University of California

Los Angeles

Efficient Probabilistic Model Based Approaches for Analysis of Human Genomic Data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Wen-Yun Yang

2013

Abstract of the Dissertation

Efficient Probabilistic Model Based Approaches for Analysis of Human Genomic Data

by

Wen-Yun Yang

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2013

Professor Eleazar Eskin, Chair

The advent of genotyping and sequencing technologies has enabled human genetics to discover numerous genetic variants and perform analysis in the level of populations. Understanding the genetic diversity of populations has broad applications in studies of human disease, history, and the relationships within and among populations. I propose a new approach, spatial ancestry analysis, for the modeling of genotypes in two and three dimensional space. I show that the explicit modeling of the allele frequency allows us to localize individuals on the geographical map based on their genetic information alone. Furthermore, a direct probabilistic interpretation of our model enables us to accurately predict geographical origins of an individual even when the individual has mixed ancestry. In addition, the analysis also identifies additional genes, e.g., *FOXP2*, *OCA2* and *LRP1B*, that have extreme allele frequency gradients that may have been due to selection.

I therefore generalize the spatial ancestry analysis based on hidden Markov models of admixture along with a model of spatial distribution of variants to infer the location of the ancestors jointly with assigning ancestry at each locus in the genome of admixed individuals. This generalized approach is able to localize their recent ancestors with an average of 470Km of the reported locations of their grandparents, for mixed European ancestries.

I propose a novel framework for haplotype inference from short read sequencing that leverages multi-SNP reads together with a reference panel of haplotypes. The basis of our approach is a new probabilistic model that finds the most likely haplotype segments from the reference panel to explain the short read sequencing data for a given individual. We devised an efficient sampling method within a probabilistic model to achieve superior performance than existing methods. Using simulated sequencing reads from real individual genotypes in the HapMap data and the 1000 Genomes projects, we show that our method is highly accurate and computationally efficient.

Finally, I introduce a novel spatial-aware haplotype copying model, which assumes that any chromosome can be modeled as a mosaic of segments copied from a set of sampled chromosomes, but chromosomes that are closest in the genetic-geographic continuum map are a priori more likely to contribute to the copying process than distant ones. This model has various potential applications. In particular, I show that this model achieves superior accuracy in genotype imputation over the standard spatial-unaware haplotype copy model. In addition, I also show the utility of this model in selecting a small personalized reference panel for imputation that leads to both improved accuracy as well as to a lower computational runtime than the standard approach.

The dissertation of Wen-Yun Yang is approved.

Jason Ernst

Wei Wang

Bogdan Pasaniuc

Eleazar Eskin, Committee Chair

University of California, Los Angeles

2013

To my mother and father

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost I would like to thank my advisors Eleazar Eskin and Bogdan Pasaniuc for their careful and insightful guidance for my research, their high standards on quality and elegance of scientific work, and their always friendly, patient and understanding in the past years. They have been a wonderful source of knowledge and encouragement. This dissertation is deeply indebted to them.

I am also very fortunate to collaborate with John Novembre and Eran Halperin. A significant portion of this dissertation is from close collaboration with them. My rotation with John Novembre in my first year exposed me to the field of population genetics, which I continued to find fascinating during my PhD years. The collaboration and discussion with Eran Halperin are always inspiring and exciting.

I am grateful to all the members of my thesis committee: Wei Wang, Jason Ernst, Bogdan Pasaniuc and Eleazar Eskin. They provided valuable feedback on my thesis and helped guide it to completion.

I give thanks to my fellow graduate students in Both Zarlab and Bogdan's group for the help they have given me, especially Farhad Hormozdiari, Zhanyong Wang, Nathaniel Parrish, Dan He, Buhm Han, Emrah Kostem, Nicholas Furlotte, Jae-Hoon Sul, Eun Yong Kang, Joanne Joo, Dat Duong, Robert Brown, Gleb Kichaev, Huwenbo Shi for making the two lab offices much more fun places to stay.

I must thank all my Chinese friends with whom I have shared many good times and from whom I have gained strength during difficult times. There are so many of them I wish to thank but hopefully they know who they are. To name just a few: Zhanyong Wang, Feng Guan, Teng Wang, Changyong Yin, Peng Wang and Xue Gao.

Finally, I would like to thank my parents and my fiancee Huiling Zhang for their patient love and consistent support. Without their love, I would never make this PhD journey so enjoyable and unforgettable. I am forever grateful to them.

Chapter 2 was published in Nature Genetics, 44(6):725-731, June 2012. Wen-Yun Yang, John Novembre, Eleazar Eskin and Eran Halperin, “A Model Based Approach for Analysis of Spatial Structure in Genetic Data” The dissertation author was the primary investigator and author of this paper.

Chapter 3 is currently in submission for publication. Wen-Yun Yang, Alexander Platt, Charleston Wen-Kai Chiang, Eleazar Eskin, John Novembre and Bogdan Pasaniuc,“Spatial Localization of Admixed Individuals” The dissertation author was the primary investigator and author of this paper.

Chapter 4 was published in Bioinformatics, 29(18):2245-2252, September 2013. Wen-Yun Yang, Farhad Hormozdiari, Zhanyong Wang, Dan He, Bogdan Pasaniuc and Eleazar Eskin, “Leveraging Reads that Span Multiple Single Nucleotide Polymorphisms for Haplotype Inference from Sequencing Data”. The dissertation author was the primary investigator and author of this paper.

Chapter 5 is currently in submission for publication. Wen-Yun Yang, Farhad Hormozdiari, Eleazar Eskin and Bogdan Pasaniuc, “A Spatial-Aware Haplotype Copying Model with Applications to Genotype Imputation” The dissertation author was the primary investigator and author of this paper.

Vita

1984	Born, Wutai, Shanxi, China
2006	Bachelor of Engineering in Computer Science, Shanghai Jiao Tong University, Shanghai, China
2009	Master of Engineering in Computer Science, Shanghai Jiao Tong University, Shanghai, China
2010-2013	Research Assistant, University of California, Los Angeles
2011-2012	Teaching Assistant, University of California, Los Angeles
2013	Doctor of Philosophy in Computer Science, University of California, Los Angeles

Publications

Wen-Yun Yang, Farhad Hormozdiari, Zhanyong Wang, Dan He, Bogdan Pasaniuc and Eleazar Eskin, Leveraging Reads that Span Multiple Single Nucleotide Polymorphisms for Haplotype Inference from Sequencing Data, *Bioinformatics*, 29(18):2245-2252, September 2013.

Farhad Hormozdiari, Zhanyong Wang, Wen-Yun Yang and Eleazar Eskin, Efficient Genotyping of Individuals using Overlapping Pool Sequencing and Imputation, *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals,* *Systems and Computers (ASILOMAR)*, pp.1023-1027, Pacific Grove, CA, USA, November 2012

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, June 2012.

Zhanyong Wang, Farhad Hormozdiari, Wen-Yun Yang, Eran Halperin and Eleazar Eskin, CNVeM: Copy Number Variation detection Using Uncertainty of Read Mapping, *Proceedings of International Conference on Research in Computational Molecular Biology (RECOMB)*, pp. 326-340, Barcelona, Spain, April 2012

Wen-Yun Yang, Yunbo Cao and Chin-Yew Lin, A Structural Support Vector Method for Extracting Contexts and Answers of Questions from Online Forums, *Information Processing and Management*, 47(6):886-898, November 2011

Wen-Yun Yang, James T. Kwok and Bao-Liang Lu, Spectral and Semidefinite Relaxations of the CLUHSIC Algorithm, *Proceedings of SIAM International Conference on Data Mining (SDM)*, pp.106-117, Columbus, Ohio, USA, April 2010

CHAPTER 1

Introduction

The high throughput genotyping and sequencing technologies for human genetic variant discernment have ushered in a new era of interdisciplinary research between computer science, statistics and biology. For the first time, we are able to collect thousands of individual's genetic data at hundreds of thousands genetic markers, and perform analysis of population genetics, genome-wide association study (GWAS) and so on. In particular, the single nucleotide polymorphism (SNP) is of the main interest in the field of genetics for the last decade, which contains significant amount of information for studies of population structure, ancestry assignment, spatial localization and GWAS study. In this thesis, I am mainly focused on the following four topics.

Understanding how genetic diversity is distributed across different populations has many important applications in modern population genomics. In particular, measures of population structure are used to correct for population stratification in genome-wide association studies [PPP06], for the discovery of novel associations of genetic variation to disease in the context of admixture mapping [SPP11], to detect regions that have undergone recent positive selection [LK73, PCN09, CPN09], and to illuminate interesting aspects of human population history [JSS08, LAT08]. In Chapter 2, I introduce a model based approach for analysis of spatial structure in genetic data, called Spatial Ancestry Analysis (SPA). SPA is a novel probabilistic model for the spatial structure of genetic variation where we explicitly model how the allele frequency of each SNP changes as a function of the location of the individual in space (i.e., the allele frequency is a function of the (x, y) coordinates of an individual on a map). Then, each individual's genotypes are assumed to follow Hardy-Weinberg equilibrium with allele frequencies defined by the individual's location. If the geographical origins of the individuals are known, we can use this information to infer their allele frequency functions at each SNP. However, if the locations are unknown, our model can infer geographical origins for individuals using only their genetic data, similar in spirit to Principle Component Analysis based approaches for spatial assignment. Using this framework we also can detect loci showing extreme patterns of spatial differentiation, for instance as a result of recent positive natural selection and/or allele surfing [ND09, ER08].

In Chapter 3, I generalize the SPA model to consider admixing process when dealing with admixed individuals. I introduce approaches for ancestry inference in recently admixed individuals in a geographic continuum within a model that flexibly handles admixture across varying number of generations and ancestries. We view admixed individuals as having recent ancestors from several locations on a genetic-geographical map. Then, we perform ancestry inference by simultaneously localizing on the map the recent ancestors of an admixed individual and partitioning the individual's genome into segments inherited from the same ancestor (locus-specific ancestry). We take advantage of the observation that if one allele is inherited from a specific ancestor, then most likely, the neighboring alleles are also inherited from the same ancestor. We validate our approach by localizing the recent ancestry of the POPRES individuals with self-reported ancestry from multiple locations in Europe. Our method is able to localize the grandparents of the admixed individuals of POPRES with an average of 470Km of their reported ancestry, ranging from 305Km for individuals with Swiss-French ancestry to 701Km for those with Spanish-Portuguese ancestry.

In Chapter 4, I introduce a novel approach called HARSH (HAplotyping with Reference and Sequencing tecHnology) for haplotype phasing. We utilize a probabilistic model to incorporate the multi-SNP read information together with a reference panel of haplotypes. We use an efficient Gibbs sampling method to find sample from the posterior distribution. This algorithm has the advantages of being computationally efficient, scalable in memory usage and accurate in genotyping and phasing prediction. We evaluate our method on simulations from real haplotypes from the HapMap project. At 1X coverage, HARSH gives around 10% improvement in terms of total error rate compared with standard phasing approaches that do not use the multi-SNP read information thus showing the benefits of modeling multi-SNP reads. We also evaluate HARSH and the basic model for varying coverage and read length, showing the benefits of our approach in higher coverage and longer read length. Additionally, we test our method on simulations starting from real sequencing data of 1000 Genomes project, where the density of SNPs is much higher than that in HapMap data. Through extensive simulations we show that the gain in performance of our approach over existing models extends to realistic read lengths (e.g. $100 - 400$ base pairs) making our approach readily applicable to existing sequencing data sets. With recent works showing that short read sequencing can dramatically increase association power in genome-wide association study (GWAS) over genotyping arrays [PRM12], we expect our approaches to further increase power in GWAS by increasing accuracy in genotype calling and phasing from short read data.

In Chapter 5, I propose a new approach to modeling genetic variation in structured populations that incorporates ideas from both the haplotype copying model [LS03] and the spatial structure framework that models genetic variation as function of geography [YNE12, BQC13]. That is, we propose a haplotype copy model that a priorly up weights the contribution of haplotypes closer in geographical distance to the copying process. We accomplish this by jointly modeling geography and the copying process. Each haplotype is associated with a geographical position; when copying into a new haplotype with known location, we instantiate an HMM that has switching transition probabilities up weighted for haplotypes closer in geographical space to target haplotype. We use real data from the 1000 Genomes project [CAA10] to show that the our spatial-aware approach fits the data significantly better than the standard model. Through a masking procedure followed by a leave-one-out experiment we show that our spatial-aware method significantly increases imputation accuracy especially for lower frequency variation (e.g. an improvement of 6%(2%) for low-frequency(common) variation in Asian data). We also show that our approach can be used to select a small personalized reference panel for imputation that increases imputation accuracy while significantly reducing imputation runtime (up to 10-fold). Finally, we show how our model can be used in a supervised manner to infer locations on the genetic-geographic map for individuals based on their genetic data.

CHAPTER 2

A Model Based Approach for Analysis of Spatial Structure in Genetic Data

2.1 Motivation

Understanding how genetic diversity is distributed across different populations has many important applications in modern population genomics. In particular, measures of population structure are used to correct for population stratification in genome-wide association studies [PPP06], for the discovery of novel associations of genetic variation to disease in the context of admixture mapping [SPP11], to detect regions that have undergone recent positive selection [LK73, PCN09, CPN09], and to illuminate interesting aspects of human population history [JSS08, LAT08].

The scale of modern SNP data has made clear that an individual's DNA encodes a considerable amount of information on the individual's ancestral origin. Multiple empirical SNP surveys have shown how an individual's geographical ancestry can be inferred using the first two principal components (PCs) of the genotype matrix (e.g., [LLN08, NJB08]). This relationship between PCs and geographic origin is expected when the underlying genetic variation is spatially structured [NS08, McV09], that is when genetic similarity decays with the geographic distance between the origins of the individuals. Spatial structure is widespread in human populations due to histories of spatial expansions and spatially restricted mating. While principal component analysis (PCA) can capture the spatial structure of the data, it is not based on an explicit probabilistic model for spatial genetic structure and as a result is less amenable for extensions compared to model-based approaches.

In this paper we develop a novel probabilistic model for the spatial structure of genetic variation where we explicitly model how the allele frequency of each SNP changes as a function of the location of the individual in space (i.e., the allele frequency is a function of the (x, y) coordinates of an individual on a map). Then, each individual's genotypes are assumed to follow Hardy-Weinberg equilibrium with allele frequencies defined by the individual's location. The family of functions we use to model allele frequency over space is deliberately simple, but leads to tractable inference algorithms with several applications.

If the geographical origins of the individuals are known, we can use this information to infer their allele frequency functions at each SNP. However, if the locations are unknown, our model can infer geographical origins for individuals using only their genetic data, similar in spirit to PCA-based approaches for spatial assignment. This provides evidence that our modeling of allele frequencies, albeit simple, is sensitive and captures the information about spatial location inherent in most variants. Since our approach is model-based, the model can predict the geographical origins of an individual even in the case where the individual is of mixed ancestry by utilizing the fact that the model provides an explicit representation of the allele frequency as a function of the map coordinates. This is not possible in other approaches such as PCA, which is based on a linear combination of genotypes and therefore, for example, will lead to an individual with an Italian and Swedish parents being assigned to Central Europe. Instead, the approach taken here can recover the disparate parental origins. We also show how our approach can be extended to model spatial structure over a sphere to predict the spatial structure of worldwide populations.

Using this framework we also can detect loci showing extreme patterns of spatial differentiation, for instance as a result of recent positive natural selection and/or allele surfing [ND09, ER08]. When we applied our approach (spatial ancestry analysis, or *SPA*) to human population genetic data, we observed that some of the outlier regions detected by SPA have been found with previous methods designed to detect recent positive selection, such as iHS [VKW06], F_{ST} [LK73, HW09] and the method Bayenv presented in Coop et al.,[CWD10]; for example, the *LCT* and *HLA* regions. In contrast to previous methods, our method is unique in being especially sensitive to strong spatial patterns, and works at the individual-level rather than partitioning individuals into populations. The SPA method is particularly sensitive to SNPs that have steep geographical gradients in allele frequency, while F_{ST}-based approaches simply highlight loci that have large variation in allele frequency.

2.2 Method

2.2.1 Genetic Spatial Structure Model

We assume we are given genotypes at a L single nucleotide polymorphisms (SNPs) from N unrelated individuals drawn from different populations distributed across the geographical region under consideration. We assume that the allele frequency of a SNP j is a function

$$f_j(x) = \frac{1}{1 + \exp(-a_j^T x - b_j)} \tag{2.1}$$

where a_j, b_j depend on the SNP j, and x is the K dimensional vector of coordinates describing the spatial positioning of an individual. Typically, $K = 2$ for geographical position. Clearly, this function has a range $[0, 1]$ that can be interpreted as a probability, and thus the likelihood of the data can be easily expressed as a function of the values of a, b and x.

Let g_{ij} represent the observed number of minor alleles at SNP j of individual i and let f_{ij} be a shorthand for $f_j(x_i)$ where x_i is the position of individual i. Since the individuals are independently sampled from the population, the log-likelihood of the entire observed sample can be calculated from the log-likelihood for each genotype

$$L(G; X, A, B) \quad \propto \quad \sum_i \sum_j \left[g_{ij} \ln f_{ij} + (2 - g_{ij}) \ln(1 - f_{ij})\right] \tag{2.2}$$

The parameter matrices $X = \{x_{ik}\}$, $A = \{a_{jk}\}$ and $B = \{b_j\}$ are $N \times K$, $L \times K$ and $L \times 1$ matrices, respectively. Specifically, each row of X contains the geographical location for each individual. Each row of A and B contains the coefficient for each allele frequency function.

2.2.2 Maximum Likelihood Estimation

Given the above likelihood model and a set of genotypes, we are interested in the matrices X, A, and B that maximize the log likelihood above. The above likelihood function is not concave, and it is therefore hard to optimize. We note however, that when X is fixed or when A, B are fixed, then the objective function (2.2) is concave. We therefore use alternative maximization in conjunction with Newton's method. Furthermore for fixed A and B, the objective function in X can be decomposed into a series of unrelated parts, each of which corresponds to one row in X, and therefore the update of X can be decomposed into a series of much smaller problems, which further simplifies the optimization.

After the simplification of the above alternative maximization and variable separations of the function (2.2), we now arrive at the following two unconstrained convex programming problems in only K variables and $K + 1$ variables, respectively.

$$\min_{x_i} \sum_j \left[g_{ij} \ln(1 + \exp(-a_j^T x_i - b_j)) + (2 - g_{ij}) \ln(1 + \exp(a_j^T x_i + b_j))\right] \tag{2.3}$$

$$\min_{a_j, b_j} \sum_j \left[g_{ij} \ln(1 + \exp(-a_j^T x_i - b_j)) + (2 - g_{ij}) \ln(1 + \exp(a_j^T x_i + b_j))\right] \tag{2.4}$$

The smooth and continuous property of this problem allows us flexibility in the choice of optimization method. We apply Newton's method which widely known for fast convergence, as it utilizes the first and second order derivatives. The details of the algorithm are given in the Supplementary Methods.

2.2.3 An Extended Model For an Admixed Individual

Instead of identifying one origin for an admixed individual, our method can infer two geographical origins for the parents. First, let x and y denote the locations of the two parents of a given admixed individual, and two shorthands $p_j = f_j(x)$ and $m_j = f_j(y)$ denote the allele frequency of those at marker j, where the function $f_j(.)$ is defined in (2.1).

Therefore, again under the assumption of independent SNPs, the genotype of the admixed individual is drawn from the following distribution

$$\begin{aligned} P(g_j = 2|x, y) &= p_j m_j \\ P(g_j = 1|x, y) &= p_j(1 - m_j) + m_j(1 - p_j) \\ P(g_j = 0|x, y) &= (1 - p_j)(1 - m_j). \end{aligned}$$

This distribution assumes that the two alleles of admixture individuals are drawn from the parents independently. Finally, we can infer the location of the parents by maximizing the log-likelihood function

$$L(g; x, y) = \sum_j \ln P(g_j|x, y) \tag{2.5}$$

This likelihood function is not concave. Thus, instead of directly using Newton's method that will cause numerical problems, we use Pseudo-Newton's method to optimize this function in x and y. The algorithm details are given in the Supplementary Methods.

2.2.4 Globe Mapping

For globe mapping, we have to extend the two dimensional vector x to three dimensions. Then, by similar derivation as two dimensional mapping, we can obtain the log-likelihood function in the same form as in (2.2), but in a different number of dimensions. To guarantee the placement of individuals in a sphere, we need to enforce the constraint $||x_i||_2 = 1$ while maximizing the log-likelihood. However, this additional constraint and its non-convexity does not allow us to use Newton's method. Instead, we turn to another widely known optimization techniques called *gradient projection* [NW00], which can handle simple constraints in the optimization problem. Basically, it modifies the line search step in gradient descent method to make sure the current solution is in the feasible region. One key step is the projection from any point to the feasible region. The projection to a sphere can be very efficiently computed by $P(x) = \frac{x}{||x||_2}$.

2.2.5 Evaluation of Individual Mapping

SPA can be applied in the case that the geographical origins of the individuals are known as well as in the case where the geographical origins are unknown. If the geographical origins are known, the slope functions parameterized by a_j and b_j are estimated using these known locations and will be concordant with actual geography. In this case, the output of individual mapping is immediately latitude and longitude.

If the geographical origins are unknown to SPA, the mapping coordinates might be different from real geography in latitude and longitude, up to an affine transformation. In order to do spatial assignment, we follow the approach taken in [NJB08], and assume the following model between mapping coordinates and geographical locations

$$\begin{aligned} u &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \\ v &= \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1^2 + \alpha_4 x_2^2 + \alpha_5 x_1 x_2 \end{aligned}$$

where u and v are latitude and longitude, respectively. $x = (x_1, x_2)$ is the coordinates from our model. The parameters α and β can be estimated from a few individuals with known mapping coordinates and geographical locations. The same model was used in [NJB08] in order to estimate the accuracy of PCA geographical assignments.

The accuracy evaluations in Tables 2.1 and 2.3 are computed based on the spatial assignment. We follow a similar leave-one-out strategy to the one used in [NJB08]. First, we can estimate the coefficients α and β by performing a least-square regression from the mapping coordinates to the true geographical location in latitude and longitude with a leave-one-out training set of individuals. Then for a test individual, we make a prediction of its geographical location using the obtained regression coefficients α and β. We also predict its population origin by assigning it to the nearest country center. The assignment accuracy for a given population is then calculated as the number of correct predictions divided by the total number of individuals in that population.

2.2.6 Characterization of Extreme Allele Frequency Gradients

The outputs of SPA model would be individual mapping coordinates X and coefficients for allele frequency slope functions A and B. Based on those two outputs, in the model, all individuals will have allele frequencies $\mathbf{f}_j = \{f_j(x_1), f_j(x_2), \ldots, f_j(x_N)\}$ organized in a slope corresponding to each SNP j.

A straightforward statistic to quantify the "steepness" of allele frequency slope is as follows

$$SPA_j = \sqrt{\sum_i \left(f_j(x_i) - \frac{\sum_i f_j(x_i)}{N} \right)^2} \tag{2.6}$$

where $f_j(x_i) = 1/(1 + \exp(-a_j^T x_i - b_j))$ stands for the allele frequency for the individual i at locus j. This score is exactly proportional to the standard deviation of $\mathbf{f}_j$ by a constant $\sqrt{1/(N-1)}$.

2.2.7 Newton's method for optimizing SPA likelihood function

Newton's method is a widely known algorithm for minimizing a convex function. In each iteration, it needs the first and second derivatives to determine the search direction. For x_i in (2.3), the first and second derivatives can be efficiently calculated as follows

$$\begin{aligned} \frac{\partial L}{\partial x_i} &= -\sum_j [g_{ij}(1-f_{ij}) - (2-g_{ij})f_{ij}] \cdot a_j \\ \frac{\partial^2 L}{\partial x_i^2} &= \sum_j 2f_{ij}(1-f_{ij}) a_j a_j^T \end{aligned}$$

Similarly, for a_j and b_j in (2.4) those can be calculated as follows

$$\begin{aligned} \frac{\partial L}{\partial a_j} &= -\sum_i [g_{ij}(1-f_{ij}) - (2-g_{ij})f_{ij}] \cdot x_i \\ \frac{\partial L}{\partial b_j} &= -\sum_i [g_{ij}(1-f_{ij}) - (2-g_{ij})f_{ij}] \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 L}{\partial a_j^2} &= \sum_i 2f_{ij}(1-f_{ij}) x_i x_i^T \\ \frac{\partial^2 L}{\partial b_j^2} &= \sum_i 2f_{ij}(1-f_{ij}). \end{aligned}$$

The computational complexity for each iteration of this algorithm is $O(NLK^2)$. The total computational time depends on the number of iterations for the algorithm to converge. In the data sets we analyzed in this paper, we used 10 to 20 iterations.

