stage we estimate the parameters of the multivariate normal distribution from the reference haplotypes. The procedure is analogous to the one described in [37], but using haplotypes instead of genotypes. Specifically, denote by H_j the matrix whose is column is haplotype h_{ij}, and by X the matrix whose is column is the corresponding location vector x_i. Then the maximum likelihood estimator of β_j has the following closed form solution. We wish to optimize (2.1), an equivalent expression can be used:

$$f(\beta_j) = \sum_{i=1}^{2n} \big(\beta_j x_i - h_{i,j}\big)^T \Lambda_j \big(\beta_j x_i - h_{i,j}\big)$$

$$\frac{\partial f}{\partial \beta_j} = 2\Lambda_j \beta_j \Big(\sum_{i=1}^{2n} x_i x_i^T\Big) - 2\Lambda_j \Big(\sum_{i=1}^{2n} h_{i,j} x_i^T\Big)$$

We set the derivative equal to 0 and obtain:

$$\begin{aligned} &\Rightarrow XX^T\beta_j^T = XH_j^T \\ &\beta_j = H_jX^T(XX^T)^{-1} \end{aligned} \tag{2.2}$$

Next, given β_j, the maximum likelihood estimator of Σ_j can also be directly derived:

$$\Sigma_j = \frac{1}{2n}\sum_{i=1}^{2n}\big(\beta_j x_i - h_{i,j}\big)\big(\beta_j x_i - h_{i,j}\big)^T \tag{2.3}$$

We follow the standard procedure of adding a small positive constant λ to the diagonal of Σ_j in order to ensure it is full rank and elilminate potential overfitting of the covariance parameters.

2.2 Correlation between positions in consecutive windows

Once the window-specific parameters are estimated, we combine the windows into an HMM which captures the correlation between positions in consecutive windows. Unless a phasing errors has occurred, the locations of the two haplotypes in two consecutive windows will remain unchanged, if a phasing error has occurred, the locations will be switched between the haplotypes.

The HMM is specified by a triplet (Q, ϵ, δ), where Q is the set of states, δ are the transition probabilities and ϵ is the emission probability functions. The set Q contains 2 x L states: For each one of the L windows there are two states, denoted (s_j^1, s_j^2), for the two possible phasing arrangements in the two window haplotypes (h_j^1, h_j^2). We define $z_j = (z_j^1, z_j^2)$ to be the indicators of the two possible phasing arrangements. Given the phasing arrangement, the states of window j emit the haplotypes in the window in probability that matches the MVN density specified in Equation (2.1) and estimated in Equations (2.1) and (2.3):

$$\begin{aligned} \epsilon_{s_j^1}(h_j^1, h_j^2, x_1, x_2) &= MVN(h_j^1; \beta_j x_1, \Sigma_j) \cdot MVN(h_j^2; \beta_j x_2, \Sigma_j) \\ \epsilon_{s_j^2}(h_j^1, h_j^2, x_1, x_2) &= MVN(h_j^2; \beta_j x_1, \Sigma_j) \cdot MVN(h_j^1; \beta_j x_2, \Sigma_j) \end{aligned} \tag{2.4}$$

The between-states transition rates are constant across windows and determined by p_s, the probability of a switch error between windows. The exact choice of this parameter, which we denote p_s, is discussed in the Results chapter 5. We therefore have:

$$\delta(s_{(j-1)}^{k1} \to s_j^{k2}) = \begin{cases} 1 - p_s & \text{if } k1{=}k2. \\ p_s & \text{otherwise.} \end{cases} \tag{2.5}$$

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, we use the Forward-Backward algorithm to compute (z_j^1, z_j^2) the expectations of the states (s_j^1, s_j^2), for every window j. We note that $z_j^1 + z_j^2 = 1$. In the maximization step we maximize the following function:

$$\begin{aligned} L(h^1, h^2 \mid \beta, \Sigma, z^1, z^2, x_1, x_2) = \sum_{j=1}^{L} \big(& z_j^1 \log(MVN(h_j^1; x_1) MVN(h_j^2; x_2)) \\ & + z_j^2 \log(MVN(h_j^1; x_2) MVN(h_j^2; x_1)) \big) \end{aligned} \tag{2.6}$$

The values of x_1, x_2 which maximize expression 2.6 have the following closed form solution. We try to maximize the likelihood expression as the function of x_1, x_2:

$$\begin{aligned} f(x_1) &= \sum_{j=1}^{L} \big(z_j^1 (h_j^1 - \beta_j x_1)^T \Lambda_j (h_j^1 - \beta_j x_1) + z_j^2 (h_j^2 - \beta_j x_1)^T \Lambda_j (h_j^2 - \beta_j x_1) \big) \\ &= \sum_{j=1}^{L} \big(-2(z_j^1 h_j^{1T} + z_j^2 h_j^{2T}) \Lambda_j \beta_j x_1 + (\beta_j x_1)^T \Lambda_j (\beta_j x_1) \big) \end{aligned}$$

$$\frac{\partial f}{\partial x_1} = \sum_{j=1}^{L} \big(-2(z_j^1 h_j^{1T} + z_j^2 h_j^{2T}) \Lambda_j \beta_j + 2 x_1^T \beta_j^T \Lambda_j \beta_j \big)$$