2.2.8 Pseudo-Newton's method for admixed individual positioning

To achieve fast convergence for the admixed individual positioning problem, we again use Newton's method to optimize the log-likelihood function (2.5). The first and second derivatives in x and y up to a constant can be computed as follows:

$$
\begin{aligned}
\frac{\partial L}{\partial x} &= \sum_j \left[I(g_j = 2)(1-p_j) + I(g_j = 1)t_1 + I(g_j = 0)(-p_j) \right] a_j \\
\frac{\partial^2 L}{\partial x^2} &= \sum_j \left[I(g_j = 2)(1-p_j)p_j + I(g_j = 1)t_2 + I(g_j = 0)(1-p_j)p_j \right] (-a_j a_j^T) \\
\frac{\partial^2 L}{\partial x \partial y} &= \sum_j I(g_j = 1) \left[\frac{m_j(1-m_j)(1-2m_j)p_j(1-p_j)(1-2p_j)}{\left[(1-m_j)p_j + (1-p_j)m_j\right]^2} \right. \\
&\qquad \left. + \frac{2m_j(1-m_j)p_j(1-p_j)}{(1-m_j)p_j + (1-p_j)m_j} \right] (-a_j a_j^T)
\end{aligned}
$$

where I is an indicator function equal to one if the condition holds and zero otherwise, and

$$
\begin{aligned}
t_1 &= \frac{(1-2m_j)(1-p_j)p_j}{p_j(1-m_j) + m_j(1-p_j)} \\
t_2 &= (1-2m_j)\frac{\frac{(1-m_j)p_j}{1-p_j} - \frac{m_j(1-p_j)}{p_j}}{\left(\frac{1-m_j}{1-p_j} + \frac{m_j}{p_j}\right)^2}
\end{aligned}
$$

Note that the first derivative for y and second derivative for y would be the same with above for x by exchanging m_j and p_j.

One minor issue about the objective function in (2.5) is that the function is not concave. Thus, directly using Newton's method will suffer from numerical problem. In practice, we employed a pseudo-Newton's method [NW00] to overcome this non-concavity while maximally preserving the advantages of Newton's method. Instead of directly using the Hessian matrix H, we subtract a constant matrix to make it strictly negative definite, i.e., $H' = H - \delta I$ where I is an identity matrix. This modification to Newton's method enables the algorithm to converge smoothly to a local optima for a non-concave problem.

2.2.9 Web Resources

The software implementation of this method is freely available to public at `http://genetics.cs.ucla.edu/spa`

2.3 Experimental Results

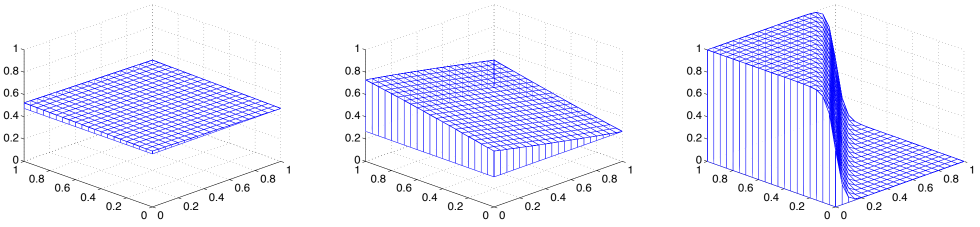
2.3.1 Datasets

We applied our methods to a data set collected from European populations, which was assembled and genotyped as part of the larger POPRES project [NBK08]. A total of 3192 European individuals were genotyped at $500,568$ loci using the Affymetrix 500K SNP chip. After removing SNPs with low-quality scores, the same stringency criteria as previous study [NJB08] were applied to avoid sampling individuals from outside of Europe, to create more even sample sizes across Europe, and to remove individuals whose grandparents have different geographical origins. When available, we use the identical geographical origins of the grandparents as the geographical origin for each individual. Otherwise, we use the self-reported country of birth. As a result, we focus our analysis on genotype data from $447,245$ autosomal loci in $1,385$ individuals from 36 populations.

For the three dimensional globe mapping, we use the Human Genome Diversity Project (HGDP) data consisting of 56 populations from Europe, Africa, Middle East, Central Asia, East Asia, Oceania and native America. In our experiments, we use genotypes at $572,139$ autosomal SNPs in 940 individuals.

2.3.2 Model implementation

The first assumption of our approach is that, the population allele frequency of each SNP can be modeled as a continuous two dimensional function of the position of the individual on the map. Put differently, when sampling a chromosome



(a) Flat Slope (b) Medium Slope (c) Steep Slope

Figure 2.1: Examples of the allele frequency slope model. (a) A SNP with nearly constant allele frequency in all regions of the map. (b) A SNP with gradual allele frequency change. (c) A SNP with a sharp frequency change.

of an individual from position (x, y) in the map, the probability of observing the minor allele in SNP j on the chromosome can be formulated as $f_j(x, y)$, where f_j is a continuous function that describes the allele frequency behavior as a function of the geographic positioning (see Methods). We then make the simplifying assumption that this function is an instance of a logistic function:

$$f_j(\mathrm{vec}x) = \frac{1}{\exp(-a_j^T \mathrm{vec}x - b_j) + 1}.$$

We refer to each of these functions f_j as the slope function of SNP j. This function encodes the "steepness" of the slope by the norm of a, assuming the offset parameter b is fixed. Moreover, the slope directionality is encoded in the value of vector a. In detail, $\theta_j = \arctan(a_j(1)/a_j(2))$ can be taken as angle degree for SNP j, where $a_j(1)$ and $a_j(2)$ are the first and second elements in a. Examples of these functions are shown in Figure 2.1, where the parameter a is set to $[0.1, -0.1]$, $[1, -1]$ and $[30, -30]$, respectively and the parameter b is set to be zero in all three slopes.

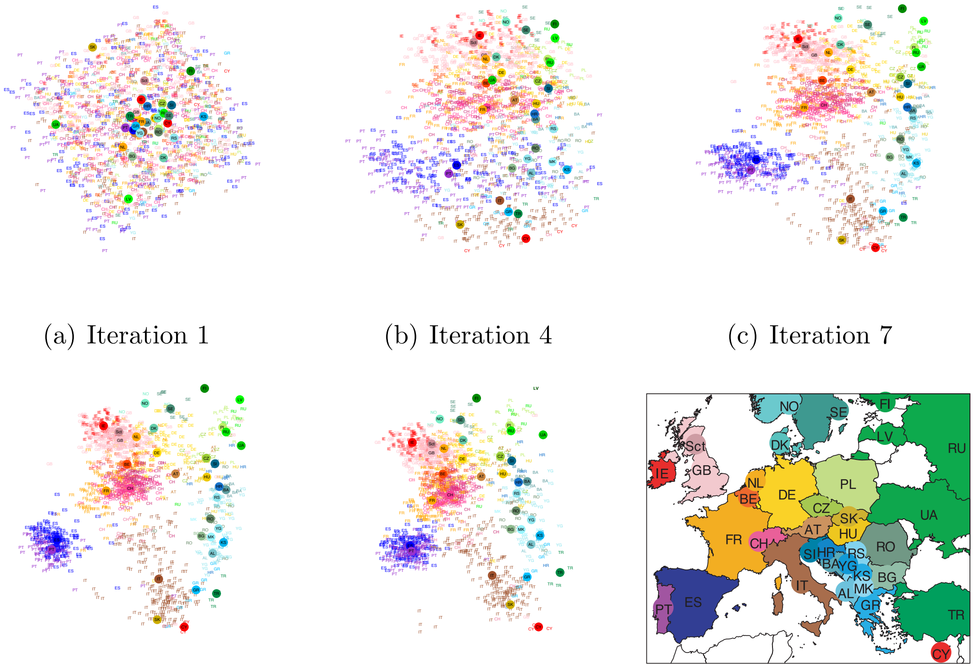
These functions clearly do not capture cases for which SNPs have complicated functions over geographic space, with, for example, multiple modes or peaks in the allele frequency surface; however, these functions should capture general trends in allele frequency where they exist. For spatial assignment applications, as we show, this behavior is not problematic - a substantial amount of information for assignment arises from SNP loci that show gradients across geographic space. Further, when we use our method for detecting extremely differentiated loci, this assumption implies that the method will only detect loci that are extreme in the sense of having steep gradients in allele frequencies (see more below).

The advantage of these functions is that they lend themselves to tractable formulations of the likelihood of genotype data, and we were able to implement efficient Newton's and pseudo-Newton's based-methods for maximizing the likelihood function for the various applications outlined below (see Methods). Utilizing other classes of functions is certainly possible in this framework but may lead to very challenging optimization problems.

2.3.3 Mapping individuals using spatial ancestry analysis (SPA) modeling

As a first application of our approach, we consider a situation similar to that encountered when running PCA on a set of individuals with unknown spatial coordinates to infer their spatial origins. For the SPA method, a challenge of this type of analysis is that neither the spatial coordinates of the individuals nor the slope function for each SNP are given and both must be inferred from the genotypes. The ability to jointly estimate both the allele frequency gradients and the spatial positions of individuals only from the genotype data provides evidence that our model captures spatial genetic structure.

We use a maximum likelihood approach to estimate simultaneously the functions f_j for every SNP j, and the spatial positioning of each of the individuals (see Methods). Roughly, we start by placing the individuals in random positions, and we then iteratively use these positions for the estimation of the slopes functions, followed by using the slope functions to update the individual positions.



(a) Iteration 1

(b) Iteration 4

(c) Iteration 7

(d) Iteration 10

(e) PCA Map

(f) Europe Map

Figure 2.2: Model-based mapping convergence with random initialization. The colors represent the true country of origin of the individual. (a) Iteration 1 starts with a random positioning of individuals. (b) By iteration 4, the northern and southern populations are separated. (c) By iteration 7, the positioning of individuals is close to convergence. (d) In Iteration 10, individuals have reached their final positions. (e) A map generated by PCA [NJB08]. (f) Map of Europe.

We applied SPA to the POPRES samples [NBK08]. The European descendant individuals in this dataset were utilized in [NJB08]. The dataset contains 3192 individuals for which $500,568$ SNPs were genotyped using the Affymetrix 500K SNP chip. For each of the individuals participating in the study the ancestry of the four grandparents is given; we only considered individuals for which all four grandparents have the same ancestry.

Figure 2.2(a)-(d) shows the convergence of the method starting from a random starting point. Figure 2.2(f) shows the map of Europe labeled with the included populations for reference. Interestingly, even though we start the optimization from a set of random positions, after a small number of iterations ($\approx$ 10) the positions of the individuals highly resemble the map of Europe (with only two exceptions, Slovakia (SK) and Russia (RU)). Figure 2.2(e) shows the results of the principal component analysis for comparison. The maps (Figures 2.2(a)-(e)) are rotated by 16 degree in counter-clock direction (similarly to the procedure performed by [NJB08]) to more closely resemble the Europe map. The x and y axes are drawn to equal scale, thus no distortion is involved in the figures. The correlation coefficient between the two maps is 0.99, and thus the two methods provide similar positioning of the individuals, up to an affine transformation. Some noticeable difference are that SPA separates Spain and Portugal more clearly from France than the PCA map. Moreover, the five outlier Italians in the PCA map are drawn closer to Italy by SPA.

Table 2.1 shows the accuracy of the individual placement compared to PCA following the evaluation procedure described in [NJB08]. We compute the accuracy based on spatial assignment (see Methods) that assigns each individual to a country of origin. The results provide support to the notion that the simplified allele frequency functions are capable of extracting the spatial information inherent in the allele frequency data, even when individual spatial coordinates are not provided.

Table 2.1: Individual localization result summary. Based on a spatial assignment method, a country origin is predicted for each individual (See Methods). The accuracy is the proportion of individuals from each country of origin correctly assigned to their true country of origin using a leave-one-out procedure.

Geographical Origin	Number of Individuals	PCA Accuracy	SPA Accuracy
Italy	219	0.70 ± 0.03	0.74 ± 0.03
United Kingdom	200	0.44 ± 0.04	0.53 ± 0.04
Spain	136	0.71 ± 0.04	0.69 ± 0.04
Portugal	128	0.20 ± 0.04	0.38 ± 0.04
Swiss-French	125	0.26 ± 0.04	0.33 ± 0.04
France	89	0.70 ± 0.05	0.66 ± 0.05
Swiss-German	84	0.23 ± 0.05	0.27 ± 0.05
Germany	71	0.25 ± 0.05	0.28 ± 0.05
Ireland	61	0.28 ± 0.06	0.28 ± 0.06
Yugoslavia	44	0.25 ± 0.07	0.30 ± 0.07
Mean	115.7	0.40 ± 0.05	0.45 ± 0.05

SPA can also be applied in the case when a subset of the individuals have known spatial origins and these coordinates are used to infer the spatial origins of a subset of the individuals with unknown origins. In this case, the known spatial origins are used for the placement of the individuals and these placements are used to estimate the functions f_j for every SNP j. We then place each individual with unknown origins using these functions. We evaluate this approach using POPRES data by performing 10-fold cross validation where we use the positions of 90% of our individuals to infer the positions of the remaining 10%. The results of SPA placement assuming known positions is shown in Supplementary Figure 2.3.

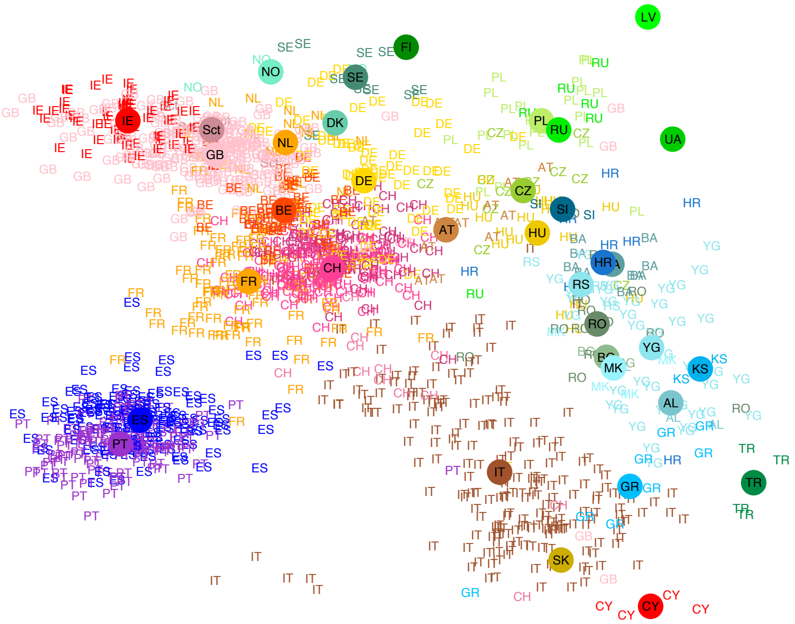


Figure 2.3: Mapping results on POPRES data set by placing individuals using country of origin information. A 10-fold cross validation is performed. In each run, we fit the slope function using the true location information of 90% of the individuals and predict the location for the remaining 10%.

2.3.4 Global Genetic Spatial Structure

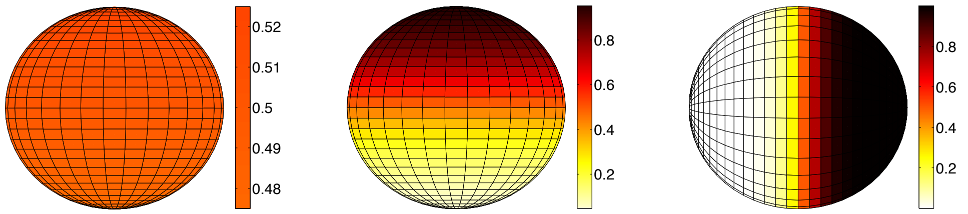
Because SPA has explicit geographical coordinates, the approach can be extended to incorporate coordinate systems beyond the 2-dimensional plane. As a demonstration we extended SPA to analyze the spatial structure of global populations where a two dimensional map cannot accurately capture the structure. We map each individual to a point on a globe in 3-dimensional space. Accordingly, we use a 3-dimensional vector vec$x$ (with the constraint $||\text{vec}x|| = \text{const}$) to represent an individual position. We also need to extend the parameter a to 3-dimensional vector in the logistic function. Examples of these functions with different parameters are shown in Figure 2.4, where the parameter a is set to $[0, 0, 0.1]$, $[0, 0, 3]$ and $[10, 0, 0]$, respectively and the parameter b is set to be zero in all three spheres. The sphere coordinates are drawn from a unit sphere, i.e. $||\mathrm{vec}x|| = 1$.

We apply our global genetic spatial structure method to data from the Human Genome Diversity Panel (HGDP [LAT08]) where 940 individuals from 52 populations worldwide were genotyped across the genome using Illumina Infinium HumanHap550 BeadChips (Figure 2.6). Remarkably, even though we start from a completely random geographical positioning (see Supplementary Figure 2.5), we observe that the resulting positioning highly resembles the world map. Particularly, individuals from the same continents are clustered together and the continents are separated.

By aligning the map in Figure 2.6, we compute the latitude and longitude for each individual and compare with actual geographical position for continents, the SPA map distorts the continent distances but correctly predicts the topology. For example, the longitudinal span of the Eurasia continent is 92 degree on the SPA globe and about 150 degrees on the actual globe. The longitudinal distance between Europe and America is 167 degrees on the SPA globe and about 90 degree on the actual globe. The summary of these comparisons is given in Table 2.2.

2.3.5 Mapping of individuals of mixed ancestry

Using a PCA-based approach, one can infer the localization of an individual with an average error of a few hundred kilometers [NJB08]. However, PCA-based methods are not designed for ancestral origin inference, and particularly if an individual is of mixed ancestry the PCA map will place the individual in the midpoint between the coordinates of the its parents.



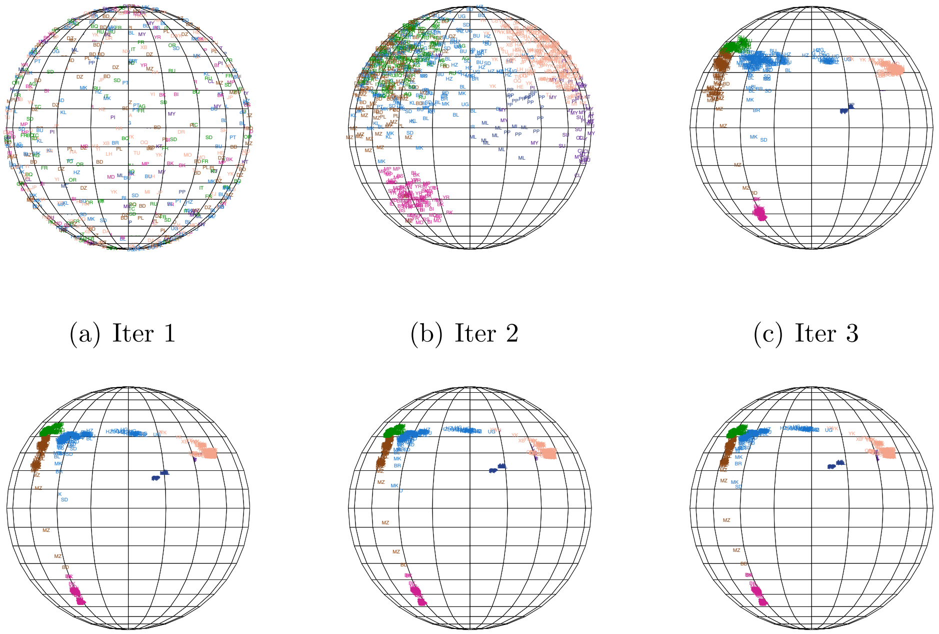
(a) Tiny change (b) Gradual change (c) Sharp change

Figure 2.4: Examples of the allele frequency model for a sphere. The allele frequency is represented by different colors (yellow are low allele frequencies, while red are high allele frequencies). (a) A SNP with constant allele frequency over the sphere. (b) A SNP with graduate allele frequency changes over the sphere. (c) A SNP with sharp frequency changes.

Table 2.2: Summary of SPA globe mapping results. The mean and standard deviation of each continent population are calculated based on an alignment to actual world globe. The positive and negative latitudes stand for north and south latitudes, respectively. The positive and negative longitudes stand for east and west longitude, respectively.

Continent	Pred. Latitude	Pred. Longitude	Actual Latitude	Actual Longitude
Africa	−44.005 ± 4.030	19.548 ± 0.885	1.845 ± 10.977	12.634 ± 16.229
America	21.206 ± 1.988	−151.095 ± 3.449	9.344 ± 16.035	−82.453 ± 18.161
Central South Asia	37.628 ± 3.625	28.047 ± 11.212	32.051 ± 4.722	69.550 ± 3.979
East Asia	30.683 ± 2.617	95.598 ± 4.925	36.685 ± 12.841	115.542 ± 14.250
Europe	38.582 ± 1.434	10.630 ± 2.751	47.732 ± 7.618	13.592 ± 16.317
Middle East	29.799 ± 7.600	12.581 ± 1.277	31.718 ± 0.451	29.307 ± 12.276
Oceania	24.203 ± 1.007	66.530 ± 2.189	−4.741 ± 0.984	147.444 ± 5.905

Because SPA is a model-based approach, it is possible to extend the method to handle individuals of admixed ancestry. As a result, SPA is able to identify which individuals have admixed ancestry and predict the origin of each of the parents by computing the maximum likelihood estimate of the origins of the father and the mother simultaneously, under the assumption that the slope functions are given



(a) Iter 1

(b) Iter 2

(c) Iter 3

(d) Iter 4

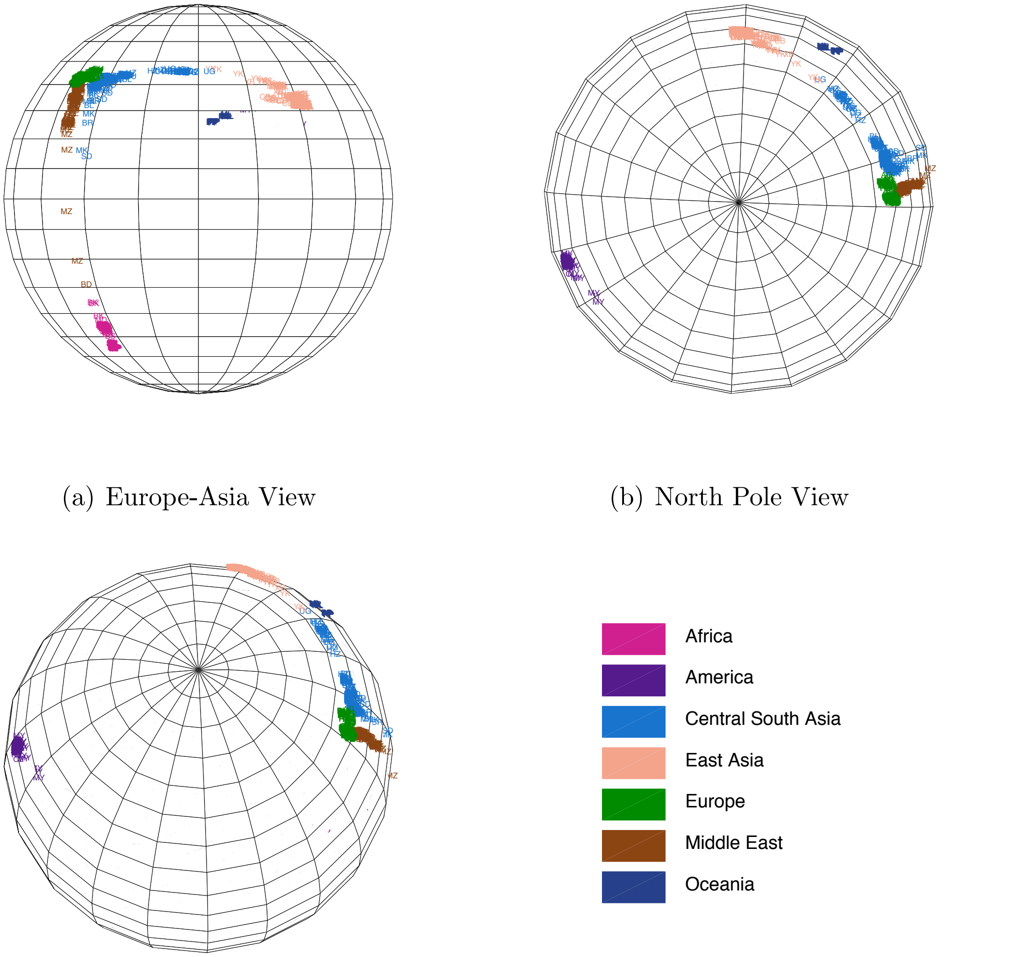
(e) Iter 5

(f) Iter 6

Figure 2.5: Globe mapping convergence with random initialization for individuals from the HGDP data set. The colors represent the continent-level origins for each individual. Iteration 1 starts from random positioning of individuals. By iteration 4, the algorithm separates the continents.

(see Methods). To test this approach, we generated $5,000$ admixed individuals by randomly selecting their parents from the POPRES data set. Each of the parents has four grandparents with the same geographical origin but the four paternal grandparents and four maternal grandparents of the simulated admixed individuals are different. Also, we ignore genders as we only use autosomal SNPs. The offspring's genotype is simulated using Mendelian segregation considering each locus independently.

We then apply SPA to predict the country of origin of the parents where the slope functions are estimated on the set of non-mixed individuals as described above and the results are shown in Table 2.3. We cannot compare the performance of PCA on this simulation since it will only predict one origin for the individual



(a) Europe-Asia View

(b) North Pole View

(c) Atlantic View

Figure 2.6: Mapping spatial structure on a globe using HGDP data. Different colors represent different continents.

Table 2.3: Admixed individual localization result summary. Using genotypes from the 5,000 simulated admixed individuals, SPA is used to predict the origin of each parent.

Origin I	Origin II	Number of Individuals	SPA Accuracy
Italy	United Kingdom	250	0.49 ± 0.03
Italy	Portugal	147	0.49 ± 0.04
Italy	Spain	142	0.68 ± 0.04
Swiss-French	United Kingdom	138	0.21 ± 0.03
Portugal	United Kingdom	137	0.41 ± 0.04
Spain	United Kingdom	128	0.45 ± 0.04
Portugal	Spain	104	0.78 ± 0.04
France	Italy	101	0.57 ± 0.05
Germany	Italy	69	0.43 ± 0.06
Germany	Portugal	60	0.30 ± 0.06
Mean		127.6	0.48 ± 0.04

which is at the midpoint of the true parental origins. Surprisingly, the accuracy for placing the parents of admixed individuals is comparable to the accuracy in placing non-admixed individuals as evident by the comparison of Tables 2.1 and 2.3.

We also evaluated our method on self-reported admixed individuals from the POPRES dataset. We considered individuals who had self reported maternal origins from one country and paternal origins from a different country. We used PCA to evaluate the accuracy of the self-reported ancestry. The PCA should localize an individual of mixed ancestry in the middle point between the parents' locations. However, out of a total of 190 individuals with mixed ancestry in the dataset, only 12 behaved as simple admixtures and were placed by PCA near the midpoint ($<$ 200 kilometers) of their parental origins. The remainder were placed further from the midpoint between the reported parental origins - perhaps suggesting more complex ancestry. By applying SPA to the individuals for which the PCA is located in the midpoint we were able to successfully infer the locations of both parents 58.3% of the time which is comparable to the simulated results.

2.3.6 Detection of loci with extreme gradients in allele frequency

The detection of genomic regions under natural selection sheds light on the functionality of these regions and it provides insights on human history and evolution. A number of methods have been suggested for the detection of selection using genetic variation data, and one particularly common approach leverages the variation in allele frequency between and within populations through the F_{ST} statistic [LK73, HW09]. The F_{ST} approach essentially leverages the insight that variation in allele frequencies across populations should follow a background neutral distribution determined by levels of gene flow and divergence, and that any regions clearly departing from this distribution are regions that putatively have experienced adaptive differentiation or balancing selection in the recent past.

A disadvantage of F_{ST}-based selection detection is that the individual genotypes have to be partitioned into discrete populations. As can be observed in Table 2.1, the definition of a population, for example, in Europe, is rather subjective. Different groupings of the individuals into populations may result in different results, and thus the interpretation of the results is again not straightforward, and particularly important signals of selection may be missed. In addition, F_{ST} is not sensitive to whether allele frequency variation is spatially organized into a steep allele frequency gradient or whether it shows a spatially incoherent pattern.

SPA can be used to detect loci with extreme frequency gradients, and it does not require grouping individuals into populations. We use SPA to identify SNPs which have steep slopes of allele frequency change, with an understanding that a portions of these may have extreme gradients because of the impact of recent positive selection. We develop a new score statistic, measuring the slope of each SNP where large score values correspond to potential regions under selection.

We analyzed the POPRES dataset by applying SPA and extracting SNPs with extreme frequency gradients (see Methods). The distribution of the frequency gradients along with a subset of the SNPs we report can be found in Figure 2.7. The spatial distribution of the typical genes are shown in Figure 2.8. We compared the SNPs found by SPA to the following methods: first, we compute F_{ST} using two types of population partitions, by country and by geographical regions as defined in [NJB08]; second, we compared SPA scores to the widely used iHS method [VKW06], which searches for SNPs with signatures of partial selective sweeps based on haplotype homozygosity, as originally suggested by [SRH02]; and third, we compare SPA to Bayenv [CWD10], which identifies alleles that correlate strongly with an environmental variable, perhaps due to natural selection. For Bayenv [CWD10] we use geographical coordinates as the environmental variable (as if one were searching for latitudinal clines for example). We obtained outlier signals using longitude, latitude, and the individual coordinates corresponding to the first five principal components as the environmental variable.

In Figures 2.9(a) and 2.9(b), we compare the top results of the four methods applied to chromosome 2 and 7. Note that SPA results in a clear cluster of extreme values in $135 - 138$ Mb of chromosome 2 which contains the lactase gene *LCT*. This region is widely noted as a target of strong selection [BSP04], and it is found by all methods. On chromosome 7, SPA detects a strong signal in the *FOXP2* region where all other methods do not.

Overall, the different scores provided by the different methods are moderately correlated ($r^2 < 0.4$, see Table 2.4) even though they each measure unique aspects of genetic variation. Most signals found by the SPA analysis were also found by the F_{ST} methods and by Bayenv. However, some of the strong signals

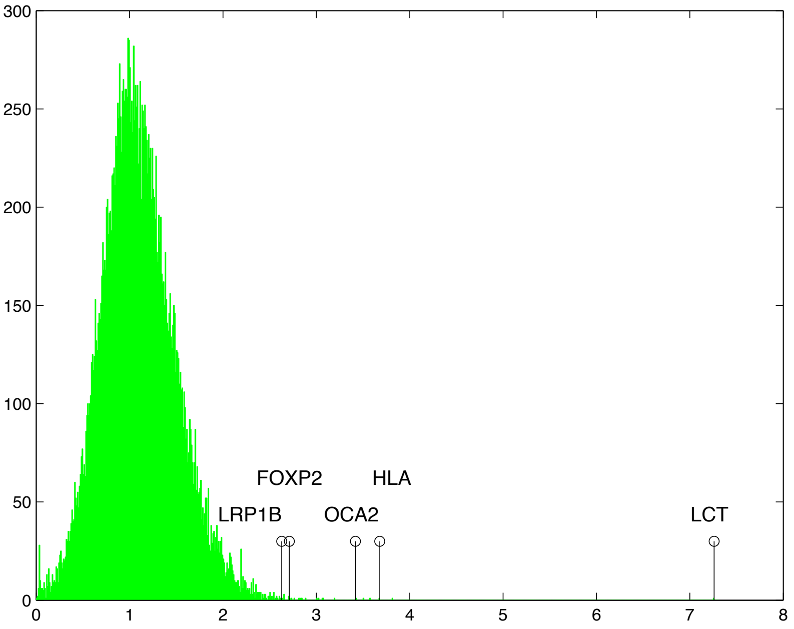
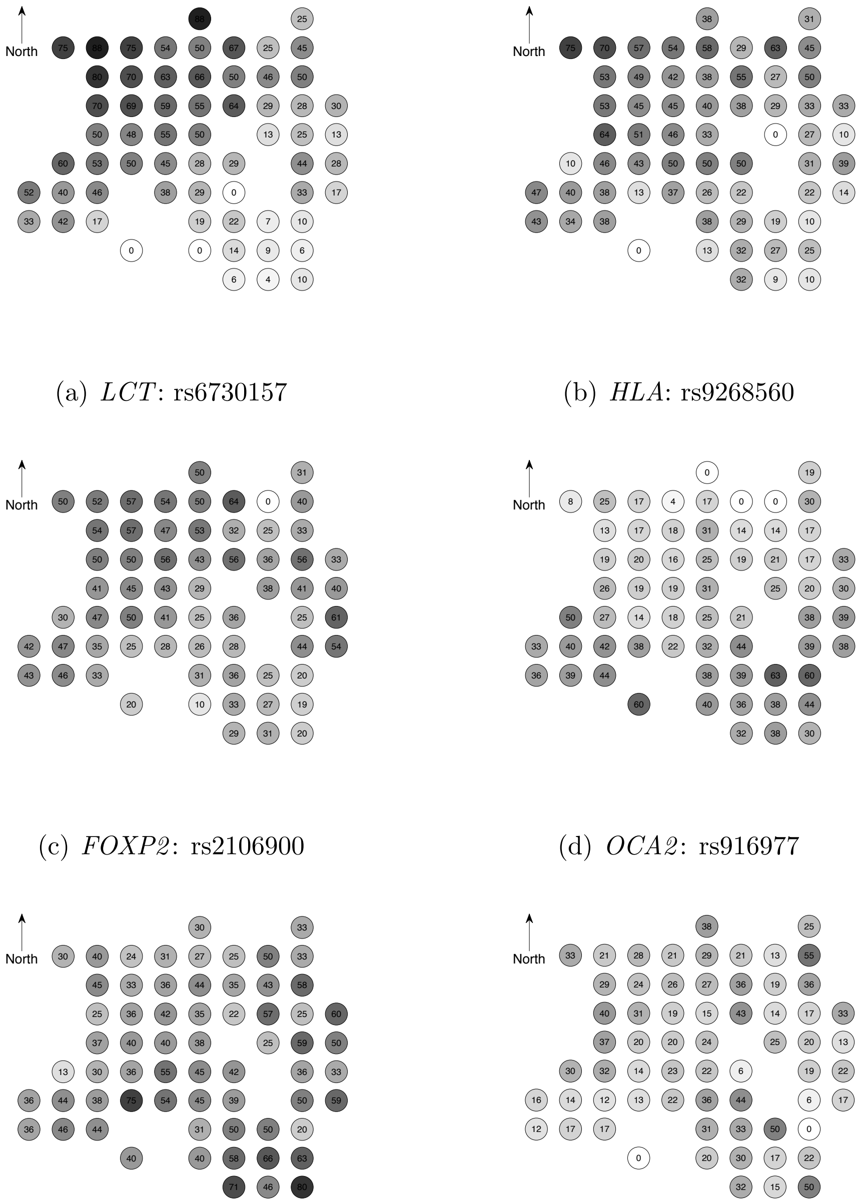


Figure 2.7: The distribution of SPA scores representing the allele frequency gradients. The marked positions correspond to genes discussed in the text.

that are found by our analysis are found by iHS and are not found by F_{ST} or by Bayenv, suggesting that our method captures loci that some regions with iHS signals that are outliers with respect to their allele frequency gradients (our method) but not with respect to overall allele frequency variation (as detected by F_{ST} or Bayenv [CWD10]). In addition, there are, as expected, signals that are found using SPA, but not using iHS (See Table 2.5). We note that the SNPs found by F_{st} and not by other methods are mostly rare SNPs with one or two occurrences of the minor allele in the data.

Importantly, we observe that the F_{ST} and the Bayenv scores are sensitive to the definition of the partition of individuals into populations. Particularly, defining populations based on country of origin leads to a different set of genes compared to the case where the populations are defined based on general geographic regions. In contrast, the analysis performed using SPA is oblivious to a partition of the



(a) *LCT*: rs6730157

(b) *HLA*: rs9268560

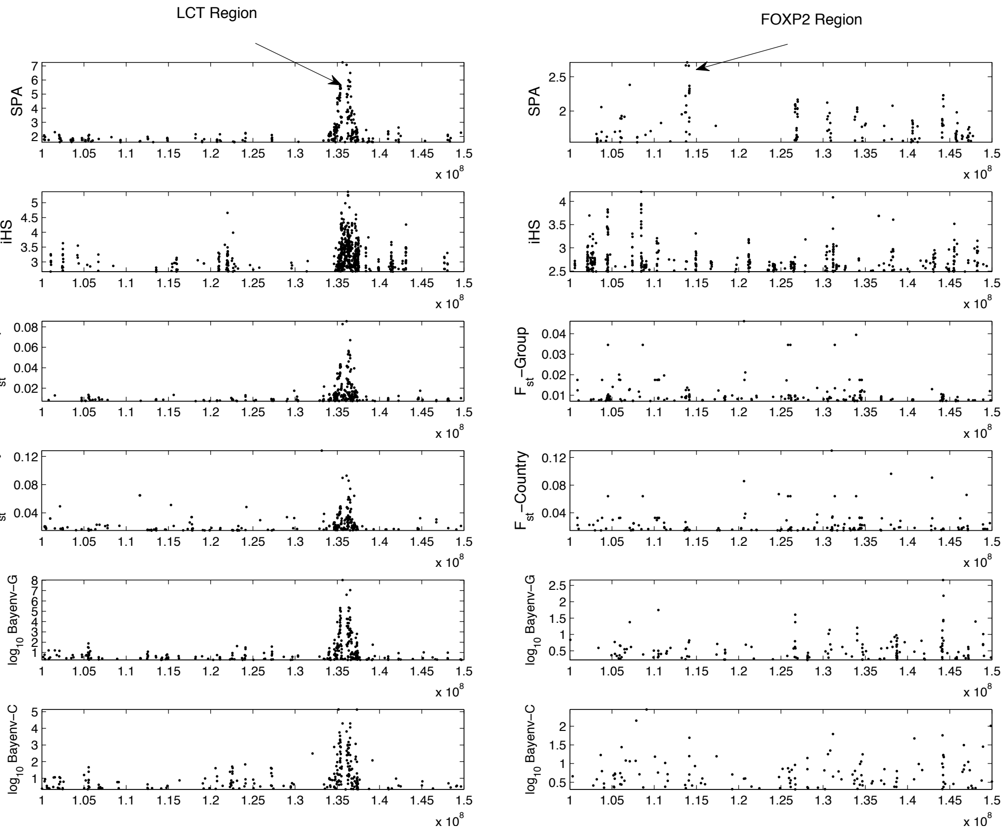
(c) *FOXP2*: rs2106900

(d) *OCA2*: rs916977

(e) *LRP1B*: rs7598314

(f) Typical SNP

Figure 2.8: Spatial distribution of SNPs with extreme allele frequency gradients. The grey scale stands for allele frequency: dark for high frequency and white for low frequency. We divide the whole map into 10 × 10 grid. We then calculate the allele frequency for each small region by averaging all individuals in the region. Regions with less than 5 individuals are removed for accurate allele frequency estimation.



(a) Chromosome 2

(b) Chromosome 7

Figure 2.9: Selection results of six methods in two chromosomes. The SPA, F_{ST} and Bayenv are run over POPRES data set, and the iHS is obtained in [VKW06] using HapMap Europe data. The plot is for 2% of POPRES SNPs and 1% of HapMap SNPs.

Table 2.4: Correlation coefficient between six methods. We check whether a SNP with top 2% scores for POPRES set or 1% scores for HapMap set in 100kb window to determine whether this window is under selection. Then we have a binary vector across the whole genome for all four methods, and we compute the correlation coefficient using the binary vector

	iHS	F_{ST}-Grp	F_{ST}-Cntry	Bayenv-Grp	Bayenv-Cntry	SPA
iHS	1.000000					
F_{ST}-Grp	0.028713	1.000000				
F_{ST}-Cntry	0.013455	0.401576	1.000000			
Bayenv-Grp	0.018016	0.305782	0.123768	1.000000		
Bayenv-Cntry	0.019460	0.203627	0.103108	0.333595	1.000000	
SPA	0.023986	0.333148	0.123805	0.446488	0.296796	1.000000

individuals into populations, since the approach treats ancestry as a continuous variable, and not as a categorial variable.

Table 2.5 in the supplementary materials provides a list of genes that are detected by our method but are not detected by iHS, F_{ST}, or Bayenv. A full list of loci with extreme frequency gradients from the SPA analysis can be found in the supplementary materials, Table 2.6. Among the most extreme gradients are the *HLA*, *LCT*, and *OCA2* regions, which are widely known to have undergone recent positive selection and show differentiation among populations. Interestingly, our analysis also indicates an extreme gradient for a SNPs in the *FOXP2* gene; *FOXP2* is associated with speech and implicated to have had important amino acid changes in early human evolution[EPF02]. In addition, *LRP1B* [LML00], a gene associated with lipid function with tumor relevance is found to have an extreme allele frequency gradient. The above are a few examples out of a longer list of genes that our method highlights as having strong gradients in allele frequencies across space (see Table 2.6).

2.4 Discussion

In this paper we present spatial ancestry analysis (SPA), a novel method for modeling spatial structure of genetic variation. Unlike previous methods which use principal component analysis to model spatial structure, our approach explicitly models the allele frequency in space and utilizes this model to place individuals in a two dimensional map or three dimensional sphere. We show that our method for localization of samples in space is slightly more accurate than principal component analysis, and importantly, unlike principle component analysis, it can be used to localize individuals of mixed ancestry in space.

Accurate spatial localization of individuals based on genetic data is important in many applications in genetics, including population stratification in genome-wide association studies, admixture mapping, and personalized genomics. We demonstrate that a model-based approach has further applications, since it characterizes the spatial behavior of each of the SNPs separately. Particularly, we demonstrate that the modeling can be used for identifying SNPs with rapidly changing allele frequencies.

We note that our proposed model for slope functions is only one natural choice for such a model, and there may be other natural choices. The fact that our algorithm converges to a map which is highly similar to the map of Europe suggests that this choice is sensible, but not necessary optimal. As a future research direction, we argue that further exploration of other choices of slope functions may potentially provide better characterization of each SNP's spatial behavior, yielding a better localization of samples to space and ability to identify SNPs with unique and interesting spatial distributions.

Table 2.5: Genes with extreme gradients detected only by SPA.

Genes	SNP with extreme gradient	highest SPA score	Position
MGLL	rs782437	2.88928	chr3:128899892
SCAND3,ZNF192,ZSCAN	rs6903535	2.74179	chr6:28525201
FOXP2	rs2106900	2.71077	chr7:113909742
TMEM104	rs2385067	2.57703	chr17:70321665
ENY2,NUDCD1	rs1380098	2.54717	chr8:110317808
TAS2R,PRH1,PRR4	rs2597996	2.54539	chr12:11099651
MIR1244,BCL2L14	rs4763782	2.51019	chr12:12146023
ITPR1	rs7637793	2.49356	chr3:4725558
NUP153	rs11753865	2.48271	chr6:17790701
ZAP70	rs6736735	2.48015	chr2:97702865
TMEM117	rs2407790	2.46361	chr12:42717760
ERC1	rs11061714	2.44995	chr12:1348831
SEMA6D	rs281297	2.44008	chr15:45472796
PTK2B,DPYSL2,TRIM35	rs6557991	2.42558	chr8:27231225
SLC45A1	rs1466654	2.42438	chr1:8298500
SLC24A3	rs4814838	2.40770	chr20:19267846
SOX6	rs7118395	2.40590	chr11:16155641
SPOCK1	rs2348605	2.39459	chr5:136838003
ADAMTSL3	rs7169595	2.38401	chr15:82196578
WWOX	rs441004	2.38284	chr16:77794331
SLC26A4,LOC286002	rs11769313	2.38096	chr7:107101183
NRXN3	rs11625485	2.36487	chr14:79311885
ZNF19	rs2288486	2.35797	chr16:70070535
SEMA3E	rs215302	2.35116	chr7:83051200
ZDHHC2	rs2959634	2.35056	chr8:17071223
CDH7	rs7237421	2.35029	chr18:61586240
KCNIP4	rs7689421	2.33900	chr4:20364338
AK5	rs11162351	2.33449	chr1:77717320
ZAK	rs3754744	2.33022	chr2:173710310
IDH2	rs12443387	2.32722	chr15:88453860
USH2A	rs2677112	2.32663	chr1:213891790
FLJ22536	rs2078482	2.32446	chr6:22105905
ERC2	rs7628951	2.32417	chr3:56432745
PRRX1	rs593479	2.32273	chr1:168909523
RBFOX3	rs4790055	2.31235	chr17:75001008
FUT11,SEC24C	rs3849969	2.30577	chr10:75196005
GRRP1	rs1335759	2.28446	chr1:26344330
SORBS1	rs526928	2.28296	chr10:97324281
NEBL	rs1340293	2.28250	chr10:21239171
HUNK	rs2833609	2.27600	chr21:32303489
TMEM170B	rs9469574	2.27332	chr6:11698126

Table 2.6: A full list of regions with extreme gradients detected by SPA. The 450 (0.1% of total) SNPs with the highest scores are listed. Results of different SNPs are merged if the two SNPs are at distance smaller than 1MB.

Genes	SNP with most extreme gradient	highest SPA score	Position
LCT region	rs6730157	7.25764	chr2:135623558
HLA-DPB1 region	rs9268560	3.67652	chr6:32497490
HLA-B region	rs2517510	3.50633	chr6:31138101
HERC2,OCA2	rs916977	3.42020	chr15:26186959
ADH1C	rs1789903	3.07810	chr4:100481064
LOC283177	rs4592433	3.06241	chr11:133853131
FAM114A1,TLR10,TLR1	rs6835514	3.02649	chr4:38570775
MGLL	rs782437	2.88928	chr3:128899892
PLA2R1,LY75, LY75-CD302,ITGB6	rs16844715	2.83868	chr2:160623352
BNC2	rs10756762	2.81414	chr9:16553123
HPS5,GTF2H1	rs4150581	2.76205	chr11:18313846
SCAND3,ZNF192, ZSCAN16,ZSCAN23	rs6903535	2.74179	chr6:28525201
TWSG1	rs8091539	2.73925	chr18:9398889
FOXP2	rs2106900	2.71077	chr7:113909742
SCN2A,SCN1A	rs1461197	2.67766	chr2:166632718
RFPL1,RFPL1S	rs5763240	2.65562	chr22:28166926
LRP1B	rs7598314	2.63123	chr2:142250435
SYT1	rs7308297	2.62652	chr12:78301975
TMEM104	rs2385067	2.57703	chr17:70321665
ZFAND3	rs10485029	2.56609	chr6:37907682
ENY2,NUDCD1	rs1380098	2.54717	chr8:110317808
TAS2R50,TAS2R19, TAS2R31,TAS2R20, PRH1,PRR4,TAS2R46	rs2597996	2.54539	chr12:11099651
MIR1244,BCL2L14,LRP6	rs4763782	2.51019	chr12:12146023
SUCLG2	rs1352657	2.50426	chr3:67535708
CHMP1A,DPEP1,C16orf55	rs164749	2.50165	chr16:88235725
ITPR1	rs7637793	2.49356	chr3:4725558
NUP153	rs11753865	2.48271	chr6:17790701
ZAP70	rs6736735	2.48015	chr2:97702865
EHBP1,OTX1	rs11125946	2.47946	chr2:63151654
TMEM117	rs2407790	2.46361	chr12:42717760
RBFOX1	rs11645481	2.45097	chr16:7021094

Continued on Next Page. . .

Genes	SNP with	highest SPA score	Position
	most extreme gradient		
ERC1	rs11061714	2.44995	chr12:1348831
ACOT6,ACOT4	rs4903128	2.44437	chr14:73145688
SEMA6D	rs281297	2.44008	chr15:45472796
PTK2B,DPYSL2,TRIM35	rs6557991	2.42558	chr8:27231225
LOC729234	rs1917890	2.42547	chr2:96035728
SLC45A1	rs1466654	2.42438	chr1:8298500
SLC24A3	rs4814838	2.40770	chr20:19267846
SOX6	rs7118395	2.40590	chr11:16155641
LOC732275	rs8051237	2.40286	chr16:84939020
LRRFIP1	rs6754972	2.39608	chr2:238206505
SPOCK1	rs2348605	2.39459	chr5:136838003
ADAMTSL3	rs7169595	2.38401	chr15:82196578
WWOX	rs441004	2.38284	chr16:77794331
SLC26A4,LOC286002	rs11769313	2.38096	chr7:107101183
EMILIN2	rs592120	2.38076	chr18:2875118
VAV3	rs10494081	2.37404	chr1:108203626
NRXN3	rs11625485	2.36487	chr14:79311885
ZNF19	rs2288486	2.35797	chr16:70070535
SEMA3E	rs215302	2.35116	chr7:83051200
ZDHHC2	rs2959634	2.35056	chr8:17071223
CDH7	rs7237421	2.35029	chr18:61586240
FOXN3	rs1952182	2.34064	chr14:88978953
KCNIP4	rs7689421	2.33900	chr4:20364338
AK5	rs11162351	2.33449	chr1:77717320
ZAK	rs3754744	2.33022	chr2:173710310
IDH2	rs12443387	2.32722	chr15:88453860
USH2A	rs2677112	2.32663	chr1:213891790
FLJ22536	rs2078482	2.32446	chr6:22105905
ERC2	rs7628951	2.32417	chr3:56432745
ITGAX	rs1106398	2.32299	chr16:31277953
PRRX1	rs593479	2.32273	chr1:168909523
RBFOX3	rs4790055	2.31235	chr17:75001008
VAT1L	rs33967759	2.30990	chr16:76546435
FUT11,SEC24C	rs3849969	2.30577	chr10:75196005
UGT2B11	rs6817250	2.28873	chr4:70123190
GRRP1	rs1335759	2.28446	chr1:26344330
SORBS1	rs526928	2.28296	chr10:97324281
NEBL	rs1340293	2.28250	chr10:21239171

Continued on Next Page...