We set the derivative equal to 0 and obtain:

$$\Rightarrow x_1^T = \big(\sum_{j=1}^{L} (z_j^1 h_j^{1T} + z_j^2 h_j^{2T}) \Lambda_j \beta_j \big) \big(\sum_{j=1}^{L} (\beta_j^T \Lambda_j \beta_j) \big)^{-1} \tag{2.7}$$

Same stages apply to x_2 as well with the corresponding equations:

$$x_2^T = \big(\sum_{j=1}^{L}(z_j^2 h_j^{1T} + z_j^1 h_j^{2T})\Lambda_j\beta_j\big)\big(\sum_{j=1}^{L}(\beta_j^T\Lambda_j\beta_j)\big)^{-1} \tag{2.8}$$

We repeat the expectation-maximization iterations until convergence. The stopping criteria was set to $\epsilon = 10^{-8}$ and the maximum number of iterations to 100.

Chapter 3

Spatial model for individuals of g-generation admixture

We extend the model to describe individuals of g-generations admixture (for instance, a case of two generations with four different ancestries).

For such individuals each of the chromosomes (paternal and maternal) is drawn from a different (but potentially overlapping) set of up to 2^{g-1} ancestries and corresponding 2^{g-1} paternal and maternal locations $(x_1^1, \ldots, x_{2^{g-1}}^1)$, $(x_1^2, \ldots, x_{2^{g-1}}^2)$. Because the admixture is recent (*i.e.*, g is small), the pair of locations from which the haplotypes in window j are drawn are correlated with the locations of window $j+1$. In order to capture these correlations, we group each K consecutive windows into a block, and assume no recombination within blocks. We do allow for phasing errors within blocks, but maintain the assumption from the section 2.2 of no phasing errors within windows.

In order to capture the structure of windows within blocks we define a two-level HMM:

1. Per each block we construct $2^{2(g-1)}$ bottom-level HMMs, one for each unordered pair of parental locations. Each of these HMMs contains $K \times 2$ states, thereby capturing all possible phasing arrangements in the block. 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 arrangements of paternal and maternal locations along the genome.

Our goal is to estimate the 2^g locations of origin, 2^{g-1} per parent, and we do so again using the Baum-Welch algorithm. The Expectation stage consists of the following two 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 gives us the emission probabilities of this block in the top-level HMM. The computation also gives us $(z^1_{br1}, z^2_{br1}) \dots (z^1_{brK}, z^2_{brK})$, the expectations of the per-window state variables in block b and for pair of locations r. We note that $z^1_{brj} + z^2_{brj} = 1$.

2. We run the Forward-Backward algorithm on the top level HMM. The transitions are assumed fixed for blocks and are determined by the average genome-wide recombination rate and by g. For each (b, r) this computation gives us the $(w_{b1}, \dots, w_{b2^{2(g-1)}})$, the expectations of the per-block state variables along the genome.

In the maximization stage we derive the locations that maximize the expected log likelihood:

$$\begin{aligned} & L(h^1, h^2 \mid \beta, \Sigma, z^1, z^2, x^1, x^2, w) \\ =& \sum_{b=1}^{B} \sum_{r=1}^{2^{2(g-1)}} \sum_{j=1}^{K} \big(w_{br} \big(z^1_{brj} \log(MVN(h^1_j; x^1_r) MVN(h^2_j; x^2_r)) \\ & + \; z^2_{brj} \log(MVN(h^1_j; x^2_r) MVN(h^2_j; x^1_r)) \big) \end{aligned} \tag{3.1}$$

We try to maximize the likelihood expression 3.1 as the function of x_i , $(i = 1 \dots 2^g)$:

$$\begin{aligned} f(x_i) = \big(\sum_{b=1}^{B} \sum_{r=1}^{2^{2(g-1)}} \sum_{j=1}^{K} \big(& w_{br} z^1_j (h^1_j - \beta_j x_i)^T \Lambda_j (h^1_j - \beta_j x_i) \\ & + w_{br} z^2_j (h^2_j - \beta_j x_i)^T \Lambda_j (h^2_j - \beta_j x_i) \big) \big) \end{aligned}$$

$$\frac{\partial f}{\partial x_i} = \sum_{b=1}^{B} \sum_{r=1}^{2^{2(g-1)}} \sum_{j=1}^{L} \big(- 2w_{br}(z_j^1 h_j^{1T} + z_j^2 h_j^{2T})\Lambda_j\beta_j + 2w_{br}x_i^T\beta_j^T\Lambda_j\beta_j\big)$$

We set the derivative equal to 0 and obtain:

$$\Rightarrow x_i^T = \big(\sum_{b=1}^{B} \sum_{r=1}^{2^{2(g-1)}} \sum_{j=1}^{K} \big(w_{br}(z_{brj}^1 h_j^{1T} + z_{brj}^2 h_j^{2T})\big)\Lambda_j\beta_j\big) \big(\sum_{b=1}^{B} \sum_{r=1}^{2^{2(g-1)}} \sum_{j=1}^{K} (w_{br}\beta_j^T\Lambda_j\beta_j)\big)^{-1} \tag{3.2}$$

The above optimization can easily be modified to capture scenarios in which each parent has a different number of ancestral locations (*e.g.*, the father resulted from a 1-generation admixture and the mother is homogeneous).

Chapter 4

Simulation setup

We used a simulated data to investigate the user-defined parameters of our model, in order to assess our algorithm's performance. We tested our methods on European samples, which are part of the POPRES data [45]. The data set consists of 364,373 SNPs that passed the criteria of allele frequency > 0.01 and missingness < 10 %. We used BEAGLE version 4 [7] for phasing and imputation.

For training and simulations, we identified individuals whose four grandparents had the same geographic origins (homogeneous ancestry). A total of 1385 individuals were collected. The dataset was split into two disjoint groups. We used some of the phased samples to simulate admixed samples. The remaining haplotypes were used as training haplotype sets to estimate model parameters.

We simulated European admixed haplotypes as a mosaic of segments randomly taken from haplotypes of homogeneous ancestry. We then assumed that admixture event occurred g generation ago independently for the maternal and paternal haplotypes. We simulated the haplotype of an admixed individual by first sampling the number of recombinations from a Poisson distribution with a parameter determined by g, and then sampling the recombination positions uniformly across the genome. After that, we set a fixed recombination rate of $\theta = 10^{-8}$ per base-pair per generation.

In order to emulate switch errors, we joined pairs of admixed haplotypes to form diplotypes and then used BEAGLE version 4 [7] to produce admixed haplotype samples with switch error events. We used those samples in order to estimate the switch error rate. We then simulated switch errors by planting a switch events in

the rate that is calculated according to BEAGLE switch error rate. We simulated 1500 samples of one generation admixture and added the switch error to them. Similarly, we simulated 500 samples of two generations admixture. Unless noted otherwise, we used those samples in the Simulation results chapter.

4.1 Performance assessment

We used several metrics to assess the performance of SLAM. The assignment's *accuracy* of each admix individual was determined by averaging distances between predicted geographical origins and their closest country of origin of that individual. Also, we define the distance errors of an individual as the distance from the true location of each of its 2^g ancestors to the estimated location. In both metrics, a few permutations may appear to set the predicted locations to the true locations. The picked permutation is chosen by finding the one with the best match.

In order to calculate the distance error, we used a transformation function on the resulting points that aims to minimize the distance between the real geographic locations and the inferred location similar to the transformation performed in [33, 35, 37]. The reference set was split into two disjoint groups where one group was used for the model's parameters and the other was used for the transformation' parameters. We applied the cross-validation method (10-fold), to validate our results on different splits of the reference set.

Chapter 5

Simulation results

5.1 Performance evaluation on simulated European data

We first examined the performance of SLAM on simulated individuals of 1-generation admixture generated from the phased POPRES dataset (see section 4). We compared SLAM to SPA's extension for admixed individuals[35]; this extension was designed to localize only 1-generation admixture, and was shown to attain a high accuracy when applied to such individuals from the POPRES dataset. A third

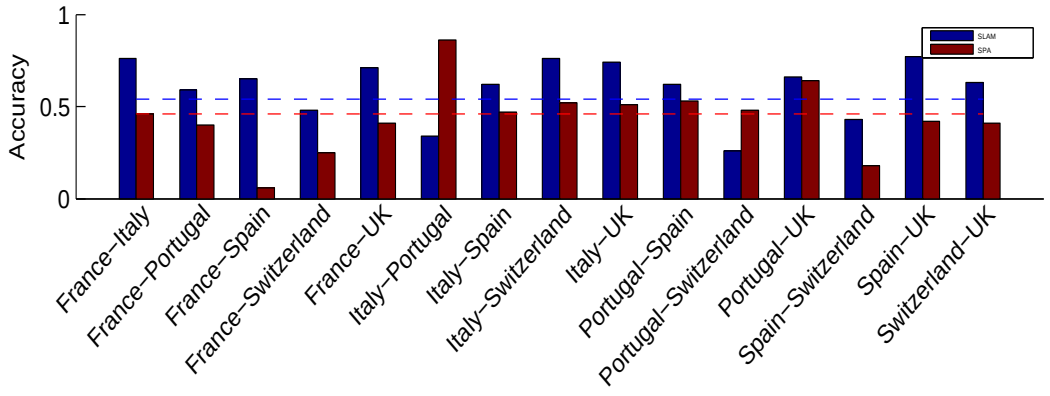

FIGURE 5.1: **Comparison of origin assignment accuracy between SLAM and SPA.** For each pair of locations we gave the fraction of admixed individuals that were correctly classified to both countries of origin. Horizontal bars mark the average accuracy per method, when the different mixture types are given equal weights.

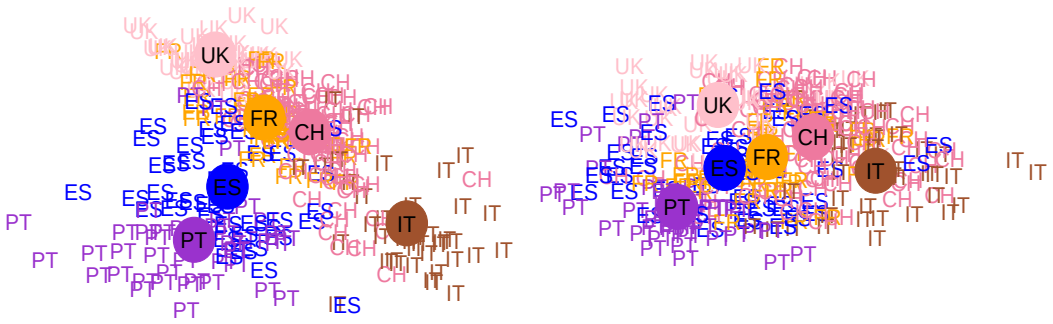

FIGURE 5.2: **SLAM's (Left) and SPA's (Right) localization results for 1-generation admixture in the POPRES data set.** The plots depict the inferred locations and contains a mark per inferred locations, two locations per individual. Abbreviations and colors code the true country of origin. Abbreviations: IT, Italy; PT, Portugal; ES, Spain; FR, France; UK, United-Kingdom; CH, Switzerland