Genes	SNP with most extreme gradient	highest SPA score	Position
LOC100188947	rs12246543	2.27656	chr10:93291408
HUNK	rs2833609	2.27600	chr21:32303489
TMEM170B	rs9469574	2.27332	chr6:11698126

CHAPTER 3

Spatial Localization of Admixed Individuals

3.1 Motivations

Inference of ancestry from genetic data is a critical aspect of genetic studies with applications in mapping genes to diseases and in the inference of population history from genetic data [PZR10, SPP11]. Although the initial large scale genetic studies have focused primarily on homogeneous populations (e.g. Europeans), in an attempt to capitalize on genetic diversity, more recent studies focus on individuals of mixed ancestry (i.e. emerging from the mixing of genetically diverged ancestors) [JSS12, HTP11, WKV11, BVK10, NCP11, PCL13]. Such studies rely on accurate and unbiased ancestry inference both at a genome-wide level as well as at each locus in the genome [SPP11, PST13].

Traditional ancestry inference from genetic data has been focused on modeling populations as discrete sources. As a result, traditional genome-wide ancestry inference estimates the proportion of sites in the genome from a set of source populations (continental or subcontinental), while locus-specific inference aims to assign each allele in the genome to one of the considered populations [FSP03, PSD00, BPS12, ANL09, PTP09, PST13]. More recent approaches model population structure in a geographic continuum capitalizing on the correlation of genetics and geography expected in isolation by distance models [PPP06, YNE12, BQC13, WMB07]. This has been usually performed through principal components analysis [NJB08, PPP06], a general procedure for reducing the dimensionality of the data, with more recent approaches focusing on explicitly modeling the relationship between patterns of genetic variation and geography [YNE12, BQC13, WMB07]. Such approaches typically make the assumption that an individual's genotype is drawn from the genetic variation present at a single geographic location. This assumption is clearly violated when individuals have ancestors from multiple geographic regions such as the recently admixed populations in the Americas (e.g. African-American) and more generally, individuals that have ancestry from multiple regions within the same continent (e.g. individuals with recent ancestors from multiple regions of Europe). As a first attempt to model individuals with mixed ancestry, the SPA approach [YNE12] included an extension which allowed for the limited scenario where an individual is a descendant of parents who are not themselves admixed, but are from different locations. While this an improvement over completely ignoring the possibility of mixing, SPA does not model the admixture process and is unable to handle the vast majority of admixed individuals (e.g. more than 1 generation admixture).

In this work, we introduce approaches for ancestry inference in recently admixed individuals in a geographic continuum within a model that flexibly handles admixture across varying number of generations and ancestries. We view admixed individuals as having recent ancestors from several locations on a genetic-geographical map. Then, we perform ancestry inference by simultaneously localizing on the map the recent ancestors of an admixed individual and partitioning the individual's genome into segments inherited from the same ancestor (locus-specific ancestry). We take advantage of the observation that if one allele is inherited from a specific ancestor, then most likely, the neighboring alleles are also inherited from the same ancestor. Specifically, we use a model-based framework for genetic variation in the geographical continuum and utilize hidden Markov modeling of the admixture process for the past few generations to segment the individual's genome into locus-specific ancestry. We propose efficient optimization algorithms that allow us to accurately predict the geographic location of the recent ancestry of an admixed individual. Furthermore, our approach is able to accurately estimate the ancestry at each locus in the genome, thus providing a localization on the geographical map of each allele in recently admixed individuals.

We validate our approach by localizing the recent ancestry of the POPRES individuals with self-reported ancestry from multiple locations in Europe. Our method is able to localize the grandparents of the admixed individuals of POPRES with an average of 470Km of their reported ancestry, ranging from 305Km for individuals with Swiss-French ancestry to 701Km for those with Spanish-Portuguese ancestry. We perform extensive simulations starting from the real genotype data of the POPRES study to show that the localization accuracy within Europe decreases with increased number of ancestors to localize (e.g. 639Km for 4 ancestries versus 550Km for 2 ancestries) and with the number of generations in the admixture (562Km for 2 ancestries 8 generations ago versus 550Km for 2 ancestries 4 generations ago). We investigate the relationship between distance among ancestors on a map and inference resolution and show that inference accuracy (at the genome-wide and locus-specific level) increases as ancestries become more distant on the geographical map. Finally, we provide an analysis of ancestry localization error across all pairs of countries in Europe as resource for community interested in subcontinental ancestry in Europe.

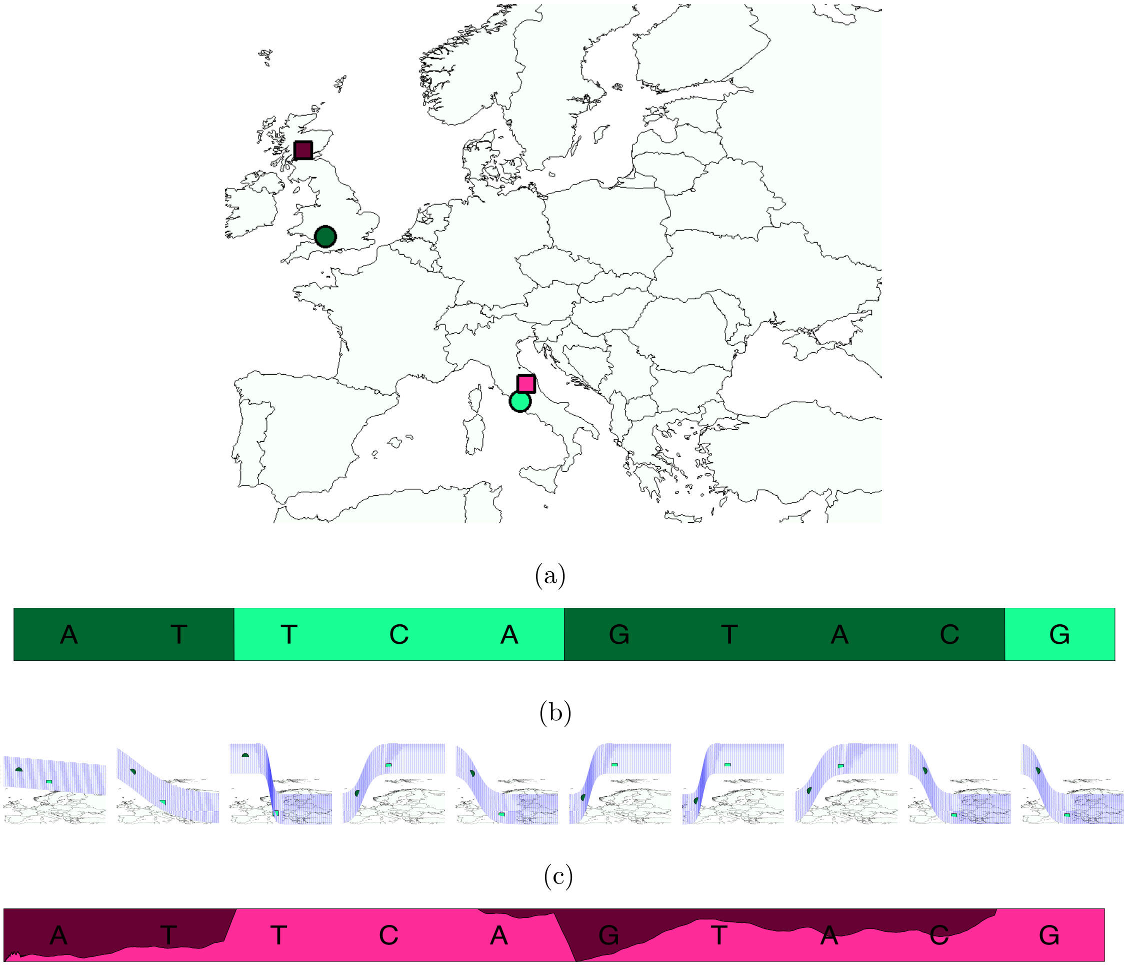
Methods

3.1.1 Overview of spatial localization for admixed individuals

We consider models of ancestry in admixed individuals in a geographical continuum. We view the mixed ancestry genome as being generated from several geographical locations on a map corresponding to the locations of their recent ancestors (see Figure 3.1). For example, consider the case of an individual with recent ancestry from Central Italy and South Great Britain (see Figure 3.1(a)). Its genome will be composed of segments originating from the two locations in Europe (see Figure 3.1(b)). In our framework, we model variation as function of geography at each position in the genome through a logistic gradient function (which can readily be inferred using public data [YNE12]). Each position in the genome has its own gradient that describes the degree of variation at that site as function of geography; for example some variants may have steep gradients while other variants may not vary at all with geographical locations (see Figure 3.1(c)). We extend standard hidden Markov models (HMM) for admixture to incorporate variation at each position on the map by allowing the emission probabilities to vary according to logistic gradients (see Methods). For example, each pair of locations on the map defines an HMM with emission probabilities at each position in the genome as function of logistic gradients. We perform inference in this model to find the ancestor locations on the map that maximize the likelihood of the observed mixed ancestry genome (Figure 3.1(a)). After finding the location of the recent ancestors, we assign each allele in the mixed genome to one of the ancestor location to provide a locus-specific ancestry call across the genome. Figure 3.1(d) shows an output of our locus-specific inference with locations in the admixed genome being labeled according to the inferred ancestral location on the map.

3.1.2 Spatial modeling of allele frequency

Although our base method for explicit modeling genetic variation as function of geography has been described elsewhere [YNE12, BQC13], we briefly present here the generative model. We view an individual's alleles as a Bernoulli draw from an allele frequency that changes across the map and we parametrize the allele frequency function through a logistic gradient as function of position (vec$x$ = (x_1, x_2)) in the map. Formally, the probability of observing the minor allele in



(a)

(b)

(c)

(d)

Figure 3.1: SPAMIX model for admixed individuals. (a) Example of haploid individual with two ancestry locations in Europe (circles denote the true ancestry locations). (b) The admixture process induces segments of different ancestry backgrounds. (c) SPAMIX uses logistic gradients to describe allele frequencies as function of geographic map to instantiate an admixture HMM for each pair of locations on a map. Each location on the map is associated to a particular allele frequency at all sites in the genome. (d) SPAMIX finds the location of ancestors on a map (denoted by squares in subfigure (a)) and the locus-specific ancestry at each site in the genome by maximizing the likelihood of genotype data.

SNP j at position vec$x$ on the map $f_j(\text{vec}x)$, is defined as:

$$f_j(\text{vec}x) = \frac{1}{1 + \exp(-\text{vec}a_j^T \text{vec}x - b_j)} \tag{3.1}$$

where veca_j and b_j are parameters specific to SNP j. We estimate veca_j and b_j from data containing individuals with known locations [YNE12] and then use these coefficients in the inference of ancestries of mixed individuals.

Although easy to manipulate mathematically, the logistic functions we employ here clearly do not capture all genetic variation (for example, variants that have multiple modes or peaks in the allele frequency surface, e.g. as may be typical of rare variants). However, these functions have been shown to capture general trends in common variant frequencies sufficiently well to produce highly accurate spatial assignment in individuals with non-mixed ancestry [YNE12]. We hypothesize that such simple-to-manipulate functions are sufficient for accurate localization of recent ancestors in individuals with mixed sub-continental ancestries.

3.1.3 Haploid data with admixed ancestry

Spatial model for admixed haploid data

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

$$L(h;X) = \sum_{Z} P(Z) \prod_{i=1}^{L} P(h_i|z_i;X) \tag{3.2}$$

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

$$\begin{aligned} P(Z) &= P(z_1) \prod_{i=1}^{L-1} P(z_{i+1}|z_i) \\ P(z_1 = j) &= 1/M \\ P(z_{i+1}|z_i) &= \begin{cases} (1-\tau_i) & z_{i+1} = z_i \\ \tau_i/(M-1) & z_{i+1} \neq z_i \end{cases} \end{aligned}$$

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

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

An illustration of the model is given in Figure 3.1. We note that our model makes the assumptions of independence of alleles conditional on local ancestry (no modeling of background LD) as well as the assumption of equally likely transition among ancestries when transition in ancestries occur along a haplotype.

Spatial ancestry inference for haploid data

Under the generative model above, spatial ancestry inference is reduced to inferring the M ancestral locations given data for an admixed haplotype, followed by posterior decoding in the HMM to obtain locus-specific predictions. This can be achieved by maximizing the likelihood function (3.2) with respect to X. By treating X as parameters and Z as hidden variables, this maximization falls within the procedure of the standard Expectation Maximization (EM) algorithm [DLR77]:

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

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

where α/β are standard forward/backward HMM functions and can be efficiently calculated.

M step: The maximization step optimizes the Q function:

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

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

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

We perform the maximization by taking advantage of the convex properties of the equation and using analytical forms for the Hessian of the function. The complete derivations are given below.

Expectation Maximization algorithm for haploid spatial ancestral inference

We would like to infer M ancestral location for a given mixed individual haplotype. This can be achieved by maximizing the likelihood function with respect to X as follows

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

By treating X as parameters and Z as hidden variables, this maximization falls in exactly the procedure of EM algorithm.

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

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

where α and β can be calculated recursively

$$\begin{aligned}
\alpha_1(j) &= (1/M)P(h_1|z_1 = j; X^{(t)}) \\
\alpha_i(j) &= \sum_{j'} \alpha_{i-1}(j')P(z_i = j|z_{i-1} = j')P(h_i|z_i = j; X^{(t)}) \\
\beta_L(j) &= 1 \\
\beta_i(j) &= \sum_{j'} P(z_{i+1} = j'|z_i = j)P(h_{i+1}|z_{i+1} = j'; X^{(t)})\beta_{i+1}(j')
\end{aligned}$$

M step. The maximization step needs to optimize the Q function, which can be done as follows

$$\begin{aligned}
& Q(X; X^{(t)}) \\
= & \sum_Z P(Z|h; X^{(t)}) \ln \left(P(Z) \prod_i P(h_i|z_i; X) \right) \\
= & \sum_j \left(\sum_i P(z_i = j|h; X^{(t)}) \ln P(h_i|z_i = j; x_j) \right) + \text{const.} \\
= & \sum_{i,j} C_{ij} \ln P(h_i|z_i = j; x_j) + \text{const.} \\
= & \sum_{i,j} C_{ij} q_i(x_j) + \text{const.} \qquad (3.3)
\end{aligned}$$

where C_{ij} denotes the constant $P(z_i = j|h, X^{(t)})$, and

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

We use Newton's method to perform the maximization step, which is a widely used optimization technique. The gradient for the Q function in (3.3) can be computed as follows

$$\frac{\partial Q}{\partial x_j} = \sum_i C_{ij} g_i(x_j)$$

where

$$g_i(x_j) = \begin{cases} \frac{1}{1+\exp(-a_i^T x_j - b_i)}(-a_i)^T & h_i = 0 \\ \frac{1}{1+\exp(a_i^T x_j + b_i)}(a_i)^T & h_i = 1 \end{cases}$$

The Hessian matrix for the Q function in (3.3) can be obtained as follows

$$\frac{\partial^2 Q}{\partial x_j^2} = \sum_i C_{ij} h_i(x_j)$$

where

$$h_i(x_j) = \frac{1}{1+\exp(-a_i^T x_j - b_i)} \cdot \frac{1}{1+\exp(a_i^T x_j + b_i)} \cdot (-a_i a_i^T)$$

Locus-specific spatial ancestral inference for haploid data

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

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

which can be efficiently solved by the Viterbi algorithm [Vit06]. In order to compute a posterior probability of each locus-specific ancestry, the forward backward algorithm can also be employed.

3.1.4 Diploid data with admixed ancestry

Spatial model for admixed diploid data

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

$$L(g; X, Y) = \sum_Z P(Z) \prod_{i=1}^{L} P(g_i | z_i^p, z_i^m; X, Y) \tag{3.4}$$

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

$$\begin{aligned}
P(Z) &= \left(P(z_1^p) \prod_{i=1}^{L-1} P(z_{i+1}^p | z_i^p)\right) \left(P(z_1^m) \prod_{i=1}^{L-1} P(z_{i+1}^m | z_i^m)\right) \\
P(z_1^p) &= 1/M \\
P(z_1^m) &= 1/N \\
P(z_{i+1}^p | z_i^p) &= \begin{cases} (1-\tau_i) & z_{i+1}^p = z_i^p \\ \tau_i/(M-1) & z_{i+1}^p \neq z_i^p \end{cases} \\
P(z_{i+1}^m | z_i^m) &= \begin{cases} (1-\tau_i) & z_{i+1}^m = z_i^m \\ \tau_i/(N-1) & z_{i+1}^m \neq z_i^m \end{cases}
\end{aligned}$$

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

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

The function f_i is the allele frequency function in logistic form (3.1). The probability $P(Z)$ models the recombination events in paternal and maternal ancestries, and the probability $P(g_i|z_i^p, z_i^m; X, Y)$ models the probability of generating the genotype from two ancestral geographical locations.

Spatial ancestry inference for diploid data

We would like to infer $M + N$ ancestral locations for a given mixed individual genotype. This can be achieved by maximizing the likelihood function (3.4) with respect to X and Y, which, analogous to the haploid case, can be performed using the EM algorithm [DLR77]:

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

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

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

M step: The maximization step optimizes the Q function:

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

where C_{ijk} denotes the posterior $P(z_i^p = j, z_i^m = k|g, X^{(t)}, Y^{(t)})$ computed from E

step, and the shorthand $q_i(x_j, y_k)$ is defined as:

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

Again, we leverage the convexity of the function and analytical forms for the Hessian to efficiently optimize the Q function. The complete derivations and optimization details are given below.

Expectation Maximization algorithm for diploid spatial ancestral inference

We would like to infer $M + N$ ancestral location for a given mixed individual genotype. This can be achieved by maximizing the likelihood function with respect to X and Y as follows

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

By treating X and Y as parameters and Z as hidden variables, this maximization falls in exactly the procedure of EM algorithm.

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

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

where α and β can be calculated recursively

$$\begin{aligned}
\alpha_1(j,k) &= 1/(MN)P(g_1|z_1^p = j, z_1^m = k) \\
\alpha_i(j,k) &= \sum_{j',k'} \alpha_{i-1}(j',k')P(z_i^p = j|z_{i-1}^p = j')P(z_i^m = k|z_{i-1}^m = k')P(g_i|z_i^p = j, z_i^m = k) \\
\beta_L(j,k) &= 1 \\
\beta_i(j,k) &= \sum_{j',k'} P(z_{i+1}^p = j'|z_i^p = j)P(z_{i+1}^m = k'|z_i^m = k)P(g_{i+1}|z_{i+1}^p = j', z_{i+1}^m = k')\beta_{i+1}(j',k')
\end{aligned}$$

M step. The maximization step needs to optimize the Q function, which can be done as follows

$$\begin{aligned}
& Q(X,Y;X^{(t)},Y^{(t)})) \\
= & \sum_{Z^p,Z^m} P(Z^p,Z^m|g;X^{(t)},Y^{(t)}) \ln \left(P(Z^p)P(Z^m) \prod_i P(g_i|z_i^p,z_i^m;X,Y) \right) \\
= & \sum_{j,k} \left(\sum_i P(z_i^p = j, z_i^m = k|g;X^{(t)},Y^{(t)})) \ln P(g_i|z_i^p = j, z_i^m = k; x_j, y_k) \right) + \text{const.} \\
= & \sum_{i,j,k} C_{ijk} \ln P(g_i|z_i^p = j, z_i^m = k; x_j, y_k) + \text{const.} \\
= & \sum_{i,j,k} C_{ijk} q_i(x_j, y_k) + \text{const.} \qquad (3.5)
\end{aligned}$$

where C_{ijk} denotes the constant $P(z_i^p = j, z_i^m = k|g, X^{(t)}, Y^{(t)})$, and

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

This function is not concave in general, since the function corresponding to heterozygous genotype $g_i = 1$ is not concave. But we can still use convex optimization techniques to get a local optimal solution. In practice, we observe that the function is concave almost all the time. Thus, this proposed algorithm can well converge to optimal solution.

Note that there is a subtle connection from the above EM algorithm to parental location inference algorithm given previously [YNE12]. For parental location inference, the hidden variables Z^p and Z^m would be fixed instead of random. Thus, the EM algorithm would reduced to the algorithm given previously, which is equivalent to one M-step in the above EM algorithm.

The gradient for the Q function in (3.5) can be computed as follows

$$\frac{\partial Q}{\partial x_j} = \sum_{i,k} C_{ijk} g_{ik}(x_j, y_k)$$

where

$$g_{ik}(x_j, y_k) = \begin{cases} -p_{ij}a_i & g_i = 0 \\ \dfrac{(1-2m_{ik})(1-p_{ij})p_{ij}}{p_{ij}(1-m_{ik}) + m_{ik}(1-p_{ij})} \cdot a_i & g_i = 1 \\ (1-p_{ij})a_i & g_i = 2 \end{cases}$$

The variables p_{ij} and m_{ik} are shorthands for the ith allele frequencies for paternal ancestry j and maternal ancestry k defined as

$$\begin{aligned} p_{ij} &= \frac{1}{1+\exp(-a_i^T x_j - b_i)} \\ m_{ik} &= \frac{1}{1+\exp(-a_i^T y_k - b_i)} \end{aligned}$$

The Hessian for the Q function in (3.5) can be computed as follows

$$\frac{\partial^2 Q}{\partial x_j^2} = \sum_{i,k} C_{ijk} h_{ik}(x_j, y_k)$$

where

$$h_{ik}(x_j, y_k) = \begin{cases} (1-p_{ij})p_{ij}(-a_i a_i^T) & g_i = 0 \\ (1-2m_{ik}) \dfrac{\dfrac{(1-m_{ik})p_{ij}}{1-p_{ij}} - \dfrac{m_{ik}(1-p_{ij})}{p_{ij}}}{\left(\dfrac{1-m_{ik}}{1-p_{ij}} + \dfrac{m_{ik}}{p_{ij}}\right)^2} (-a_i a_i^T) & g_i = 1 \\ (1-p_{ij})p_{ij}(-a_i a_i^T) & g_i = 2 \end{cases}$$

and

$$\begin{aligned}\frac{\partial^2 Q}{\partial x_j \partial y_k} &= \sum_i I(g_i = 1)\left[\frac{m_{ik}(1-m_{ik})(1-2m_{ik})p_{ij}(1-p_{ij})(1-2p_{ij})}{[(1-m_{ik})p_{ij}+(1-p_{ij})m_{ik}]^2}\right. \\ &\qquad \left. + \frac{2m_{ik}(1-m_{ik})p_{ij}(1-p_{ij})}{(1-m_{ik})p_{ij}+(1-p_{ij})m_{ik}}\right](-a_i a_i^T)\end{aligned}$$

Locus-specific spatial ancestral inference for diploid data

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

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

which can be efficiently solved by the Viterbi algorithm [Vit06]. The posterior of local ancestries for each allele can be obtained using a forward-backward algorithm following the *E step* in the algorithm.

Homogeneous paternal and maternal ancestries

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

3.1.5 POPRES data set

We applied our methods to a data set collected from European populations, which was assembled and genotyped as part of the larger POPRES project [NBK08] and accessed via dbGAP accession number phs000145.v4.p2. A total of 3192 European individuals were genotyped at $500,568$ loci using the Affymetrix 500K SNP chip. The same stringency criteria as in [NJB08] were applied to create the training data. We removed SNPs with low-quality scores and high missigness[NJB08]. We filtered individuals to avoid sampling individuals from outside of Europe, to create more even sample sizes across Europe, and to remove individuals whose self-reported data have grandparents with different origins. We note that this is the same data set used in [NJB08, BQC13]. For the remaining individuals who have observed grandparental data, we use that origin for the individual. Otherwise, we use the individual-level self-reported country of birth. As a result, we infer logistic gradients starting from genotype data from $447,245$ autosomal loci in $1,385$ individuals from 36 populations. 77.4% of SNPs are common SNPs (allele frequency > 0.05), and the rest 22.6% are rare SNPs. For testing, we identified an additional 470 individuals from the POPRES data that have self-reported grandparental ancestry from 2 or more countries in Europe. A summary of homogeneous ancestry individuals used in estimating logistic gradients (1,385) and with sub-continental European admixed ancestry (470) are given in Table 3.1. Although our approach models admixture LD, it assumes that markers are independent conditional on local ancestry. To account for LD, we performed LD pruning at a level of $r2 < 0.2$ (72,418 SNPs retained) and we adjusted the transition rates in our model by a factor of 10^{-2} to remove spurious short ancestry windows induced by residual LD (see Table 3.3 and 3.5 for results at different pruning levels).

3.1.6 Simulation setup

We use BEAGLE to phase the POPRES data then simulate offspring admixed individuals by modeling recombinations within the last couple of generations. The recombination probability between each SNPs is approximated as $(g-1)\phi(d_{i+1} - d_i)$ where d_i's are the locations of each SNPs in bp, g denotes the number of generations and ϕ is the probability of one recombination per generation per base-pair [PSK09]. As recombination map, we assumed a flat recombination rate of $\phi = 10^{-8}$ per base-pair. For given number of M paternal ancestries and N maternal ancestries, we randomly select from POPRES data set a set of $M + N$ individuals, each of which has 4 grandparents from the same locations and randomly select one haplotype from each individual. We simulate the admixed haplotypes independently for the maternal and paternal haplotypes using the standard Poission process of admixture block distribution [PTP09]. If specified as homogeneous paternal and maternal ancestries, we pick M instead of $M + N$ ancestries, and use the same M ancestries for both paternal and maternal haplotype simulation.

Table 3.1: Self-reported grandparental ancestry (location of origin) of the POPRES data individuals (1906 in total). For individuals with grandparental ancestry from 2 different countries, we also report the number of individuals with 2 grandparents from one location and 2 from the other (2/2), versus individuals with 3 grandparents from one country (3/1).

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

For SPAMIX haploid model, the simulated haplotypes are used as input directly. Also, we always use the correct number of ancestries M or N as input. For the SPAMIX diploid model, the combined genotype from two simulated paternal and maternal haplotypes are used as input. To avoid testing bias, we estimate the allele frequency logistic gradients each time using the POPRES individuals with the $M + N$ simulation ancestors excluded from the training set.

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

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

3.1.7 Web Resources

The software implementation of this method is freely available to public at `http://genetics.cs.ucla.edu/spamixG`

Table 3.2: Average distance between inferred and true ancestry locations in simulated admixed individuals from POPRES data. Simulations assume 4 generations in the mixture process. Naive model denotes the extension of SPA that ignores admixture-LD. SPAMIX (logistic) represents simulation results starting from haplotypes generated at a location on a map using a Bernoulli sampling from the logistic gradients (see Methods). Parenthesis denote the standard deviation while standard error of the mean is computed as standard deviation divided by square of number of simulations in each category.

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

3.2 Results

3.2.1 Performance of continuous ancestry inference for admixed individuals in simulations

We investigated the performance of our inference through simulations starting from real POPRES data [NBK08]. We randomly selected individuals with known locations, used them to simulate admixed individuals, and employed the remaining individuals as training data to infer the logistic gradients and perform inference (see Methods). We find that our approach is able to infer the ancestry locations for individuals with two recent ancestors in Europe to an average of 550 Km of the true ancestral locations (see Table 3.2). We observe a very large variance (334Km) across pairs of samples thus showing the high variability in performance across data. A potential cause for this effect is the denser sampling of individuals within the center of Europe combined with higher errors in fitting the logistic gradients in some regions of the map (e.g. boundaries).

Table 3.3: Average distance between inferred and true ancestry locations in simulated admixed individuals from POPRES data. Simulations assume 4 generations in the mixture process. Naive model denotes the extension of SPA that ignores admixture-LD. SPAMIX (logistic) represents simulation results starting from haplotypes generated at a location on a map using a Bernoulli sampling from the logistic gradients (see Methods). Parenthesis denote the standard deviations. We found that Linkage Disequilibrium (LD) effect significantly affects the ancestry inference as well as the local ancestry inference in unaccounted for. We observe more recombination events than expected if using the correct recombination probability (used in simulations). We circumvent this bias multiplying the transition probability by a factor 10^{-1}, 10^{-2}, 10^{-4} and 10^{-5} for the pruned SNP list with 0.1, 0.2, 0.5 and 0.8 pruning thresholds. 44,699, 72,418, 136,284, 194,432 SNPs were retained at the 4 pruning thresholds.

No. of ancestry		1	2	3	4
Independent SNP model	Pruned SNP (0.1)	425(252)	961(540)	977(599)	982(655)
	Pruned SNP (0.2)	443(265)	880(491)	898(530)	880(578)
	Pruned SNP (0.5)	420(245)	823(448)	855(502)	810(494)
	Pruned SNP (0.8)	421(259)	810(429)	845(491)	813(505)
SPAMIX	Pruned SNP (0.1)	425(252)	558(314)	596(353)	621(405)
	Pruned SNP (0.2)	443(265)	550(326)	591(367)	639(423)
	Pruned SNP (0.5)	420(245)	557(359)	630(522)	657(617)
	Pruned SNP (0.8)	421(259)	589(557)	809(895)	878(848)

To test how much is gained by explicit modeling of admixture LD (i.e. correlations among SNPs induced by segments of recent shared ancestry), we also inferred the recent ancestry location using a naive model that assumes all SNPs to be independent (as in [YNE12]). We observe a significant decrease in local ancestry prediction error (880 Km for Naive model versus 550 Km for SPAMIX), thus showing that modeling admixture LD significantly increases performance. We also quantified the effect of background LD (correlations among markers conditional on local ancestry) in our approach. We observe increased performance after LD pruning (see Tables 3.3 and 3.5); therefore, all results in the main text are obtained after LD pruning (r2<0.2, see Methods).

It is often the case that due to access to pedigree data, haplotypes can be determined with high accuracy. Therefore, we quantified the gain in accuracy arising from having access to accurately phased haplotype data (i.e. haploid data) as compared to diploid data. Table 3.2 shows that having access to accurate phasing significantly increases localization accuracy. For example, the haploid model is able to localize a pair of ancestries within 557Km of true simulated location as compared for the diploid model that localizes the 4 ancestries of its two haplotypes with an average of 639Km of its simulated locations. As expected, when phasing is ignored, higher accuracies are attained in localizing similar number of ancestry locations from diploid model as compared to haploid model. For example, when localizing two ancestors the haploid model attains an error of 557 Km as compared to 550 Km for the diploid model, increasing to 620 Km versus 591 Km for localizing three ancestors. This is due to the diploid model having access to more data (both haplotypes) to localize the same number of ancestries. Therefore, for fixed number of ancestry locations, it is better to use the whole genome and ignore the phasing.

An important parameter of our model is the number of generations; with more generations, more recombination events have the opportunity to shuffle ancestry across the genome thus reducing the average length of the ancestry segment. We observe a slight decrease in performance from 2 to 8 generations (548 to 562 Km) which we expect to continue as the number of generations increases (in the limit of extremely large number of generations, our model is equivalent to the naive model that does not model admixture-LD) (see Table 3.4).

Our framework models genetic variation as function of geography by imposing a logistic gradient to the generating functions (see Methods). That is, the fre-

Table 3.4: Average distance between inferred and true ancestry locations in simulated admixed individuals from POPRES data as function of number of generations in the mixture process. Two ancestral locations were assumed for this simulation. Parenthesis denote the standard deviation while standard error of the mean is computed as standard deviation divided by square of number of simulations in each category.

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

quency of a given variant is allowed to change in a given direction on a map only according to a parametrized logistic function. Although this approach has been shown to provide a good approximation of common variation leading to accurate ancestry inference, we hypothesize that the error in fitting logistic gradients to real data limits the methods accuracy. To assess this scenario, instead of using real individual's haplotype data, we simulated admixed haplotypes directly from the logistic gradients we inferred from POPRES data (see Methods). We observe a large increase in accuracy in this idealized scenario as compared to simulations from real haplotype data (e.g. 236 vs 550 Km for 2 ancestries 4 generations ago, Table 3.2), thus indicating that logistic gradients do not account for all the correlation between geography and genetic variation. This suggests that more complicated functions linking geography to genetics within our framework may yield further improvements (see Discussion).

We investigated the performance of our approach as we increase the number of ancestral locations ($M + N$) to estimate for a given admixed individual. For a fixed number of generations (4), we varied the number of ancestry locations to estimate. The parental inference is different from 2 ancestry inference, as the parental inference assumes that one haplotype is from paternal ancestry and one from maternal ancestry. However, the 2 ancestry inference assumes that both

Table 3.5: Average distance between inferred and true ancestry locations in simulated admixed individuals from POPRES data. Naive model denotes the extension of SPA that ignores admixture-LD. SPAMIX (logistic) represents simulation results starting from haplotypes generated at a location on a map using a Bernoulli sampling from the logistic gradients (see Methods). Parenthesis denote the standard deviations. We found that Linkage Disequilibrium (LD) effect significantly affects the ancestry inference as well as the local ancestry inference in unaccounted for. We observe more recombination events than expected if using the correct recombination probability (used in simulations). We circumvent this bias multiplying the transition probability by a factor 10^{-1}, 10^{-2}, 10^{-4} and 10^{-5} for the pruned SNP list with 0.1, 0.2, 0.5 and 0.8 pruning thresholds. 44,699, 72,418, 136,284, 194,432 SNPs were retained at the 4 pruning thresholds.

No. of generation		2	4	6	8
Independent SNP model	Pruned SNP (0.1)	995(550)	961(540)	974(539)	987(537)
	Pruned SNP (0.2)	899(487)	880(491)	864(466)	927(491)
	Pruned SNP (0.5)	809(444)	823(448)	819(436)	837(444)
	Pruned SNP (0.8)	834(441)	810(429)	812(442)	799(447)
SPAMIX	Pruned SNP (0.1)	549(318)	558(314)	567(334)	546(326)
	Pruned SNP (0.2)	548(329)	550(326)	541(295)	562(336)
	Pruned SNP (0.5)	551(390)	557(359)	590(371)	588(467)
	Pruned SNP (0.8)	580(478)	589(557)	634(576)	586(538)

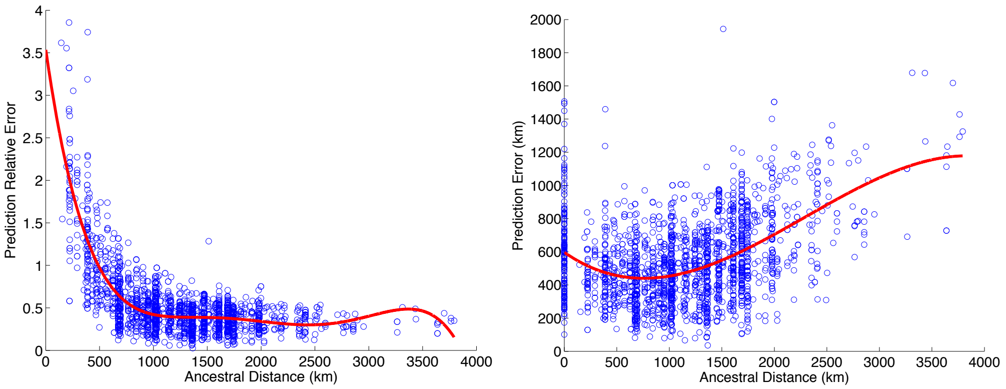
of the haplotypes are mosaic of two ancestries ($M = N = 2$) . As expected, we observe decreases in performance as the number of ancestries increases; this is due to the fact that the same genetic data is used to infer more and more locations (Table 3.2). For example, the average prediction error increases from 550 for 2 ancestries to 639 Km for 4 ancestral locations.

3.2.2 Increased distance between ancestral locations improves performance

It is well known that accuracy of ancestry inference correlates with genetic distance between ancestral populations. For example, discrete local ancestry can be inferred with very high degree of accuracy in mixtures of highly diverged populations (e.g. African Americans) as compared to closely related ones (e.g. subcontinental mixtures) [PSK09]. Since geography correlates with genetic distance, we hypothesized that continuous ancestry inference in recently admixed individuals also correlates with distance among ancestries on the map. Indeed, we observe that the relative prediction error decreases with the distance between ancestries in Europe (Figure 3.2(a)). For example, if the ancestries are 500 Km apart, we observe a relative prediction error of 0.75 as compared to 0.5 when the ancestries are located 2000 Km apart. Interestingly, when not normalizing for the distance between ancestries (Figure 3.2(b)), we observe that prediction error increases with increased accuracy. This shows that although the task of separating the ancestry locations becomes simpler, the localization accuracy becomes poorer (e.g. two ancestors located 500Km apart are localized within 450Km of their true locations, while two ancestors located 3000Km apart are localized within 1000Km of their true locations). This effect is presumably due to assignment errors in the locus-specific ancestry that have a much bigger impact if the ancestral locations are further apart. Although fewer locus-specific errors are being made with increased distance (see below), these errors introduce more noise in the ancestral localization due to their higher distance to true location.

3.2.3 Locus-specific inference

An advantage of our framework is that, in addition to identifying the most likely locations of the recent ancestry of admixed individuals, it can also provide an



(a) Prediction Error

(b) Prediction Relative Error

Figure 3.2: Ancestral location prediction error as a function of distance between ancestral locations in simulations over POPRES data. Left shows the prediction error normalized by the distance between the ancestral locations used in simulations and right plots the the prediction error. Simulations use the haploid model with 2 generations in the mixture.

assignment of each allele in the genome to each ancestry location. We observe that local ancestry prediction accuracy accuracy increases with the distance (Figure 3.3) from 55% of loci assigned accurately for very closely related ancestries (less than 500Km apart) to more than 70% for ancestries 2500Km apart (Figure 3.4). Similar to the ancestor localization, we observe that although the total number of assignment errors is reduced with increased distance, these errors have a bigger impact when averaging across all sites to compute the average allele localization error. Therefore, we observe that the average local ancestry prediction error is increased as the ancestral distance is increased.

3.2.4 Map of accuracy across Europe

We also investigated the variance in performance according to the ancestor's labeled origin ((i.e. typically to level of country). Figure 3.5 shows the prediction error for admixed individuals with ancestry from pairs of origins in Europe. In

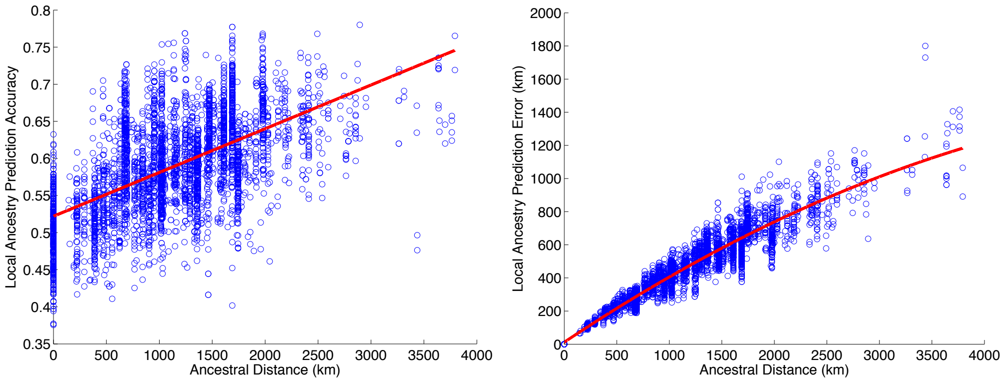
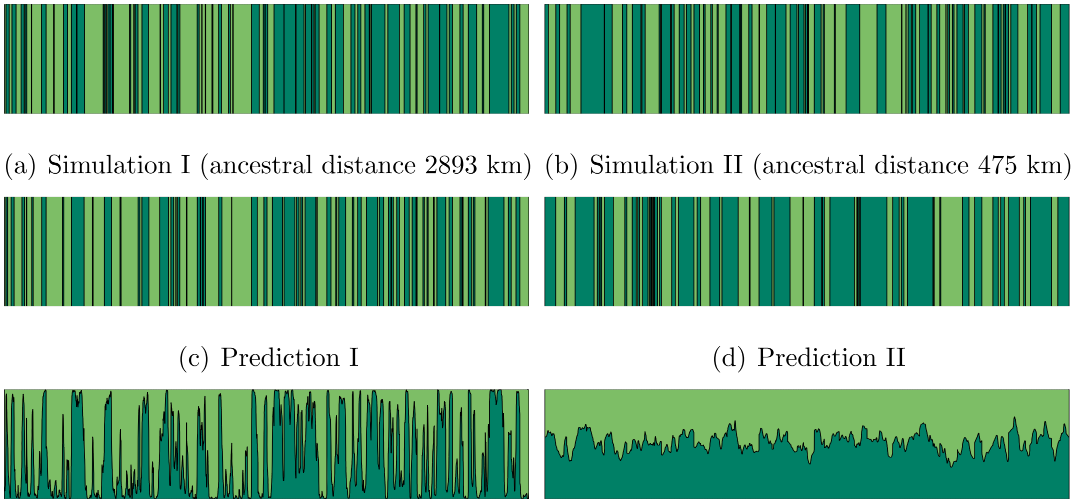


Figure 3.3: SPAMIX locus-specific ancestry prediction accuracy as function of distance between ancestral locations. Left displays the local ancestry prediction accuracy, defined as the percentage of all loci with correct assignment of ancestry. Right plot displays the average distance to true locations for each allele in the genome (local ancestry prediction error). Simulations use the haploid model with 2 generations in the mixture.



(a) Simulation I (ancestral distance 2893 km) (b) Simulation II (ancestral distance 475 km)

(c) Prediction I

(d) Prediction II

(e) Posterior I

(f) Posterior II

Figure 3.4: Example of local ancestry prediction results for distant and close ancestors.

general, we observe decreased performance for populations at the boundary of the European map (e.g., Portugal, Spain, Italy), and increased performance for subcontinental admixtures from populations located geographically in the center of Europe (e.g. France, Switzerland) (Figures 3.6 and 3.7). This can be an effect of biased sampling in the POPRES data, that sampled more individuals from Europe center, but also can be an effect of having more information to localize individuals in SPAMIX. In general we observe a prediction accuracy ranging from 411 Km for admixtures from Spain and Italy to 641 Km for individuals with recent ancestors from Spain and the UK.

3.2.5 Analysis real admixed individuals from POPRES data

Finally, we investigated whether high accuracies observed in simulations can also be attained in real data. Using SPAMIX, we localized the recent ancestry of all admixed European individuals from POPRES (see Methods). A total of 470 admixed individuals are analyzed using SPAMIX (see Table 3.1). As "ground truth" ancestral locations, we used the the center of the self reported grandparent location of origin. Therefore, we assume the mixed individuals from POPRES have 2 to 4 ancestry locations to infer. Across all 470 individuals, we observe a average prediction error distance is 426Km excluding outlier admixed individuals and 470Km including them; the outlier individuals were defined as those with prediction errors larger than $1,000$ Km and are reported in Table 3.6. We note that this error distance is lower than simulated experiments, but this is likely due to the dominating proportion of the admixed individuals of French and Swiss ancestries which can be accurately localized (average of 305 Km). As above, we note that SPAMIX ancestor localization performance varies greatly across Europe with ancestors from pairs of countries localized at the boundary of European map being harder to localize (e.g. an average of 701 Km for ancestor localization for Spanish Italian mixed individuals).

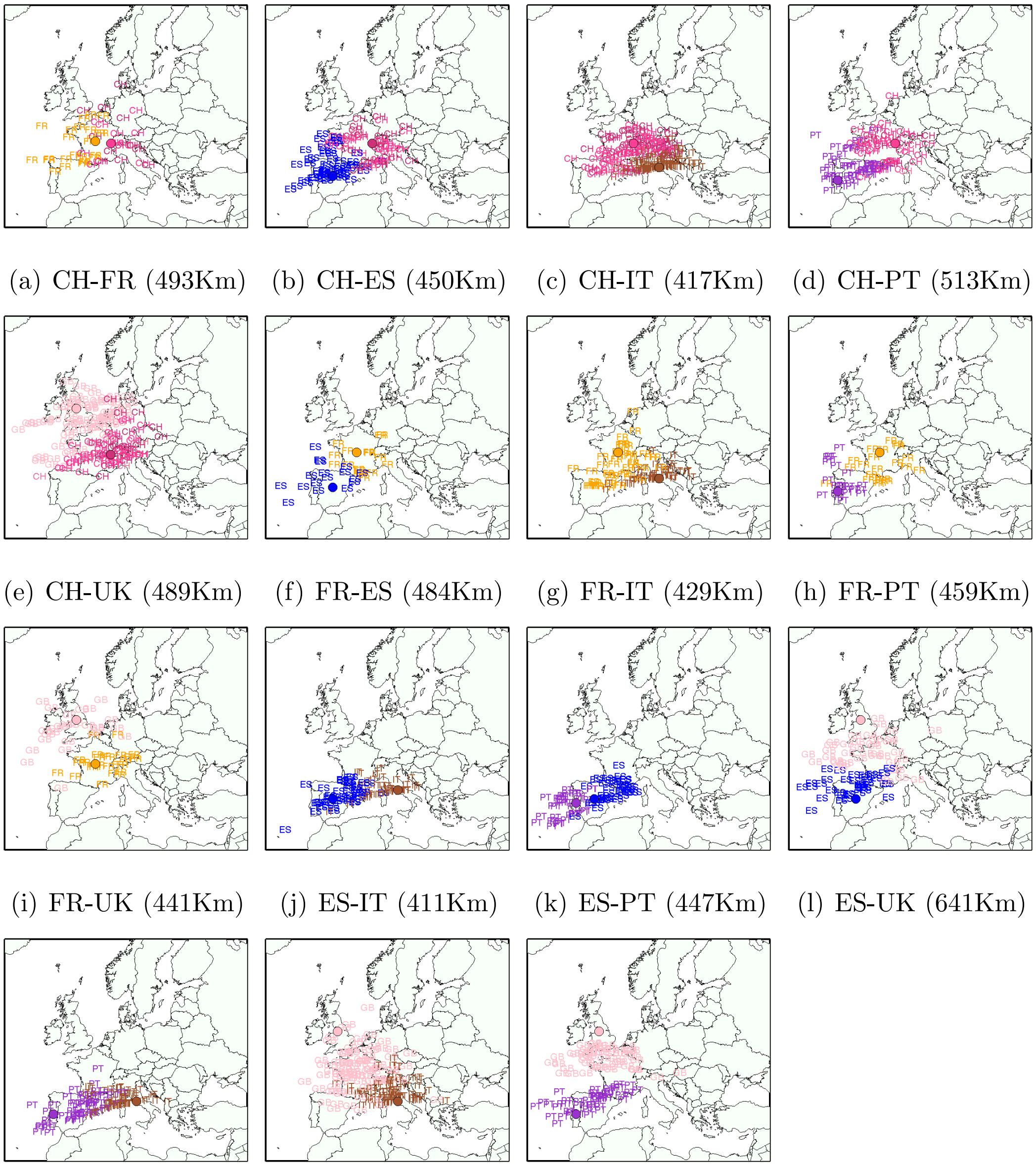
3.3 Discussion

We have introduced a novel method SPAMIX for predicting the geographical origins of multiple recent ancestors for individuals with recent mixed ancestry. Existing methods for ancestry inference in admixed population focus on discrete ancestry assignment and do not account for the continuous genetic variation within each continent. We introduce models that leverage the spatial structure of genetic variation using hidden Markov models for the admixture process to achieve high accuracy in localizing the recent ancestry of a given individual on a geographical map. We proposed computational efficient algorithms that enable us to infer the location on the map with great precision. Our proposed model can be viewed as a generalization of the parental localization model proposed in [YNE12] to account for admixture-LD while allowing for multiple generations and ancestries.

Although in our framework we use standard logistic gradient functions that were previously used to link geography and genetic variation, it is worth mentioning that such functions do not capture the whole variability observed in empirical data. To that extent, introducing more flexibility in these functions within the framework for admixture we described here are likely going to provide considerable improvements in accuracy with a tradeoff of computational time. We view this as a promising direction for future study. This is especially important for handling sequencing data, as rare variants rarely are fit well by the gradient functions (results not shown). Another area for further developments is extending the framework to model background LD (correlations among variants on the same ancestral backgrounds). Such LD adjustments have proved fruitful in improving localization accuracy for un-admixed individuals [BQC13] and are likely to improve inference for admixed individuals as well. Although we leave this for future work, one potential approach would be perform inference within short windows (to account for the local structure of LD) and merge the information within each window into the overall likelihood.

We also note that our approach uses a simplified model of ancestry switching along chromosomes that ignores the pedigree configuration of the recent ancestries; future work that explicitly considers the pedigree structure would have more structured ancestry transition matrices and could allow one to address questions regarding the specific configuration of ancestries. For example, for a mixed individual with recent Italian and British ancestry, we could ask whether the 4 grandparents were admixed from the 2 ancestries, or that whether the 4 grandparents were two Italians and two British. This question could be investigated by assigning local ancestry followed by analyzing the length distribution of the ancestry blocks and we leave this as future work.

Finally, we note that throughout this work we assumed that the number of different ancestries is known for a given admixed individual. Although this prohibits the direct application of our method for individuals where the number of different ancestries is unknown, in principle, model selection (balancing overall likelihood with number of parameters) could be employed to select the number of ancestries for individuals with unknown number of ancestors.