method named SPAMIX [44] that was designed to estimate recent ancestry locations has been recently published, however its code is currently not available and we therefore could not include its results in this report.

For both SPA and SLAM, we first used a training set of homogeneous ancestry individuals with their known origins to estimate model parameters, and then localized the simulated admixed individuals.

Figure 5.1 depicts the accuracy of assignment to country of origin as achieved by both methods. The panel includes results for populations that are represented by at least 40 individuals in the training data. SPA's average success rate is 42%, whereas SLAM's success rate is 57% (see Figure 5.1). In two cases SPA performed better than SLAM (Italy-Portugal and Portugal-Switzerland), and more generally, SLAM obtained considerably lower success rates on some countries. In these cases most of SLAM's predictions were the nearest countries to the origins: For example, Portugal's successful classification rate was only 26%, but 90% of the misclassified origins were assigned to Spain.

SLAM also outperforms SPA in terms of km error, attaining a median error of 374 km SLAM compared with SPA's 1141 km. Figure 5.2 depicts the inferred locations for all individuals in the test set of both methods, with the transformation stage (see section 4.1) omitted in order to show the exact clustering patterns. SLAM's predictions can be seen to cluster better within each country, thereby supporting the claim that modeling short-range correlations improves the localization of individuals of 1-generation admixture.

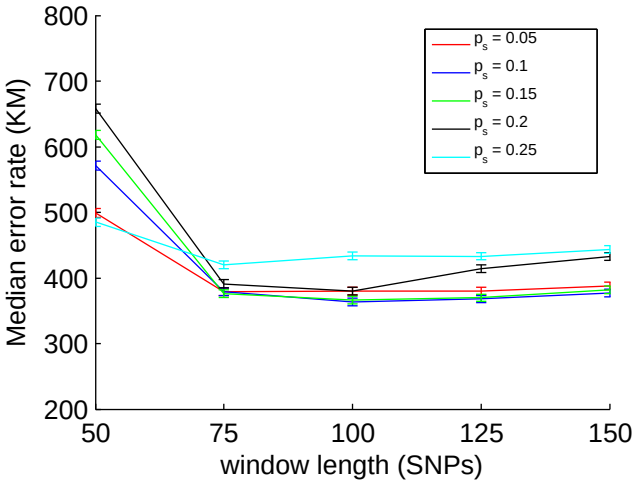

FIGURE 5.3: **The effect of the window's length l and the switch error probability parameter p_s on SLAM's accuracy.**

5.2 Robustness of SLAM to parameter setting

SLAM partitions the genome into short windows and assumes no phasing errors within them. Each window provides LD information, however as the window length increases the probability of violating the perfect phasing assumption increases as well. In addition, the number of covariance parameters increases quadratically with window length, and it becomes harder to accurately estimate them using a limited number of samples. Setting the window length therefore involves a tradeoff, and so we used the simulated POPRES data to test the robustness of SLAM's performance to this parameter's values, as well as to the value of the phasing error probability parameter p_s.

The results depicted in Figure 5.3 reveal that SLAM's performance is robust to the choice of p_s for window lengths in the range of 75 to 125 SNPs, corresponding to 600-900 Kb on average. Fixing the window length to 100, the p_s value which yields the highest accuracy is 0.1. According to the results of this section we used the values $l = 100$, $p_s = 0.1$ in all our experiments unless mentioned otherwise. We note that p_s depends strongly on the phasing quality. Based on our simulations, the phasing error rate of BEAGLE on the POPRES dataset is sufficiently low to assume there are no errors across multiple windows of length (75-125). We also note that the actual error rate as simulated by BEAGLE was measured to be 0.2, twice the length of the optimal value we have estimated; we believe that the reason for the discrepancy is the large variance in phasing error rates across different genomic regions.

TABLE 5.1: **Comparison of the effect of LD using different algorithms**

Comparison between algorithms	Distance: 2^{nd} $[1^{st}, 3^{rd}]$ Quartile (KM)
SLAM	374 [263.0 519.0]
SLAM (perfect phasing)	226.7 [158.4 307.3]
SLAM with no LD (covariance = I)	576.5 [425.2 733.0]
SLAM window length 1	630.5 [573.1 829.3]

5.3 The effect of modeling LD

We quantified the effects of modeling LD on the spatial-localization results of admixed individuals. To eliminate the account for LD, we generated LD-pruned-data. We removed SNPs within a 50 SNP window that had a cutoff of 0.2 for the pairwise r^2 using PLINK [32]. We compared several approaches:

1. Using the SLAM model, but ignoring the presence of LD by updating the covariance matrix values to include only the variances (diagonal values) and using LD-pruned-data.
2. Using LD-pruned-data and setting window length to 1.
3. Using SLAM's model with the window's length set to 100.
4. SLAM's model with window length set to 100 and perfect phasing data.