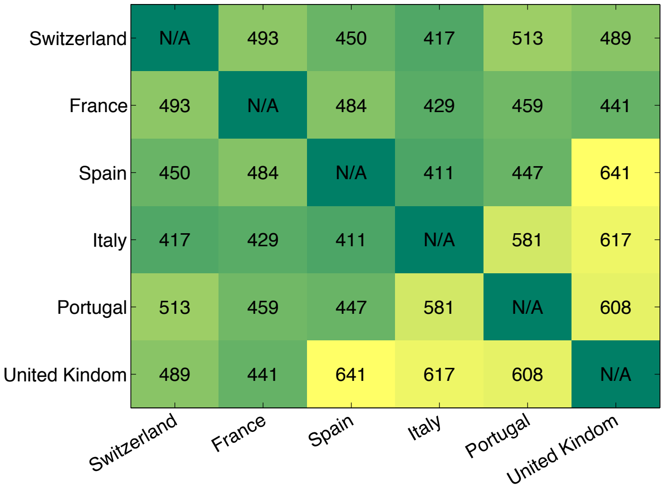
(a) CH-FR (493Km) (b) CH-ES (450Km) (c) CH-IT (417Km) (d) CH-PT (513Km)

(e) CH-UK (489Km) (f) FR-ES (484Km) (g) FR-IT (429Km) (h) FR-PT (459Km)

(i) FR-UK (441Km) (j) ES-IT (411Km) (k) ES-PT (447Km) (l) ES-UK (641Km)

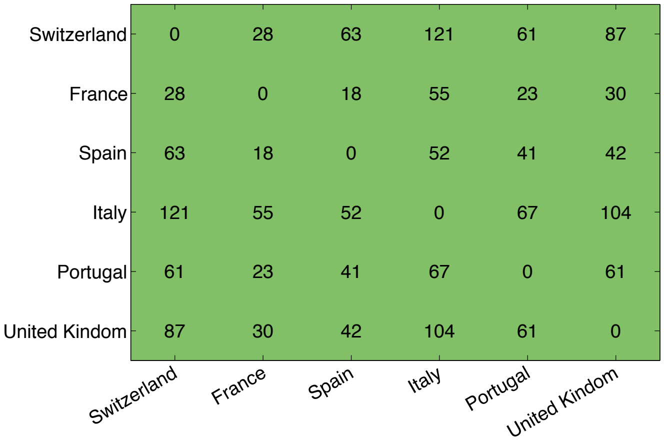
(m) IT-PT (581Km) (n) IT-UK (617Km) (o) PT-UK (608Km)

Figure 3.5: Ancestral location prediction error in simulations of European individuals with ancestry from two locations in Europe, stratified by the country of origin of each location (the country of origin is displayed in different colors). The assumed true locations are displayed by shaded circles. Results in parenthesis denote the average ancestral location prediction error across all simulations. In each simulation the reference data (used to estimate logistic gradients) is disjoint from data used to simulate admixed genomes (see Methods). The admixed genome is simulated as 4 generations ago, and SPAMIX diploid model is used for the inference. The number of simulated pairs can be found in Figure 3.7.



	Switzerland	France	Spain	Italy	Portugal	United Kindom
Switzerland	N/A	493	450	417	513	489
France	493	N/A	484	429	459	441
Spain	450	484	N/A	411	447	641
Italy	417	429	411	N/A	581	617
Portugal	513	459	447	581	N/A	608
United Kindom	489	441	641	617	608	N/A

Figure 3.6: Average Prediction error (Km) for six country pairs with largest populations.

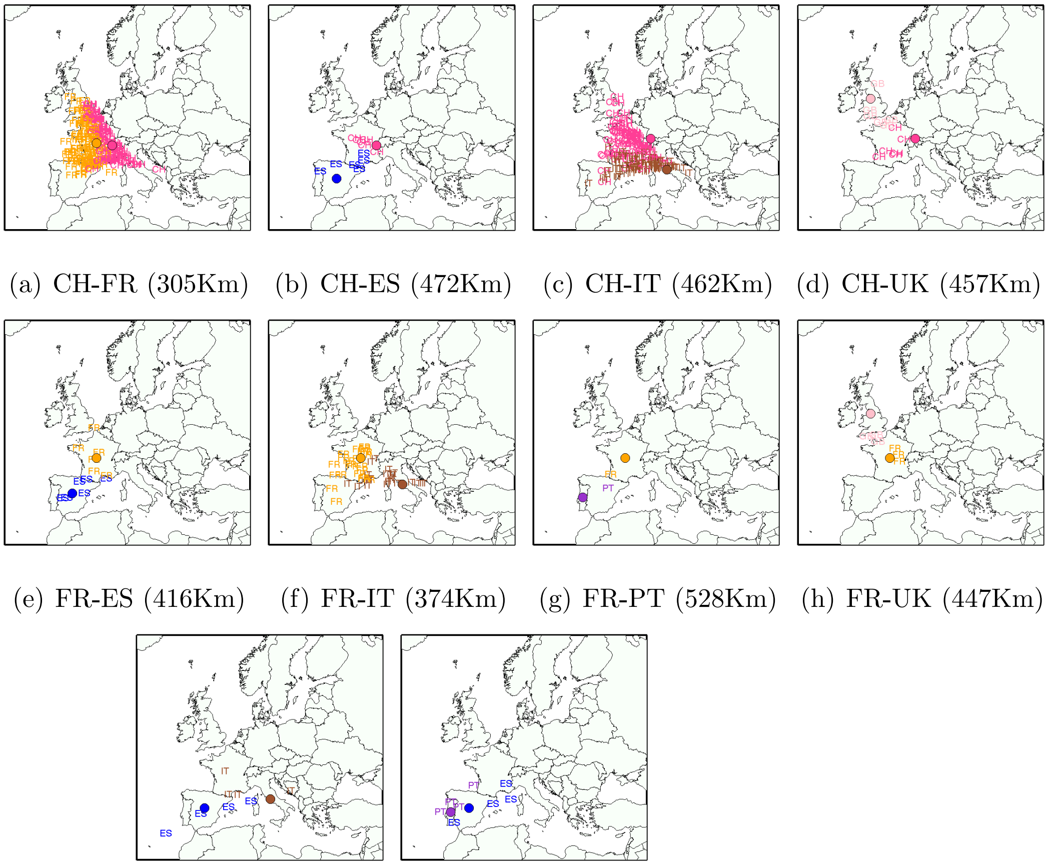


	Switzerland	France	Spain	Italy	Portugal	United Kindom
Switzerland	0	28	63	121	61	87
France	28	0	18	55	23	30
Spain	63	18	0	52	41	42
Italy	121	55	52	0	67	104
Portugal	61	23	41	67	0	61
United Kindom	87	30	42	104	61	0

Figure 3.7: Number for simulations for six country pairs with largest populations.

Table 3.6: POPRES admixed individuals with ancestral predictions inconsistent with their self-reported ancestries.

POPRES ID	self	father	PGF	PGM	mother	MGF	MGM	Pred. Locations
4183	Austria	Austria	Austria	Czech Republic	Poland	Russia	Switzerland	(52.26 12.61),(43.78 -9.81),(41.63 16.61),(52.44 13.30)
28710	France	Poland	Germany	Poland	France	France	France	(39.47 5.82),(37.97 14.65),(40.99 15.31)
24943	Sweden	Sweden	Sweden	Sweden	Finland	Russia	France	(48.03 6.91),(48.09 6.67),(57.35 8.79)
20086	France	France	France	France	France	Switzerland	France	(35.62 13.04),(43.65 16.14)
5550	Germany	Switzerland	Switzerland	Switzerland	Germany	Germany	Germany	(46.45 23.86),(55.66 -7.40)
47799	Germany	Germany	Russia	Germany	Germany	Germany	Germany	(50.20 14.77),(53.02 3.44)
32002	France	France	France	France	Turkey	Turkey	France	(35.85 10.05),(40.75 13.59)
27995	Poland	Poland	Poland	Poland	Russia	Russia	Poland	(55.65 12.49),(47.67 10.56)
38489	Russia	Russia	Germany	Germany	Switzerland	Switzerland	Russia	(46.59 5.62),(46.67 7.38),(49.86 -0.22)
7251	Austria	Switzerland	Switzerland	Switzerland	Austria	Austria	Austria	(38.64 -6.89),(62.19 30.99)
17323	Switzerland	Russia	Russia	Russia	Switzerland	Switzerland	Switzerland	(50.99 -1.41),(46.35 10.78)
20046	France	Russia	Russia	Russia	Poland	Poland	Poland	(37.25 14.09),(44.40 9.91)
24429	France	Russia	Russia	Russia	France	France	France	(50.59 1.93),(44.02 4.97)
39106	Israel	Greece	Greece	Greece	Russia	Germany	Sweden	(41.68 4.76),(36.35 23.47),(41.78 5.04)
49793	France	Romania	Romania	Romania	Russia	Russia	Russia	(41.31 9.54),(40.83 19.25)
47137	France	France	Germany	Germany	Austria	Switzerland	Switzerland	(38.01 12.40),(41.96 12.49)
34848	Egypt	Turkey	Turkey	Turkey	France	France	France	(38.75 9.47),(37.47 14.32)
10635	France	France	France	France	Russia	Russia	Bulgaria	(53.28 5.54),(45.50 5.56),(50.11 7.61)
18548	Czech Republic	Germany	Germany	Germany	Russia	Russia	Russia	(42.54 13.24),(38.67 9.20)
22423	Russia	Ukraine	Ukraine	Ukraine	Russia	Russia	Russia	(53.32 7.55),(50.79 17.17)
13411	France	Russia	Russia	Russia	France	France	France	(39.22 9.31),(46.77 5.19)
42867	Switzerland	Switzerland	Switzerland	Russia	Switzerland	Switzerland	Switzerland	(47.42 16.49),(49.47 -4.12)
33744	Switzerland	Switzerland	Russia	Germany	Spain	Switzerland	Switzerland	(42.39 13.98),(40.81 5.44),(50.07 2.19)
31350	Israel	Romania	Romania	Romania	Russia	Russia	Russia	(42.23 7.63),(38.30 19.55)
15990	France	Russia	Russia	Russia	Greece	Greece	Greece	(38.60 5.52),(38.88 15.46)



(a) CH-FR (305Km) (b) CH-ES (472Km) (c) CH-IT (462Km) (d) CH-UK (457Km)

(e) FR-ES (416Km) (f) FR-IT (374Km) (g) FR-PT (528Km) (h) FR-UK (447Km)

(i) ES-IT (659Km) (j) ES-PT (701Km)

Figure 3.8: Ancestral location prediction error in real POPRES admixed individuals, stratified by the country of origin of each location. Letters are the inferred locations, and the shaded circles are the assumed true locations.

CHAPTER 4

Leveraging Multi-SNP Reads from Sequencing Data for Haplotype Inference

4.1 Motivation

Humans are diploid organisms with two copies of each chromosome, one inherited from the father and the other from the mother. The two copies are very similar to each other and only differ at a small fraction ($\sim 0.1\%$) of sites. Most of the variation is contained at single nucleotide polymorphic (SNP) sites. The sequence of alleles on each chromosome is referred to as the haplotype. Haplotype information is centrally important for a wide variety of applications, including association studies and ancestry inference [Laz01, HCZ01, RDS01, FD01, SRH02, MG03]. Unfortunately, standard methods for probing genetic variation are able to collect only genotype information but not haplotypes. A large number of computational methods, referred to as haplotype phasing approaches, have been proposed to infer haplotypes from genotypes. The most successful methods utilize a set of reference haplotypes to build a probabilistic model of the haplotypes in the population [HDM09, BY09, LMN09, LWD10, KZE10, HMS11]. Using a population genetics model for the haplotype distribution, these models predict the most likely haplotype data that can explain the observed genotypes.

Rapid advances in hight throughput sequencing (HTS) technologies provide new opportunities for haplotype phasing methods. HTS yields short segments of the DNA (reads) where each read originates from one of the pair of chromosomes.

Therefore, all of the alleles in this read are from the same haplotype. Although reads that cover multiple SNPs (multi-SNP reads) could be used to improve haplotype inference, existing methods generally ignore this information, partially due to computational difficulty associated to modeling such reads.

Several methods have been proposed to predict haplotypes directly from the reads. These methods, referred to as haplotype assembly methods, utilize overlapping reads to construct the haplotype [BB08, BHA08, DHM10, HCP10, AI12, XWJ12, DMH12]. The most commonly used objective function for haplotype assembly is the minimum error correction (MEC). The MEC objective function aims at finding the minimum number of edits such that the reads can be partitioned into two disjoints sets and each set of reads originate from one of the haplotypes. However, since these methods do not use the information in the reference haplotype panel, they significantly underperform standard phasing methods that ignore read information but utilize reference panel [HCP10]. Recently, one of these methods has been extended to utilize the reference [HHE12, HE13]. Unfortunately, this method has prohibitive memory and time requirements, thus making it unfeasible for moderate to large data sets.

Here we propose a novel approach called HARSH (HAplotyping with Reference and Sequencing tecHnology) for haplotype phasing. We utilize a probabilistic model to incorporate the multi-SNP read information together with a reference panel of haplotypes. We use an efficient Gibbs sampling method to find sample from the posterior distribution. This algorithm has the advantages of being computationally efficient, scalable in memory usage and accurate in genotyping and phasing prediction. We evaluate our method on simulations from real haplotypes from the HapMap project. At 1X coverage, HARSH gives around 10% improvement in terms of total error rate compared with standard phasing approaches that do not use the multi-SNP read information thus showing the benefits of modeling multi-SNP reads. We also evaluate HARSH and the basic model for varying coverage and read length, showing the benefits of our approach in higher coverage and longer read length. Additionally, we test our method on simulations starting from real sequencing data of 1000 Genomes project, where the density of SNPs is much higher than that in HapMap data. Through extensive simulations we show that the gain in performance of our approach over existing models extends to realistic read lengths (e.g. $100 - 400$ base pairs) making our approach readily applicable to existing sequencing data sets. With recent works showing that short read sequencing can dramatically increase association power in genome-wide association study (GWAS) over genotyping arrays [PRM12], we expect our approaches to further increase power in GWAS by increasing accuracy in genotype calling and phasing from short read data.

4.2 Methods

The best performing approaches for haplotype inference rely on Hidden Markov Models for describing the distribution of haplotypes in the population. These approaches generally ignore multi-SNP information in the reads thus implementing the model as a linear chain graph. The model structure becomes complicated when we are considering multi-SNP information as it is not trivial to perform standard operations (e.g. Viterbi decoding) to a non-linear chain graph. Previous methods (e.g. Hap-SeqX [HE13]) have attempted to extend the Viterbi algorithm to the complex graph induced by multi-SNP reads and reference haplotypes but the approach is very expensive in both time and memory usage. As opposed to previous approaches, in this work we use a Gibbs sampler based method for fast inference. The main advantage of this approach is that the computations are efficient and it can achieve the optimal or close to optimal solution in a feasible amount of time. However, all other current methods are either not optimal or not practical in terms of computational time or memory usage.

4.2.1 Gibbs Sampler Preliminaries

A Gibbs sampler serves as the basis for our method. We first introduce the general idea of Gibbs sampling before we use it to solve the haplotype problem. Consider the following distribution typically used to perform optimization in graphical models

$$P(X) \;=\; \frac{1}{Z}\exp\left(\mu\sum_{i=1}\sum_{j=1}\phi_{ij}(x_i, x_j)\right)$$

where $X = (x_1, x_2, \cdots x_d)$ is a d-dimensional vector and Z is a normalization factor. The function ϕ specifies the edge potential for two variables with an edge between them. We would like to collect samples of X based on this distribution $P(X)$. However, sampling directly from the full distribution is not trivial. Gibbs sampler is one of such methods designed for efficiently sampling from the $P(X)$ distribution.

Gibbs sampler is a special case of Monte Carlo Markov Chain (MCMC) method [GG84], which is guaranteed to converge to the equilibrium distribution after sufficient burn-in rounds. In each round, it randomly samples one variable x_i based on the conditional probability $P(x_i|x_{[-i]})$ when all other variables $x_{[-i]} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$ are fixed. Formally, this conditional probability can be written as follows

$$P(x_i = t|x_{[-i]}) \;=\; \frac{P(x_i = t, x_{[-i]})}{\sum_{t'} P(x_i = t', x_{[-i]})}. \qquad (4.1)$$

A more complete treatment of MCMC is available in [Liu08].

4.2.2 Haplotype Assembly with Sequencing Data

Sequencing technologies provide us with a set of reads, each of which is a short fragment from one of the chromosomes. Haplotype assembly aims to assemble the whole haplotype based on only read information. An illustrative example is given in Figure 4.1.

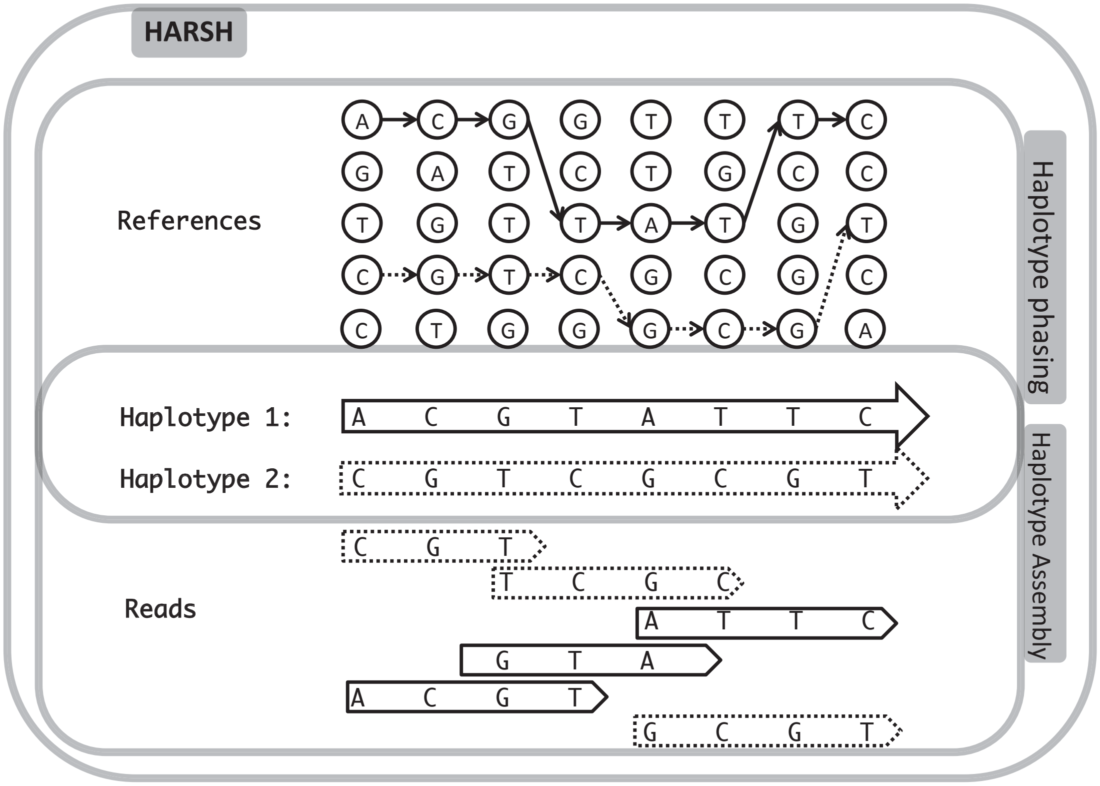


Figure 4.1: An illustration of haplotype inference problems. The two chromosomes for an individual are unknown to us at first. Sequencing technology produces a set of reads, each of which originates from one of the two chromosomes. We also have a set of reference haplotypes, which are from the same population as the donor. Haplotype assembly aims to assemble the two donor haplotypes by only using the read information. Haplotype phasing problem aims to phase the two haplotypes by mosaic copies from the reference haplotypes. However, our approach HARSH takes into account both read information and reference panel for more accurate haplotype inference.

We can formalize this problem as follows. Suppose that we only consider L biallelic SNPs and M reads. Each read is represented by $X_j = \{-1, 1, 0\}^L$, where 0 stands for unobserved SNP in jth read, -1 and 1 stand for observed minor and major alleles, respectively. Since the homozygous site does not affect the haplotype phasing, we only consider heterozygous sites. Therefore, the objective is to find a sequence of haplotype and its complementary $\{h, \bar{h}\}$ where $h = -\bar{h} \in \{-1, 1\}^L$, to minimize the total number of flipped loci within reads, such that every read can be perfectly assigned to one of the haplotypes. Another necessary variable for the model is the read origin indicator $r_j \in \{-1, 1\}$. If $r_j = 1$, the j-th read is assumed to have been generated from haplotype h, and if $r_j = -1$, the j-th read is from the complementary haplotype $\bar{h}$. We assume the read generation process is as follows. First we randomly pick one of the haplotypes $(h, \bar{h})$ with equal probability, and then sample the read starting position from one of the L possible positions in the genome. If we consider the read generation processing is error free then we have $x_{ij} = h_i r_j$. However if the read generation process is error-prone and ϵ indicates the rate of sequencing error then with probability $1-\epsilon$ we have $x_{ij} = h_i r_j$ and with probability ϵ we have $x_{ij} = -h_i r_j$. An illustrative example is given in Figure 4.2.

We can formalize the connection between the haplotypes and read origin variables into the following probabilistic distribution. For each possible values of the haplotypes and read origin variables, we can calculate its probability as:

$$\begin{aligned} & P(R, H; X) \\ = \; & \frac{1}{Z} \exp\left(\mu\left(\sum_{ij:x_{ij}=1} \theta_{ij}(h_i, r_j) + \sum_{ij:x_{ij}=-1} \eta_{ij}(h_i, r_j)\right)\right) \end{aligned} \tag{4.2}$$

where

$$\theta_{ij}(h_i, r_j) = \begin{cases} \ln(1-\epsilon) & h_i = r_j \\ \ln \epsilon & h_i \neq r_j \end{cases}$$

$$\eta_{ij}(h_i, r_j) = \begin{cases} \ln \epsilon & h_i = r_j \\ \ln(1-\epsilon) & h_i \neq r_j \end{cases}$$

and the variables $R = (r_j)_{j=1}^{M}$, $H = (h_i)_{i=1}^{L}$ and $X = (x_{ij})_{ij}$ are vectors and matrix composed of scalar variables r, h and x. The variable Z is a normalization constant to ensure $\sum_{R,H} P(R, H; X) = 1$. The functions θ and η specify edge potentials that favor h and r to be of equal values and opposite values, respectively. The model parameter μ controls the "heat" of the probabilistic model. Generally speaking, the probability distribution is more smooth when μ is small and more sharp when μ is large.

Lemma 1. *The maximum a posteriori (MAP) assignment of (4.2) corresponds to the MEC haplotype for any $\epsilon < 0.5$.*

Proof. We can prove by constructing the MEC haplotype from MAP assignment. Let H^* and R^* denote the MAP assignment of our probabilistic model and the corresponding probability calculated from (4.2) will be

$$P(H^*, R^*; X) = \frac{1}{Z} \exp(\mu(n \ln(1-\epsilon) + m \ln \epsilon))$$

where n is the number of edges getting potential $\ln(1-\epsilon)$ and m is the number of edges getting potential $\ln \epsilon$ based on the configuration H^* and R^*. As $\ln(1-\epsilon) > \ln \epsilon$ for $\epsilon < 0.5$ and the number of edges is fixed, this MAP assignment H^* and R^* is actually minimizing the number of edges getting potential $\ln \epsilon$. We can use this haplotype H^* and flip every read bit corresponding to the edge getting potential $\ln \epsilon$. The resulting MEC score for H^* will be m, which is minimized.

Suppose that there exists another haplotype H' with MEC score $m' < m$. It suggests that we can flip only m' read bit then all the reads will be perfectly

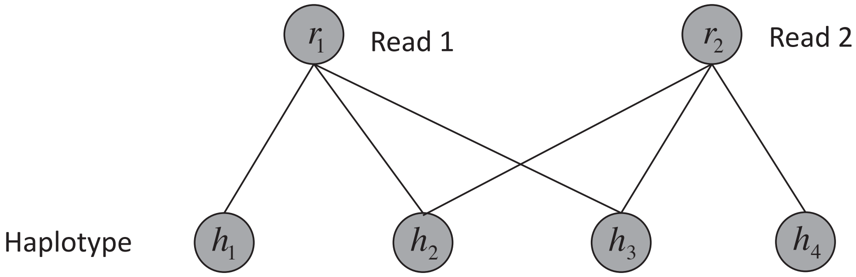


Figure 4.2: A graphical model for haplotype assembly. In this example, two reads and four heterozygous SNPs are considered. Read 1 covers the SNPs 1, 2 and 3. Read 2 covers SNPs 2, 3 and 4. The variables $h \in \{1, -1\}$ stands for the haplotype. The variable $r \in \{1, -1\}$ stands for whether the read is from haplotype h or the complementary $\bar{h}$.

assigned to one of the haplotypes. We keep those assignments into the variable R'. Thus, we should have

$$P(H', R'; X) = \frac{1}{Z} \exp(\mu((n + m - m') \ln(1 - \epsilon) + m' \ln \epsilon)).$$

By definition $m' < m$, thus $P(H', R'; X) > P(H^*, R^*; X)$, which contradicts the fact that H^* and R^* is the MAP assignment maximizing the configuration probability. By this contradiction, we can conclude that there does not exist H' and R' with MEC score $m' < m$. □

4.2.3 Haplotype Phasing with Sequencing Data and Reference

Current haplotype assembly methods mainly focus on *de novo* assembly, which uses short reads as the only information source. This is partially due to the complexity of extending the method to the scenario of assembly using reference. On the other hand, current haplotype phasing methods only use the reference panel and genotype likelihood in each SNP but ignore the multi-SNP information in the reads. We aim to utilize both the reference panel and sequencing data to perform haplotype phasing as shown in Figure 4.1. Formally, suppose that we are only considering L biallelic SNPs, M reads and N reference haplotypes. Each read is represented by $X_j = \{-1, 1, 0\}^L$, where 0 stands for unobserved SNP in jth read. The objective is to find two haplotypes $H = \{h^1, h^2\}$ where $h^1, h^2 \in \{-1, 1\}^L$. We want to find the two haplotypes with small number of inconsistent loci with reads, as well as more consistent with reference haplotypes. We use another set of variables $S = \{s^1, s^2\}$ where $s^1, s^2 \in \{1, 2, \ldots, N\}^L$ to stand for the assignment of each loci to reference haplotypes. We also need a set of variables $R = \{r_1, r_2, \ldots, r_M\}$ where $r_i \in \{-1, 1\}$ stands for the haplotype that each read originates from. An illustrative example of the graph structure is given in Figure 4.3.