Options 1., 2. do not model LD, while options 3., 4. do. Moreover, option 4. is an ideal case that provides an upper bound on what we could achieve using our model. Table 5.1 suggests that accounting for LD significantly improves the results. Interestingly, SLAM with no LD (covariance = I) showed better results than setting window length 1. This implies that assuming long stretches of SNPs with no switch error may be more resilient to noise. Also, we note that with perfect phasing, SLAM results are similar to haplotype version of LOCO-LD [37], implying that the phasing error incur a major reduction in accuracy.

5.4 Localization of 2-generations admixed individuals

We simulated individuals of 2-generations admixture using the phased POPRES homogenous individuals (See chapter 4). In the simulations, we fixed the number of generations to two and set the number of different ancestries to four. Similarly to the situation for windows described in 5.2, a tradeoff exists when setting the number of windows within a block. We fixed the block size to be 20 and window length is 100 SNPs, which indicates a recombination event every second block. SLAM succeeded to localize individuals of 2-generation admixture of 4 different ancestries with median distance error of 478 km. This is 27% increase error compared to 1-generation admixture.

5.5 Running time

We recorded the average run time that it takes for SPA and SLAM methods to infer the locations of a new recent admixed individual. We did not include the training stage as part of the experiment, though SLAM (same as LOCO-LD method) is 140 times faster than SPA [37]. The experiment was done on a machine containing eight Quad-Core AMD Opteron 2354 processors. We estimated the running time of inferring the location of individual of 1-generation admixture. As expected, SPA was faster with running time of 20 seconds, while SLAM running time was 45 seconds on average. The number of SNPs that each method handles is different. SPA input includes only 84462 pruned-LD-data, whereas SLAM input uses all 364373 SNPs. In addition, SLAM localized individual of 2-generation admixture with running time of 2 minutes.

Chapter 6

Real data results

So far we have investigated the performance of SLAM on simulation data. We also used SLAM to localize 1-generation admixed individuals from the PORPES dataset for whom locations of origin are known. We analyzed 254 such individuals, and achieved median error of 368 km.

In addition, we questioned how accurate the results of simulation to the real data. To evaluate that, we used the same admixed populations and the same amount of samples in both datasets. Figure 6.1 shows that SLAM's localization results on the real data are closely matched to those observed in the simulation study. For example, SLAM achieved median error of 396 km compared to 399 km for admixed population group Switzerland-France (CH-FR). In simulation the standard error results are higher compared to the real data. This can be explained by our method of cross-validation (10-fold). On each run we use different set of simulated samples that are differ from the training set, whereas, the real data use the same samples in all runs.

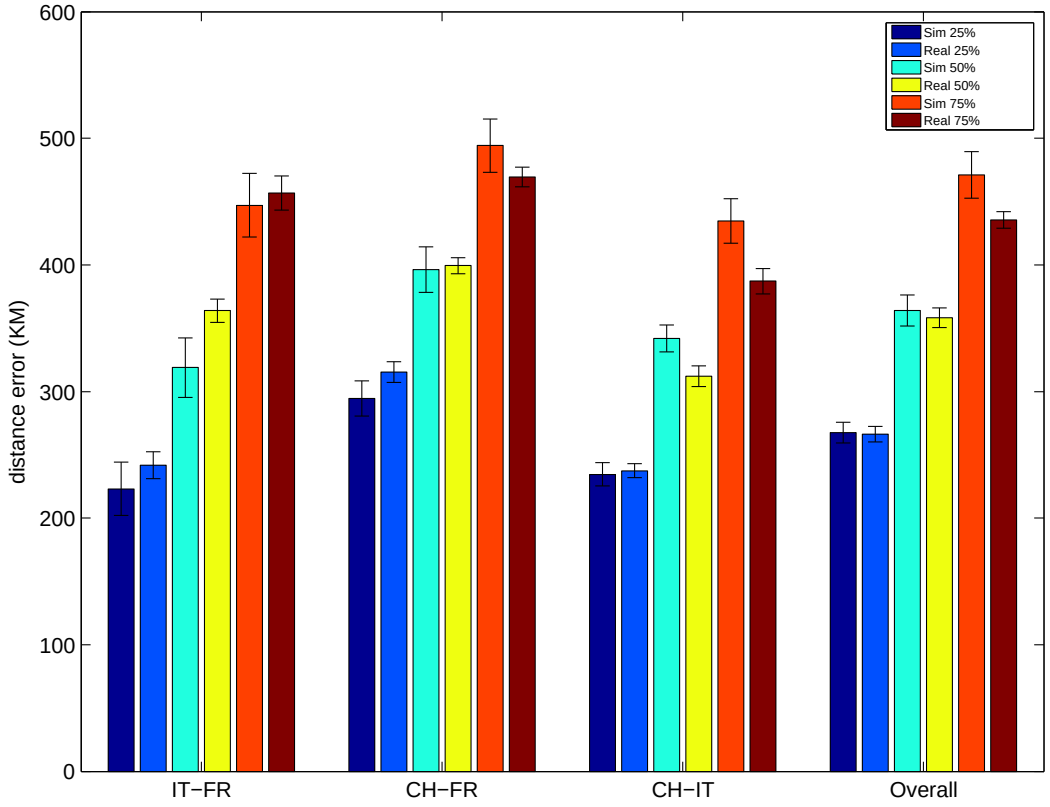

FIGURE 6.1: **Distance error of simulation and real data**. We tested SLAM accuracy on the 3 main groups of admixed populations that exist in the real data (IT-FR, CH-FR, CH-IT). We compared the quantile results of 25%, 50% (median) and 75% of the samples on both datasets. The last group ("Overall") includes samples from all 3 groups.

Chapter 7

Spatial ancestry inference in NHL dataset

The NHL (Non Hodgkin Lymphoma) study is a case control study based on interviews and biological samples among NHL patients and controls in Israel and the West Bank. Genotyping was performed using the Illumina platform at the Genome Institute of Singapore. After quality control procedures the dataset consists of 587,609 SNPs. We used SLAM to estimate the locations of origin of the Jewish samples from the NHL dataset.

We collected a total of 193 individuals of homogeneous ancestry, who originated from the following Jewish communities: Ashkenazi (Russian, Romanian and Hungarian), Middle Eastern (Iranian and Iraqi), North African (Moroccan, Tunisian and Libyan), Sephardi (Bulgarian and Turkish) and Yemenite Jewish community.

First, we used PCA [33] and LOCO-LD [37] to infer the location of the homogeneous individuals. Both methods were able to correctly cluster the different individuals according to their origins. Similarly to the results of previous studies [48, 49], we obtained a tight cluster of Ashkenazi Jews, with no clear substructure between the different European regions.

Next, we simulated admixed haplotypes in a manner identical to the POPRES simulation described in 4. We localized 250 simulated individuals of 1-generation admixture, considering Ashkenazi Jews to originate from a single origin (see Figure 7.1).

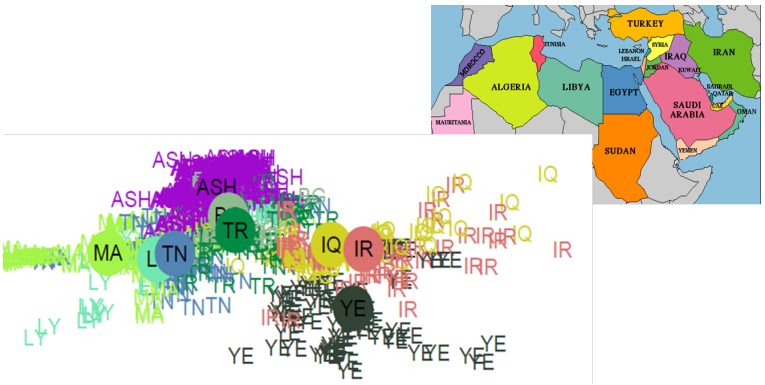

FIGURE 7.1: **SLAM's results on simulated one generation mixture on the NHL dataset (250 individuals)**. The abbreviations are: ASH, Ashkenazi; YE, Yemen; IR, Iran; IQ, Iraq; TN, Tunisia; LY, Lybia; MA, Morocco; TR, Turkey; BG, Bulgaria;

We set SLAM's parameters to $l = 50$, $\lambda = 0.01$ and $p_s = 0.1$; the smaller window length compared to POPRES was chosen due to the higher phasing error rate in the NHL simulations (7.5% in NHL vs. 7.1% in POPRES for chromosomes 18-21), as well as due to the higher density of the SNPs in the NHL dataset (a SNP every 4750 bases on average compared to 7650 in POPRES).

SLAM achieved an overall success rate of 58% in classifying haplotypes to origin (See Figure 7.2), which is similar to the success rate on the simulated European individuals presented in section 5.1. As expected, admixed individuals whose origins are far apart (*e.g.* Ashkenazi and Yemeni) were handled more successfully than individuals whose two origins are close to each other (*e.g.* Lybian and Yemeni).