Similar to the previous section, we can formalize the connection between the three variables H, R and S into the following probabilistic distribution. For each possible values of H, R and S, we can calculate its probability as follows

$$\begin{aligned}
& P(H, R, S; X) \\
= \; & \frac{1}{Z} \exp \Bigg[\mu \cdot \Bigg(\sum_{ij:x_{ij}=1} \theta(h_i^1, -r_j) + \sum_{ij:x_{ij}=-1} \eta(h_i^1, -r_j) \\
& \qquad + \sum_{i=1}^{L} \xi(h_i^1, s_i^1) + \sum_{i=1}^{L-1} \tau(s_i^1, s_{i+1}^1, i) \\
& \qquad + \sum_{ij:x_{ij}=1} \theta(h_i^2, r_j) + \sum_{ij:x_{ij}=-1} \eta(h_i^2, r_j) \\
& \qquad + \sum_{i=1}^{L} \xi(h_i^2, s_i^2) + \sum_{i=1}^{L-1} \tau(s_i^2, s_{i+1}^2, i) \Bigg) \Bigg] \qquad (4.3)
\end{aligned}$$

where we have four edge potential functions. The functions θ and η are defined similarly as in (4.2) except that there would no penalty if the read is assigned by r to the other haplotype.

$$\theta(h_i, r_j) = \begin{cases} \ln(1-\epsilon) & r_j = 1, h_i = 1 \\ \ln \epsilon & r_j = 1, h_i = -1 \\ 0 & r_j = -1 \end{cases},$$

$$\eta(h_i, r_j) \;=\; \begin{cases} \ln\epsilon & r_j = 1, h_i = 1 \\ \ln(1-\epsilon) & r_j = 1, h_i = -1 \\ 0 & r_j = -1 \end{cases} \cdot$$

The edge potential function ξ specifies the "haplotype copying" which is motivated that the predicted haplotype is a mosaic of reference haplotypes with a small number of differences. In this case, the predicted haplotypes are similar to reference haplotype s^1 and s^2 at position i.

$$\xi(h_i^1, s_i^1) \;=\; \begin{cases} \ln(1-\omega) & h_i^1 = G_{s_i^1, i} \\ \ln\omega & h_i^1 \neq G_{s_i^1, i} \end{cases}$$

where G_{ij} stands for the jth allele in ith reference haplotype. Thus, $G_{s_i^1,i}$ stands for the ith allele in s_i^1-th reference haplotype. Moreover, we use the following function to model the transition probability in haplotype copying model [LS03].

$$\tau(s_i, s_{i+1}, i) = \begin{cases} \exp(-\frac{\rho_i}{N}) + (1 - \exp(-\frac{\rho_i}{N}))/N & s_i = s_{i+1} \\ (1 - \exp(-\frac{\rho_i}{N}))/N & s_i \neq s_{i+1} \end{cases}$$

where $\rho_i = 4N_e r_i$ and r_i is the per generation genetic distance between site i and site $i+1$, and N_e is a constant.

This probabilistic model provides us a disciplined way to infer the most probable haplotype given a set of reads and a set of reference haplotypes. It extends the haplotype copying model [LS03] from genotype input to sequencing data input. It also extends the haplotype assembly problem in previous section to a more general case where the reference panel can be utilized to improve the phasing. We are then able to design efficient sampling approach to find the most possible configurations of H, R, and S that maximize the probability given in Equation (4.3).

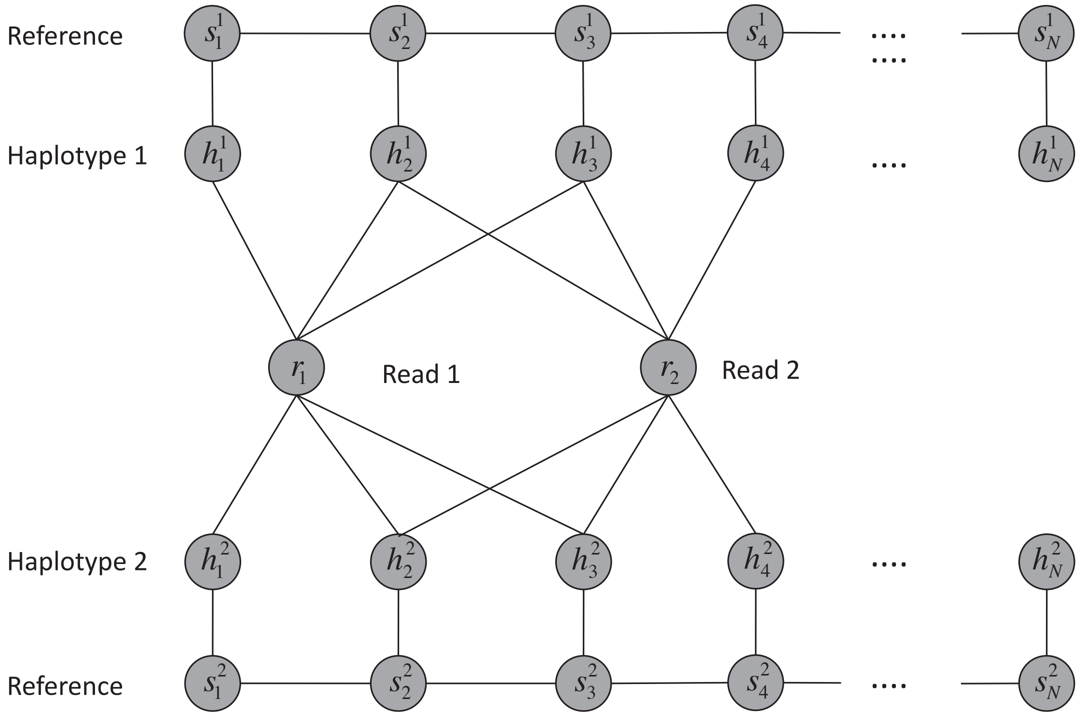


Figure 4.3: A graphical model for haplotype phasing with reference. The variables h^1 and h^2 stand for the first and second haplotypes. The variables $r_i = \{-1, 1\}$ specify whether the read comes from the first haplotype or second haplotype. The variable s^1 and s^2 specify which haplotype in the reference is generating the haplotype h^1 and h^2, respectively.

4.2.4 Efficient Sampling

Haplotype assembly without reference. The bipartite structure in Figure 4.2 suggests an efficient procedure for sampling. For fixed one layer of the bipartite graph, the variables in the other layer will be independent on each other. Thus, the conditional probability in Equation (4.1) of Gibbs sampler can be significantly reduced. Formally, following the standard procedure of Gibbs sampling, we can sample haplotype from the conditional probability for fixed read origins. The sampling ratio $\delta_i = P(h_i = -1|R)$ can be calculated as follows

$$\delta_i = \frac{\exp\left(\sum_{j:X_{ij}=1}\theta(-1,r_j) + \sum_{j:X_{ij}=-1}\eta(-1,r_j)\right)}{\begin{pmatrix} \exp\left(\sum_{j:X_{ij}=1}\theta(-1,r_j) + \sum_{j:X_{ij}=-1}\eta(-1,r_j)\right) \\ +\exp\left(\sum_{j:X_{ij}=1}\theta(1,r_j) + \sum_{j:X_{ij}=-1}\eta(1,r_j)\right) \end{pmatrix}}. \tag{4.4}$$

Similarly, we can also do a similar Gibbs sampling step for read origin for fixed haplotype. The sampling ratio $\rho_j = P(r_j = -1|H)$ can be calculated as follows

$$\rho_j = \frac{\exp\left(\sum_{i:X_{ij}=1}\theta(h_i,-1) + \sum_{i:X_{ij}=-1}\eta(h_i,-1)\right)}{\begin{pmatrix} \exp\left(\sum_{i:X_{ij}=1}\theta(h_i,-1) + \sum_{i:X_{ij}=-1}\eta(h_i,-1)\right) \\ +\exp\left(\sum_{i:X_{ij}=1}\theta(h_i,1) + \sum_{i:X_{ij}=-1}\eta(h_i,1)\right) \end{pmatrix}}. \tag{4.5}$$

The complete sampling algorithm for haplotype assembly is shown in Algorithm 1. As default, we use $10,000$ rounds for sampling.

Haplotype phasing with reference. The sampling for haplotype phasing with both sequencing data and reference from the graph in Figure 4.3 is more challenging. However, we can still take advantages of the special structure of the graph and perform efficient sampling procedure. Following the idea of Gibbs sampler, we will alternatively (a) sample read origin R for fixed haplotype H and reference assignment S; (b) sample S for fixed R and H; (c) sample H for fixed R and S. The step (a) is similar with that in haplotype assembly. Formally, the sampling

Algorithm 1 Sampling Algorithm for Haplotype Assembly

```
1: Randomly initialize haplotype $H$.
2: For fixed haplotype $H$, sample read origin $R$. For probability $\rho_j$, we get $r_j = -1$ and for
   probability $1 - \rho_j$, we get $r_j = 1$ where the ratio $\rho$ can be calculated as in (4.5).
3: For fixed read origin $R$, sample haplotype $H$. For probability $\delta_i$, we get $h_i = -1$ and for
   probability $1 - \delta_i$, we get $h_i = 1$ where the ratio $\delta$ can be calculated as in (4.4).
4: Repeat step 2 and 3 for sufficient rounds until equilibrium.
5: Collect samples by repeating step 2 and 3, and output the one with highest probability.
```

ratio $P(r_j = -1|H, S)$ for read origin can be calculated by

$$\rho_j = \frac{\exp\left(\sum_{i:X_{ij}=1} \theta(h_i^1, 1) + \sum_{i:X_{ij}=-1} \eta(h_i^1, 1)\right)}{\left(\begin{array}{c} \exp\left(\sum_{i:X_{ij}=1} \theta(h_i^1, 1) + \sum_{i:X_{ij}=-1} \eta(h_i^1, 1)\right) \\ + \exp\left(\sum_{i:X_{ij}=1} \theta(h_i^2, 1) + \sum_{i:X_{ij}=-1} \eta(h_i^2, 1)\right) \end{array}\right)}. \tag{4.6}$$

The step (c), sampling of haplotype H for fixed read origin R and reference assignment S is a straightforward extension from Equation (4.4). The modification is based on the extra edge between reference penal variables S and haplotype H. Formally, the sampling ratio $P(h_i^1 = -1|R, S)$ for the first haplotype can be calculated by

$$\delta_i^1 = \frac{\alpha(-1)}{\alpha(-1) + \alpha(1)} \tag{4.7}$$

where

$$\alpha(h) = \exp\left(\sum_{j:X_{ij}=1} \theta(h, -r_j) + \sum_{j:X_{ij}=-1} \eta(h, -r_j) + \xi(h, s_i^1)\right).$$

The sampling ratio $P(h_i^2 = -1|R, S)$ is similar with $P(h_i^1 = -1|R, S)$. Similarly, we can obtain the sampling ratio for the second haplotype as follows

$$\delta_i^2 = \frac{\beta(-1)}{\beta(-1) + \beta(1)} \tag{4.8}$$

where

$$\beta(h) = \exp\left(\sum_{j:X_{ij}=1} \theta(h, r_j) + \sum_{j:X_{ij}=-1} \eta(h, r_j) + \xi(h, s_i^2)\right).$$

The step (b), sampling for the haplotype reference panel variables S for fixed read origin R and haplotype H is very challenging. The difficulty comes from the dependency between the variables s_i and s_{i+1}, and the large number of possible values for each s_i. Note that unlike the binary variable h and r, the variable $s_i \in \{1, 2, \ldots, N\}$ where N is the number of reference haplotypes. Thus, straightforward Gibbs sampler would be very inefficient in this case. To tackle this computational challenge, we resort to the following Markov chain sampling procedure [Liu08]. The joint distribution over all variables in S can be written as follows

$$P(S|H) \quad = \quad \frac{1}{Z} \exp\left(\phi_0(s_1) + \sum_{i=1}^{L-1} \phi_i(s_i, s_{i+1})\right) \tag{4.9}$$

where

$$\begin{aligned} \phi_0(s_1) &= \xi(h_1, s_1) \\ \phi_i(s_i, s_{i+1}) &= \tau(s_i, s_{i+1}, i) + \xi(h_{i+1}, s_{i+1}). \end{aligned}$$

Sampling directly from $P(S|H)$ is still tedious in practice. However, we can convert the $P(S|H)$ to multiplication series of probability functions as follows $P(s_1|s_2, H)P(s_2|s_3, H) \cdots P(s_{L-1}|s_L, H)P(s_L, H)$. Then sampling from $P(s_L)$ and sampling backward using those conditional probabilities becomes trivial. We can use dynamic programming to convert the $P(S|H)$ distribution to the alternative form. We define

$$V_1(s_2) = \sum_{s \in S} \exp\left(\phi_0(s)\phi_1(s, s_2)\right)$$

and

$$V_i(s_{i+1}) = \sum_{y \in S} V_{i-1}(y) \exp(\phi_i(y, s_{i+1})) \text{ for } i = 2, \cdots, L.$$

Thus, we can compute the normalization factor $Z = \sum_{s_L \in S} V_{L-1}(s_L)$ efficiently using dynamic programming, and then we can compute the marginal probability $P(s_L, H) = (V_{L-1}(s_L))/Z$. Moreover, we can backward compute $P(s_i|s_{i+1}, H)$ similarly. Note that a naive implementation of this step would result in a complexity of quadratic in the number of reference haplotypes. We take advantage

Algorithm 2 Sampling Algorithm for Haplotype Phasing

```
1: Randomly initialize haplotype $H$
2: For fixed haplotype $H$, sample read origin $R$ using sampling ratio $\rho_j$ in (4.6).
3: For fixed haplotype $H$ sample haplotype reference $S$ following Markov chain
   sampling procedure described after (4.9).
4: For fixed read origin $R$, and haplotype reference $S$, sample haplotype $H$ using
   sampling ratio $\delta_i$ in (4.7).
5: Repeat step 2, 3 and 4 for sufficient rounds until equilibrium.
6: Collect samples by repeating steps 2, 3 and 4. Output samples with highest
   probability.
```

of the symmetry in the haplotype coping model to reuse computation to achieve runtime linear in the number of reference haplotypes.

An outline of the sampling algorithm for haplotype phasing with sequencing data and a reference panel is given in Algorithm 2. As default, we use $10,000$ rounds of sampling.

4.2.5 Web Resources

The software implementation of this method is freely available to public at `http://genetics.cs.ucla.edu/harsh`

4.3 Experimental Results

4.3.1 Data Sets and Experimental Settings

We performed simulation experiments using HapMap Phase II data [Int05] and 1000 Genomes data [Dur10]. For our simulations we used the 60 parental individuals of CEU populations from HapMap Phase II as well as 60 individuals randomly chosen from the European populations for 1000 Genomes data. Though our method is scalable to the whole genome, for the purpose of demonstration, we use only chromosome 22 as representative of the rest of the genome, as it is the shortest chromosome. Because we are performing many simulations, we restrict our results to the $35,421$ SNPs in chromosome 22 of the HapMap data, and the first $30,000$ SNPs in chromosome 22 of 1000 genomes data, which span around 3 Mb. The datasets are publicly available at http://mathgen.stats.ox.ac.uk/impute/ and http://hapmap.ncbi.nlm.nih.gov/.

We evaluate our method using a leave-one-out procedure. In each round, we infer the haplotype for one individual using simulated sequencing data and the haplotypes of the other 59 individuals as reference panel. This procedure is repeated 60 times and all the evaluation metrics are averaged. The reads are simulated uniformly across chromosome 22 for a given coverage. The read length in each end of a pair-end read is fixed but the gap between the two ends follow a normal distribution with fixed mean and standard deviation. Errors are inserted in the read at a rate ϵ.

We evaluate our method HARSH using the standard metric for genotyping and phasing accuracy: genotyping error rate and switching error rate. The genotyping error rate is the proportion of wrongly predicted genotypes, and the switching error is the proportion of switches in the inferred haplotypes to recover the correct phase in an individual. The total error rate is the sum of genotyping error rate and switching error rate. We also use percentage improvement when comparing two methods. The percentage improvement is computed as the error rate difference between two methods normalized by the error rate of baseline method. For example, suppose that HARSH has error rate x and baseline method has error rate y, the improvement of HARSH over the baseline method would be $(y-x)/y$.

We fixed the parameters $\mu = 1$, $\omega = 0.002$ and $\epsilon = 0.01$ for all our experiments. From our experience, the performance of the proposed method is not sensitive to parameter tuning. Using μ from 1 to 10 and ω from 0.001 to 0.005 does not affect the performance significantly. The sequencing error $\epsilon = 0.01$ is standard sequencing error rate.

All experiments are performed in a cluster machine where each node has 8 to 16 cores 3.0GHz CPU and 1G to 16G memory. Jobs are submitted in a parallel manner but each job uses only one node.

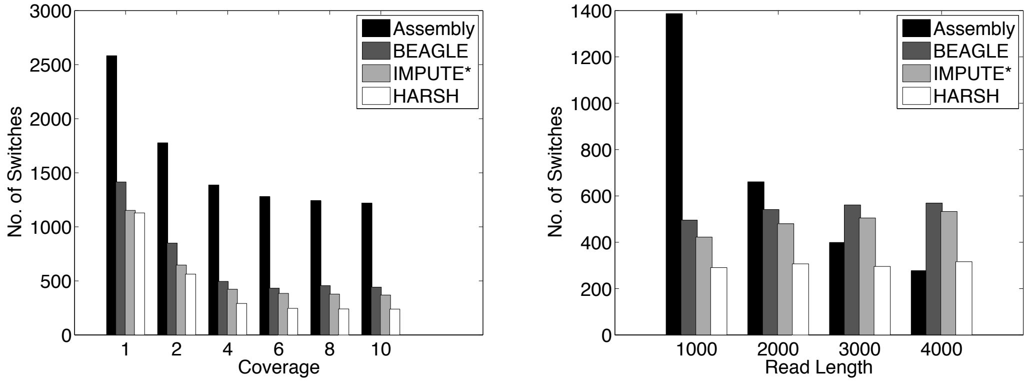
4.3.2 HapMap Simulations

We use HapMap data set to evaluate our method HARSH. We compare our method with three other state-of-the-art methods: the Hidden Markov Model (HMM) at the core of the IMPUTE method [HDM09], BEAGLE [BB09] and Hap-SeqX [HE13]. Since IMPUTE does not support haplotype phasing for uncovered SNPs, for a fair comparison, we re-implemented the basic HMM model of the IMPUTE v1.0, which uses the pre-defined genetic map information for transition probability. We will refer to our implementation of the HMM model in IMPUTE method as IMPUTE*. In our modified version, we use the read count for each SNP as input to IMPUTE* method. The likelihood of read count from genotype is used as the emission probability for the HMM model. Then the Viterbi algorithm is utilized to decode two paths from the reference panel which are most likely to generate the read counts in each SNP. The two paths in reference panel also give the two predicted haplotypes. Since the latest implementation of IMPUTE [HDM09] is not able to phase, we also compared our approach to BEAGLE 3.3.2 [BB09], a widely used approach for haplotype phasing and imputation.

We first use the HapMap data set to show that haplotype assembly without a reference panel will underperform haplotype phasing with a reference panel. The main reason is that there are not enough long reads covering all continuous heterozygous SNPs. Thus, haplotype assembly can not do more than random guess between two continuous heterozygous SNPs if there is no read spanning them. We can compute a lower bound of the number of switches for haplotype assembly as $K/2$ where K is the number of those gaps, assuming the Minimum Error Correction (MEC) score to be zero. For pair-end reads with fixed length $1,000$ bp mean and 100 bp standard deviation, we evaluate our method using six levels of sequencing coverages: 1X, 2X, 4X, 6X, 8X and 10X. As shown in Figure 4.4(a), higher coverage does not help haplotype assembly to achieve similar performance than haplotype phasing methods. At fixed coverage 4X, we simulated pair-end reads with $1,000$ bp, $2,000$ bp, $3,000$ bp and $4,000$ bp in each end. As shown in Figure 4.4(b), we can observe that the lower bound of haplotype assembly achieves similar performance as haplotype phasing only under the very unrealistic read length $4,000$ bp. Also, at 4X coverage, we can observe that our method can improve around 44% over BEAGLE and around 37% over IMPUTE in terms of numbers of switches.

For simulated pair-end reads with $1,000$ bp for each end at 1X coverage. Only 32% reads contain one SNP and around 26% of the reads contain more than 3 SNPs. On average, every read contains around 2.8 SNPs. Following the same procedure as [HE13], we divide the chromosome into overlapping chunks containing $1,200$ SNPs each and run our method on each chunk independently. The final haplotypes are then constructed by stitching together the haplotypes from each chunk. Chromosome 22 is divided into a total of 36 chunks. The total error rate for both IMPUTE* and HARSH are shown in Figure 4.5. We can observe from the figure that HARSH consistently performs better than IMPUTE* across all 36 chunks. The average improvement over IMPUTE* is 7.6%. We then concatenated those haplotype chunks by minimizing the mismatches in the overlap region between two adjacent chunks. After concatenation, the overall error rate for HARSH is 4.01% for chromosome 22, compared with 4.42% for IMPUTE*. The overall improvement is 9.3% over IMPUTE*.

We compare HARSH with a previous method for combining multi-SNP reads



(a) Varying coverage for fixed read length1, 000bp

(b) Varying read length for fixed coverage4X

Figure 4.4: The number of switches within heterozygous SNPs for haplotype assembly, BEAGLE, IMPUTE* and HARSH. The number of switches of haplotype assembly is estimated by the lowest bound.

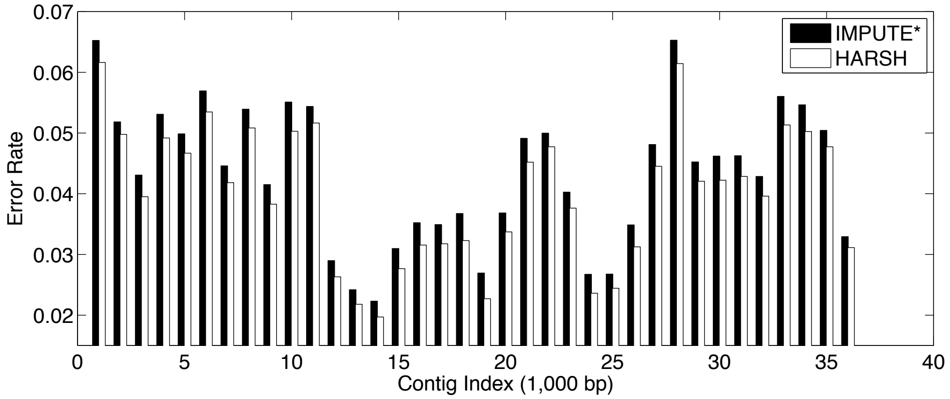


Figure 4.5: The error rate for IMPUTE* and our method for each chunk of length $1,200$ SNPs in chromosome 22. The error rate consists of both genotyping error for all SNPs and switch error within heterozygous SNPs.

with a reference panel, Hap-SeqX [HE13]. Hap-SeqX is an approximation to the dynamic programming approach of the Hap-Seq method [HHE12] which optimizes a similar objective function to HARSH. Hap-SeqX only searches a fraction of the search space compared to Hap-Seq by only storing the top values at each state.

However, Hap-SeqX is still a very expensive method in both time and memory usage. In this experiment, we use the default parameters of Hap-SeqX, where $t = 0.01$ specifies that the algorithm saves the top 1% of values for each state. On addition, Hap-Seq and Hap-SeqX, unlike HARSH, can only handle up to three SNPs in a read and split reads containing more SNPs into multiple reads. The performance comparisons are shown in Table 4.1. HARSH and IMPUTE* have similar running time. HARSH takes about 10 minutes compared to IMPUTE* 5 minutes on chromosome 22. Both of these methods compare very favorably to Hap-SeqX which takes 5 hours for the same dataset. Cross validation of 60 individuals would be prohibitive for Hap-SeqX. Thus, We compare all these three methods using only the first individual in HapMap data set. The results averaged over 36 chunks. We can see that Hap-SeqX improves by around 12.53% from the baseline method IMPUTE*, and HARSH significantly improves by 21.34% from IMPUTE*. We conducted significance test (paired-sample t-test) on the improvement of HARSH over Hap-SeqX and IMPUTE*. The test results show that HARSH significantly outperforms both Hap-SeqX and IMPUTE* with p-value $< 1 \times 10^{-3}$ and p-value $< 1 \times 10^{-7}$, respectively. Overall, the comparison shows that HARSH is the most accurate and practical method among existing methods.