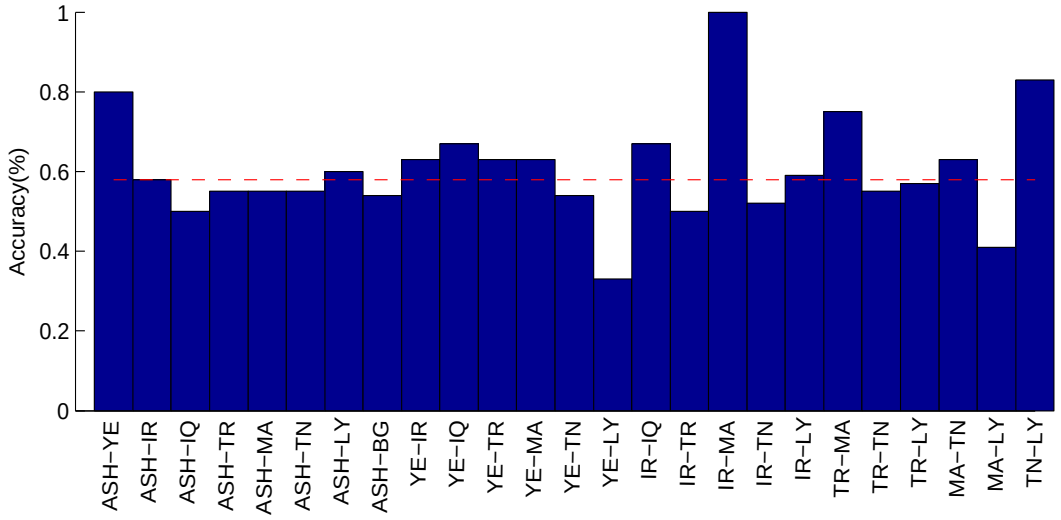

FIGURE 7.2: **Prediction accuracy results of assigning simulated admixed populations to their origins in the NHL dataset**. The abbreviations are: ASH, Ashkenazi; YE, Yemen; IR, Iran; IQ, Iraq; TN, Tunisia; LY, Lybia; MA, Morocco; TR, Turkey; BG, Bulgaria; The average accuracy (red cross lines on the plot) are calculated across admixed populations given equal weights.

Chapter 8

Discussion

In this thesis we presented SLAM, a novel method for the inference of ancestral locations of individuals of recent admixed ancestry. SLAM is the first method capable of accurately inferring the spatial location of individuals with multiple generations of admixture. Its accuracy is achieved by modeling both long-range correlations due to recent admixture, and short-range correlations due to linkage disequilibrium.

There are a few natural potential extensions to SLAM. First, SLAM makes some simplifying assumptions, such as lack of phasing errors within windows and lack of recombinations with block, in addition to setting fixed window length and block size across the genome. A more flexible model in which the window and block lengths vary across the genome, and a small number of recombinations and phasing errors are allowed within a block and a window respectively, should lead to more accurate results.

Second, prior information about some of the ancestral locations or about proximity between subsets of the locations could potentially enable accurate origin inference in individuals with 3-generation admixture and above.

Finally, even though this was not the focus of the current work, SLAM provides locus-specific location estimates which could potentially be used in population genetic studies, analogously to the discrete local ancestry estimates produced for recently admixed populations by methods such as [29, 30].

Bibliography

[1] Luigi Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza. *The history and geography of human genes*. Princeton university press, 1994.

[2] Peristera Paschou, Jamey Lewis, Asif Javed, and Petros Drineas. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of medical genetics*, pages jmg–2010, 2010.

[3] Mikako Mori, Ryo Yamada, Kyoko Kobayashi, Reimi Kawaida, and Kazuhiko Yamamoto. Ethnic differences in allele frequency of autoimmune-disease-associated snps. *Journal of human genetics*, 50(5):264–266, 2005.

[4] Todd Bersaglieri, Pardis C Sabeti, Nick Patterson, Trisha Vanderploeg, Steve F Schaffner, Jared A Drake, Matthew Rhodes, David E Reich, and Joel N Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6):1111–1120, 2004.

[5] Laurie J Ozelius, Geetha Senthil, Rachel Saunders-Pullman, Erin Ohmann, Amanda Deligtisch, Michele Tagliati, Ann L Hunt, Christine Klein, Brian Henick, Susan M Hailpern, et al. Lrrk2 g2019s as a cause of parkinson's disease in ashkenazi jews. *New England Journal of Medicine*, 354(4):424–425, 2006.

[6] Neil Risch, Kathleen Merikangas, et al. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[7] Sharon R Browning and Brian L Browning. 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.

[8] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35, 2010.

[9] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[10] Michael F Seldin, Bogdan Pasaniuc, and Alkes L Price. New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, 12(8):523–528, 2011.

[11] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[12] Jun J Yang, Cheng Cheng, Meenakshi Devidas, Xueyuan Cao, Yiping Fan, Dario Campana, Wenjian Yang, Geoff Neale, Nancy J Cox, Paul Scheet, et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nature genetics*, 43(3):237–241, 2011.

[13] K. Bryc, C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C.D. Bustamante, and H. Ostrer. 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.

[14] J.P. Jarvis, L.B. Scheinfeldt, S. Soi, C. Lambert, L. Omberg, B. Ferwerda, A. Froment, J.M. Bodo, W. Beggs, G. Hoffman, et al. Patterns of ancestry, signatures of natural selection, and genetic association with stature in western african pygmies. *PLoS genetics*, 8(4):e1002641, 2012.

[15] A.G. Hinch, A. Tandon, N. Patterson, Y. Song, N. Rohland, C.D. Palmer, G.K. Chen, K. Wang, S.G. Buxbaum, E.L. Akylbekova, et al. The landscape of recombination in african americans. *Nature*, 476(7359):170–175, 2011.

[16] D. Wegmann, D.E. Kessner, K.R. Veeramah, R.A. Mathias, D.L. Nicolae, L.R. Yanek, Y.V. Sun, D.G. Torgerson, N. Rafaels, T. Mosley, et al. Recombination rates in admixed individuals identified by ancestry-based inference. *Nature genetics*, 43(9):847–853, 2011.

[17] S. Gravel, B.M. Henn, R.N. Gutenkunst, A.R. Indap, G.T. Marth, A.G. Clark, F. Yu, R.A. Gibbs, C.D. Bustamante, D.L. Altshuler, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

[18] Kristin G Ardlie, Leonid Kruglyak, and Mark Seielstad. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309, 2002.

[19] R Kaeuffer, D Réale, DW Coltman, and D Pontier. Detecting population structure using structure software: effect of background linkage disequilibrium. *Heredity*, 99(4):374–380, 2007.

[20] David B Goldstein and Michael E Weale. Population genomics: linkage disequilibrium holds the key. *Current Biology*, 11(14):R576–R579, 2001.

[21] Buhm Han, Brian M Hackel, and Eleazar Eskin. Postassociation cleaning using linkage disequilibrium information. *Genetic epidemiology*, 35(1):1–10, 2011.

[22] Melissa J Hubisz, Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inferring weak population structure with the assistance of sample group information. *Molecular ecology resources*, 9(5):1322–1332, 2009.

[23] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS genetics*, 8(1):e1002453, 2012.

[24] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[25] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[26] Sriram Sankararaman, Srinath Sridhar, Gad Kimmel, and Eran Halperin. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303, 2008.

[27] Bogdan Paşaniuc, Sriram Sankararaman, Gad Kimmel, and Eran Halperin. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):i213–i221, 2009.

[28] Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller, David A Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J OBrien, David Altshuler, et al. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.

[29] Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, 5(6):e1000519, 2009.

[30] Yael Baran, Bogdan Pasaniuc, Sriram Sankararaman, Dara G Torgerson, Christopher Gignoux, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G Ford, Pedro C Avila, et al. Fast and accurate inference of local ancestry in latino populations. *Bioinformatics*, 28(10):1359–1367, 2012.

[31] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[32] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[33] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[34] Chaolong Wang, Sebastian Zöllner, and Noah A Rosenberg. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS genetics*, 8(8):e1002886, 2012.

[35] Wen-Yun Yang, John Novembre, Eleazar Eskin, and Eran Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nature genetics*, 44(6):725–731, 2012.

[36] Alkes L Price, Michael E Weale, Nick Patterson, Simon R Myers, Anna C Need, Kevin V Shianna, Dongliang Ge, Jerome I Rotter, Esther Torres, Kent D Taylor, et al. Long-range ld can confound genome scans in admixed populations. *American journal of human genetics*, 83(1):132, 2008.

[37] Yael Baran, Inés Quintela, Ángel Carracedo, Bogdan Pasaniuc, and Eran Halperin. Enhanced localization of genetic samples through linkage-disequilibrium correction. *The American Journal of Human Genetics*, 92(6):882–894, 2013.

[38] John E Pool and Rasmus Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719, 2009.

[39] Jeffrey C Long. The genetic structure of admixed populations. *Genetics*, 127(2):417–428, 1991.

[40] Julian Adams and Richard H Ward. Admixture studies and the detection of selection. *Science*, 180(4091):1137–1143, 1973.

[41] Emile R Chimusa, Noah Zaitlen, Michelle Daya, Marlo Möller, Paul D van Helden, Nicola J Mulder, Alkes L Price, and Eileen G Hoal. Genome-wide association study of ancestry-specific tb risk in the south african coloured population. *Human molecular genetics*, 23(3):796–809, 2014.

[42] Stig E Bojesen, Karen A Pooley, Sharon E Johnatty, Jonathan Beesley, Kyriaki Michailidou, Jonathan P Tyrer, Stacey L Edwards, Hilda A Pickett, Howard C Shen, Chanel E Smart, et al. Multiple independent variants at the tert locus are associated with telomere length and risks of breast and ovarian cancer. *Nature genetics*, 45(4):371–384, 2013.

[43] Katherine A Drake, Dara G Torgerson, Christopher R Gignoux, Joshua M Galanter, Lindsey A Roth, Scott Huntsman, Celeste Eng, Sam S Oh, Sook Wah Yee, Lawrence Lin, et al. A genome-wide association study of bronchodilator response in latinos implicates rare variants. *Journal of Allergy and Clinical Immunology*, 133(2):370–378, 2014.

[44] Wen-Yun Yang, Alexander Platt, Charleston Wen-Kai Chiang, Eleazar Eskin, John Novembre, and Bogdan Pasaniuc. Spatial localization of recent ancestors for admixed individuals. *bioRxiv*, 2014.

[45] Matthew R Nelson, Katarzyna Bryc, Karen S King, Amit Indap, Adam R Boyko, John Novembre, Linda P Briley, Yuka Maruyama, Dawn M Waterworth, Gérard Waeber, 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.

[46] Ekrame Boubtane, Jean-Christophe Dumont, et al. Immigration and economic growth in the oecd countries 1986-2006: A panel data analysis. 2013.

[47] Jianzhong Ma and Christopher I Amos. Principal components analysis of population admixture. *PloS one*, 7(7):e40115, 2012.

[48] Shai Carmi, Ken Y Hui, Ethan Kochav, Xinmin Liu, James Xue, Fillan Grady, Saurav Guha, Kinnari Upadhyay, Dan Ben-Avraham, Semanti Mukherjee, et al. Sequencing an ashkenazi reference panel supports population-targeted personal genomics and illuminates jewish and european origins. *Nature communications*, 5, 2014.

[49] Doron M Behar, Bayazit Yunusbayev, Mait Metspalu, Ene Metspalu, Saharon Rosset, Jüri Parik, Siiri Rootsi, Gyaneshwer Chaubey, Ildus Kutuev, Guennady Yudkovsky, et al. The genome-wide structure of the jewish people. *Nature*, 466(7303):238–242, 2010.