To fully evaluate the performance of our method, we apply our method to cases

Table 4.1: Comparison between IMPUTE*, Hap-SeqX and HARSH on a HapMap data set with 1 donor individual, 59 reference individuals and $35,421$ SNPs. $1,000$bp read length and 1X coverage are simulated.

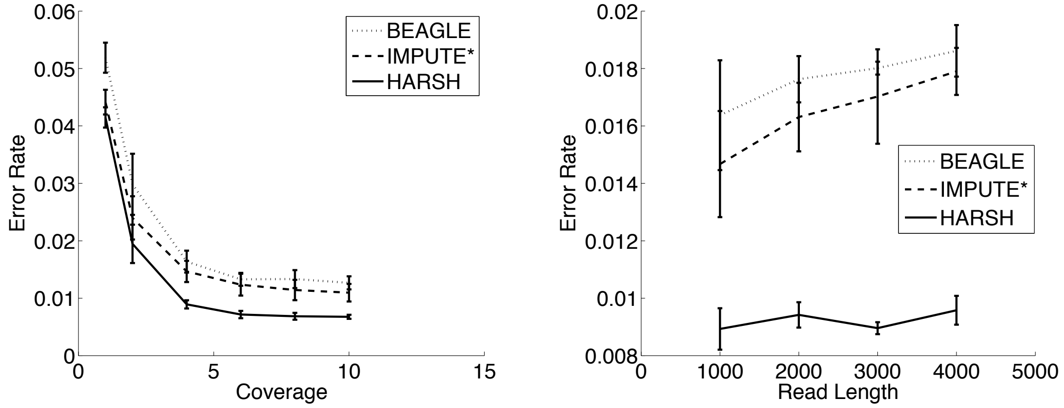
Methods	Error Rate (Switch, Genotyping)	Time
IMPUTE*	0.04836 (0.00804, 0.04033)	∼ 5 minutes
Hap-SeqX	0.04230 (0.00726, 0.03504)	∼ 5 hours
HARSH	0.03804 (0.00664, 0.03140)	∼ 10 minutes

with different coverages and read lengths. For pair-end reads with fixed length $1,000$ bp mean and 100 bp standard deviation, we evaluate our method using six levels of sequencing coverages: 1X, 2X, 4X, 6X, 8X and 10X. The result is shown in Figure 4.6(a). As expected, the performance improvement of HARSH over BEAGLE and IMPUTE* becomes more significant when the coverage increases. The reason we expect this is that the higher the coverage, the larger number of reads that span multiple SNPs. HARSH is able to take advantage of the multi-SNP information within those reads but BEAGLE and IMPUTE* can not take advantage of that. In Table 4.2, we show the genotyping and switching error rate of HARSH and IMPUTE* method for different coverages. It can be observed that both genotyping error and switching error are significantly reduced by HARSH over BEAGLE and IMPUTE*. It is also worth mentioning that 4X seems to be the best choice in terms of the compromise between the cost of coverage and achieved accuracy. The coverage 4X gives 0.28% genotyping error and 0.62% switching error. However, the improvement of higher coverage than 4X is limited.

We also evaluate HARSH with different read lengths. At fixed coverage 4X, we simulated pair-end reads with $1,000$ bp, $2,000$ bp, $3,000$ bp and $4,000$ bp in each end. The results are shown in Figure 4.6(b). It is not immediately intuitive

Table 4.2: Genotyping and switching errors (%) for varying coverages on HapMap data set. Read length is fixed to be $1,000$ bp.

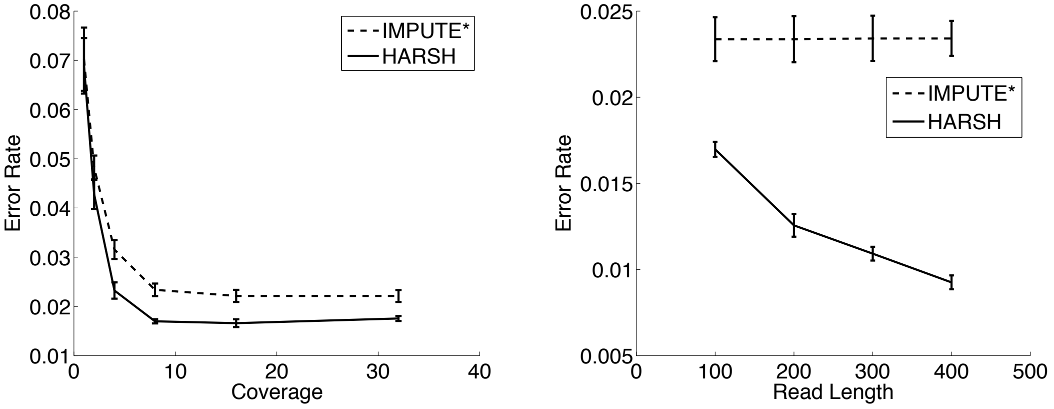
Coverage		1X	2X	4X	6X	8X	10X
Genotyping Error	BEAGLE	4.21	1.94	0.59	0.22	0.10	0.04
	IMPUTE*	3.59	1.53	0.56	0.30	0.17	0.12
	HARSH	3.42	1.28	0.28	0.08	0.04	0.02
Switching Error	BEAGLE	0.97	1.04	1.05	1.11	1.23	1.23
	IMPUTE*	0.82	0.87	0.90	0.94	0.97	0.98
	HARSH	0.72	0.67	0.62	0.63	0.65	0.65



(a) Varying coverage for fixed read length 1,000bp (b) Varying read length for fixed coverage 4X

Figure 4.6: Performance of BEAGLE, IMPUTE* and HARSH for varying coverage and read length on HapMap

why the genotyping error rates for BEAGLE, IMPUTE* and HARSH increase when the read length increases. A possible reason is that longer reads for a fixed coverage result in fewer total reads and larger gaps without any coverage. In other words, longer reads result in less random read bits across the chromosome. An extreme example is that the gap will be half of the genome on average if the read length is equal to the genome size and coverage is 1X. Sequentially, larger gap where no reads cover will potentially harm the imputation and haplotype phasing accuracy. But we can still see that the performance gap between BEAGLE or IMPUTE* and HARSH is enlarged while the read length increases. This is attributed to the ability of HARSH to leverage the multi-SNP information in longer reads. In Table 4.3, we show that the improvement of HARSH over BEAGLE and IMPUTE*. The improvement is basically from the reduced switching error, which is reduced from 0.62% to 0.48% by HARSH but not by IMPUTE*. The genotyping error for both methods increase at the same pace, due to the larger gaps caused by longer reads. The error rates for BEAGLE, IMPUTE* and HARSH increase from 0.59% to 0.79%, from 0.56% to 0.85% and from 0.28% to 0.48%, respectively,



(a) Varying coverage for fixed read length 100 bp

(b) Varying read length for fixed coverage 8X

Figure 4.7: Performance of IMPUTE* and HARSH for varying coverage and read length on 1000 genomes

when the read length increases from 1,000 bp to 4,000 bp. But HARSH consistently performs better than BEAGLE and IMPUTE even while the genotyping error rate is increasing.

Table 4.3: Genotyping and switching errors (%) for varying read lengths on HapMap data set. Coverage is fixed to be 4X.

Read Length		1,000bp	2,000bp	3,000bp	4,000bp
Genotyping Error	BEAGLE	0.59	0.67	0.74	0.79
	IMPUTE*	0.56	0.70	0.77	0.85
	HARSH	0.28	0.37	0.40	0.48
Switching Error	BEAGLE	1.05	1.10	1.07	1.07
	IMPUTE*	0.90	0.93	0.94	0.94
	HARSH	0.62	0.57	0.49	0.48

4.3.3 1000 Genomes Simulations

The 1000 Genomes project is an on-going project that uses high throughput sequencing technology to collect the genetic variant data across many individuals with the goal of characterizing rare variants which are not present in HapMap. This provides us the opportunity to evaluate our method using simulations which will realistically capture the distributions of rare variants and more accurately reflect a tubal performance. We simulate realistic paired end reads, which have 100 bp for each end, and a gap size following a normal distribution with 100 bp mean and 10 bp standard deviation. Only 22% reads contain only one SNP and around 55% reads contain more than 3 SNPs. On average, every read covers around 3.1 SNPs. Following the same settings as what we did for HapMap data, we test HARSH for different coverages and read lengths. The results for coverage 1X, 2X, 4X, 8X, 16X and 32X are shown in Figure 4.7(a). We observe that the error rate does not further drop after coverage 8X. At coverage 8X, the improvement of HARSH over IMPUTE* is 29% from 0.021 to 0.015. Thus for fixed coverage 8X, we simulate pair-end reads with 100 bp, 200 bp, 300 bp and 400 bp in each end. The results are shown in Figure 4.7(b). We observe that, HARSH, unlike IMPUTE*, benefits from using longer reads as they contains more multi-SNP reads than shorter reads. Thus, as expected, the performance gap between IMPUTE* and HARSH increases as the read length increases. However, in Figure 4.7(b) we do not see that the error rate increases when the read length increases as in Figure 4.6(b). A possible reason is that the SNPs are much denser in 1000 genomes data than HapMap data, and we simulated much shorter reads for 1000 genomes data. Thus, the gap caused by 400 bp read length would be much shorter than previous $4,000$ bp read length for HapMap data set. The reference haplotype panel could well take advantage of Linkage Disequilibrium (LD) effect to recover those gaps. Therefore, the error rate for IMPUTE* keeps almost the same for different read lengths but our method HARSH reduces the error rate by incorporating more multi-SNP read information when the read length increases.

4.4 Discussions

Haplotype phasing plays an important role in a wide variety of genetic applications. Although it is possible to determine haplotypes using laboratory-based experimental techniques, these approaches are expensive and time-consuming. Recently, [KMA11] were able to generate the complete phased sequence of a Gujarati individual using a Fosmid library. Unfortunately this method is not easily scalable to phasing more than one individual. Thus, the need for a practical computational method for haplotype phasing remains.

We have presented HARSH, an efficient method that combines multi-SNP read information with reference panels of haplotypes for improved genotype and haplotype inference in sequencing data. Unlike previous phasing methods that utilize read counts at each SNP as input, our method takes into account the information from reads spanning multiple SNPs. Following a novel sampling method based on Gibbs sampling, HARSH is able to efficiently sample the posterior distribution of the probabilistic model given the sequencing data and a reference panel. Thus, HARSH is able to efficiently find the likely haplotypes in terms of the marginal probability over the genotype data. Using simulations from HapMap and 1000 Genomes data we show that our method achieves superior accuracy than existing approaches with decreased computational requirements. In addition we evaluate our method as function of coverage and read length showing that our method continues to improve as read length and coverage increases.

CHAPTER 5

A Spatial-Aware Haplotype Copying Model

5.1 Motivation

Complex population demography coupled with the presence of recombination hotspots have shaped genetic variation in the human genome into blocks of markers with similar recent ancestry [GBH03, CAA10, DRS01]. This recent ancestry sharing induces dependencies among variants in the form of linkage disequilibrium (LD), i.e. the non-random association of alleles at two or more loci [Kru99]. Therefore, the observed LD patterns across the genome are the result of a population's demographic history and are modeled in a wide-range of problems from population genetic inferences [LBC09, PHJ10] to medical population genetics [MHM07, LWD10]. Most notably, LD has enabled the era of genome-wide association studies that use a small number of variants (as compared to all variation in the genome) to assay variation across the entire human genome [dYP05]. Thus, modeling population LD is a fundamental problem in computational genetics with applications ranging from genotype imputation and haplotype inference to locus-specific and genome-wide ancestry inference [MHM07, HDM09, HFS12, CKW13, SMW13, PSK09, PTP09].

Although many approaches for modeling LD have been proposed [DRS01, LS03], one of the most successful framework has been introduced by Li and Stephens (widely referred to as the *haplotype copy model* [LS03]). Drawing on coalescent theory, in this model, a haplotype sampled from a population is viewed as a mosaic of segments of previously sampled haplotypes. This mosaic structure can be efficiently modeled within a hidden Markov model to achieve very accurate solutions to many genetic problems such as genotype imputation [MHM07, HDM09, HFS12], mapping admixed populations [PSK09, PTP09], quality control in genome-wide association studies [HKE09], detection of identity by descent (IBD) segments [Bro06, BB10], calculating the recombination rates [WKV11], haplotype phasing [DMZ12], migration rates [RS07] and calling of genotypes at low coverage sequencing [PRM12, LSK11].

At the core of the Li and Stephens [LS03] model lies a hidden Markov model (HMM) that emits haplotypes through a series of segmental copies from the pool of previously observed haplotypes. The hidden states in the HMM indicate which haplotype from the reference panel to copy from while emission probabilities allow for potential mutation events observed since the most recent common ancestor of the target and the reference copy haplotype. Recombination events are modeled through the transition probabilities; the probability of copying from the same reference haplotype at successive loci is much higher than switching to another haplotype, based on the idea the probability of having a recombination between two neighboring loci is low. Motivated by coalescent theory in randomly mating populations, the a priori probability of switching the copy process to another haplotype is equally likely among all the previously observed haplotypes. However, since human populations show a tremendous amount of structure across geography [NJB08, YNE12, BQC13] (inline with isolation-by-distance models), it is likely that haplotypes physically closer in geography to the target haplotype contribute significantly more to the copy process. Furthermore, with the emergence of high-throughput sequencing that is generating massive amount of data [Mar08, Sch08, SMV04], existing methods are increasingly computationally intensive due to the ever larger samples of haplotypes that can be used as reference. Although a commonly used approach for reducing computational burden is to downsample the reference panels [HMS11, PAG10, LLW13] (often in an ad-hoc manner) a principled approach for selection of a reference panel for optimizing performance is currently lacking.

In this paper, we propose a new approach to modeling genetic variation in structured populations that incorporates ideas from both the haplotype copying model [LS03] and the spatial structure framework that models genetic variation as function of geography [YNE12, BQC13]. That is, we propose a haplotype copy model that a priorly up weights the contribution of haplotypes closer in geographical distance to the copying process. We accomplish this by jointly modeling geography and the copying process. Each haplotype is associated with a geographical position; when copying into a new haplotype with known location, we instantiate an HMM that has switching transition probabilities up weighted for haplotypes closer in geographical space to target haplotype.

We use real data from the 1000 Genomes project [CAA10] to show that the our spatial-aware approach fits the data significantly better than the standard model. Through a masking procedure followed by a leave-one-out experiment we show that our spatial-aware method significantly increases imputation accuracy especially for lower frequency variation (e.g. an improvement of 6%(2%) for low-frequency(common) variation in Asian data). We also show that our approach can be used to select a small personalized reference panel for imputation that increases imputation accuracy while significantly reducing imputation runtime (up to 10-fold). Finally, we show how our model can be used in a supervised manner to infer locations on the genetic-geographic map for individuals based on their genetic data.

5.2 Methods

5.2.1 The standard haplotype copying model

We start by briefly introducing the standard haplotype copying model [LS03] for modeling LD in a population. Let $H \in \{0,1\}_{N \times L}$ be a matrix of haplotypes (which we will refer to as *reference panel*), where $h_{ij} \in \{0,1\}$ indicates if the i-th individual at the j-th position (SNP) contains the reference or the alternate allele. N denotes the number of haplotypes in the reference panel and L the number of SNPs in the data. Let $h \in \{0,1\}_{1 \times L}$ be a multi-locus haplotype which we will refer to as the *target haplotype* where $h_i \in \{0,1\}$ indicates the i-th SNP. The haplotype copy model views the target haplotype as being composed of a mosaic of segments from haplotypes of the reference panel.

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

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

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

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

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

The likelihood of the target haplotype h is defined as:

$$P(h|S, H; \lambda) = P(S) \prod_k P(h_k|s_k, H) = \left(\prod_k \tau_k(s_{k-1}, s_k)\right) \left(\prod_k \omega(h_k, s_k; H)\right) \tag{5.1}$$

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

5.2.2 A spatial-aware haplotype copying model

A drawback of the standard haplotype copying model comes from the equal treatment of reference haplotypes; that is, a priori all haplotypes from the reference panel are equally likely to contribute to the target haplotype. This effect motivates us to propose the following approach to take spatial effect into account in the model. Let $X = \{x_1, \dots, x_N\}$ indicate the locations for all N reference haplotypes and x indicate the location for target haplotype. In a scenario where the location of the individuals are not known, we estimate their locations from genotype data using methods such as PCA [NJB08], SPA [YNE12] or LOCO-LD [BQC13]. Then, instead of using uniform switching probability across all reference haplotypes, we assign higher probability to haplotypes located closer to the target haplotype.

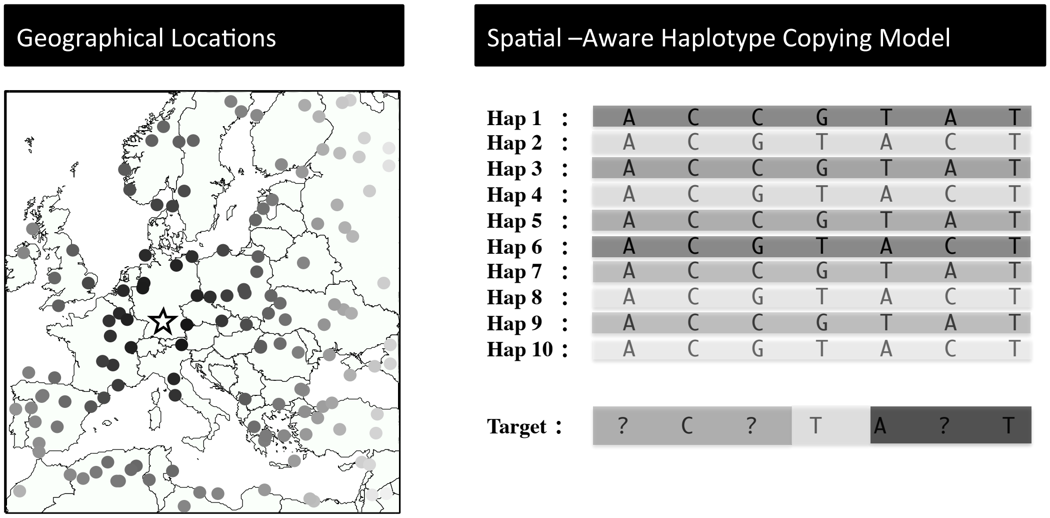


Figure 5.1: An illustration of spatial haplotype copying model. In the left panel, the location for target haplotype is shown using the star. All haplotypes in the data are color coded using the distance to the target location (light more distant, darker are closer). We enforce the transition rates (that encode the copy switching) to give higher weight to haplotypes closer to the target haplotype. A haplotype at the target location is more likely to contain mosaic segments from haplotypes that are closer to the target location.

Formally, we redefine the transition rate τ between SNP k and SNP $k+1$ as:

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

The function $\psi(x, x_j)$ denotes a distance function between x and x_j (e.g. Euclidean distance) and Z is a normalization factor to ensure the probability definition. The parameter λ specifies the effect of geographical distance. It is worth mentioning that this spatial-aware model can be reduced to standard haplotype copying model by setting $\lambda = 0$, such that $p_j = 1/N$; therefore our approach can be viewed as a generalization of the standard Li and Stephens model. An illustration of our model is shown in Figure 5.1. Intuitively a large value for λ indicates a more pronounced spatial effect (less probability to copy from distant haplotypes), while

$\lambda = 0$ reverts to assigning equal a priori probability.

The likelihood of the target haplotype is defined as before by summing on all paths in the model (Eq 5.1). Inference in this model can be performed as in the standard haplotype copy model using a combination of Viterbi and posterior decoding as function of the particular application.

5.2.3 Estimation of spatial effect parameter λ

A pre-requisite step in applying our model is the specification of λ. It is necessary to estimate the λ before using the model for various applications, as the value of λ could vary significantly across individuals or populations. We estimate λ through maximum likelihood estimation (MLE). Starting from the likelihood of the target haplotype h (Eq 5.1), we marginalize over all possible values of hidden variables S to obtain likelihood as function of λ:

$$L(h; \lambda) = \sum_{S} P(h|S, H) \tag{5.2}$$

However, this overall likelihood function is infeasible to optimize directly, as the number of all possible values of S is very large L^N. Although the likelihood computation can be reduced by forward-backward algorithm to $O(NL)$, the gradient is still very expensive to compute, as the calculation would involve a forward-backward in $O(NL)$ and a summation of $O(N^2L)$ terms. When the number of reference haplotypes is large, this gradient would be infeasible to compute. Fortunately, the gradient for the Q function in EM algorithm is much simpler to compute than the gradient of likelihood function in (5.2). It is also guaranteed that the gradient of the Q function will be an increasing direction for the original likelihood function, which is a theoretical property of the EM algorithm. Thus, we resort to compute the gradient of the Q function instead of the gradient of original likelihood function.

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

$$\begin{aligned} Q(\lambda, \lambda^{(t)}) &= \sum_S P(S) \ln P(h, S; \lambda) \\ &\propto \sum_{kij} P(s_{k-1} = i, s_k = j; \lambda^{(t)}) \ln \tau_k(i, j; \lambda) \end{aligned} \tag{5.3}$$

The gradient for this Q function can be calculated as follows

$$\frac{\partial Q}{\partial \lambda} = -\sum_{kij} P(s_{k-1} = i, s_k = j; \lambda^{(t)}) \left(\frac{\psi(x, x_j) - \sum_l \psi(x, x_l) p_l}{1 + I(i = j)\left(\frac{\theta_k}{(1-\theta_k)p_j}\right)} \right) \tag{5.4}$$

where the identity function $I(i = j)$ is equal to 1 when $i = j$ and 0 otherwise. However, simply calculation of this gradient will also be inefficient with the complexity $O(N^2L)$, which is still expensive for thousands of reference haplotypes and millions of SNPs. We resort to computing a stochastic gradient for the Q function, and apply it to the original likelihood function as a searching direction. We estimate the gradient by sampling over N haplotypes, instead of enumerating all of them. In practice, between each pair of SNP k and SNP $k+1$, we randomly sample 1000 pairs of $s_{k-1} = i$ and $s_k = j$, instead of all N^2 pairs. The overall algorithm for efficient optimization of the spatial effect parameter λ is described in Algorithm 3.

5.2.4 Localization of individuals based on their genetic data

Another appealing application for spatial-aware haplotype copying model is to localize individuals on the map. That is, given locations X for all reference panel haplotypes, we seek to find the best location x for the target haplotype to maximize the likelihood of the data. The algorithm follows similar procedure as above section 5.2.3. The difference mainly comes from a different Q function as follows

$$\begin{aligned} Q(x, x^{(t)}) &= \sum_S P(S) \ln P(h, S; x) \\ &\propto \sum_{kij} P(s_{k-1} = i, s_k = j; x^{(t)}) \ln \tau_k(i, j; x) \end{aligned} \tag{5.5}$$

Algorithm 3 Learning Algorithm for Parameter λ Estimation

```
Setting optimization parameters $R$ and $C$ (e.g., $R = 1 \times 10^3$ and $C = 20$)
Pre-computing $\psi(x, x_j)$ for all reference haplotype $j$, and $\theta_k$ for all $k$.
Randomly initialize $\lambda^{(0)} > 0$
for $t$ from 0 to $T$ do
    Perform forward-backward algorithm to get the forward/backward probability
    Compute stochastic gradient $g(\lambda^{(t)})$ by sampling $R$ pairs of $i$ and $j$ in (5.4)
    Setting $\lambda^{(t+1)} = \lambda^{(t)} + \frac{1}{t+C} \cdot g(\lambda^{(t)})$
end for
Output $\lambda^{(T+1)}$
```

which is parameterized by x instead of λ as in Equation (5.3). However, this change leads to non-concavity of the function in general. But since there is only one parameter to estimate, and the function is well behaved in practice, we can still compute the gradient for the Q function and apply it to the stochastic gradient descent method. The gradient for the Q function in Equation (5.5) can be calculated as follows

$$\frac{\partial Q}{\partial x} \quad = \quad -\sum_{kij} P(s_{k-1} = i, s_k = j; x^{(t)})\lambda \left(\frac{\frac{\partial \psi(x, X_j)}{\partial x} - \sum_l p_l \cdot \frac{\partial \psi(x, X_j)}{\partial x}}{1 + I(i = j)\left(\frac{\theta_k}{(1-\theta_k)p_j}\right)} \right) \tag{5.6}$$

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

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

The overall algorithm is similar as Algorithm 3 for optimizing λ, except for replacement of λ by x and the gradients correspondingly.

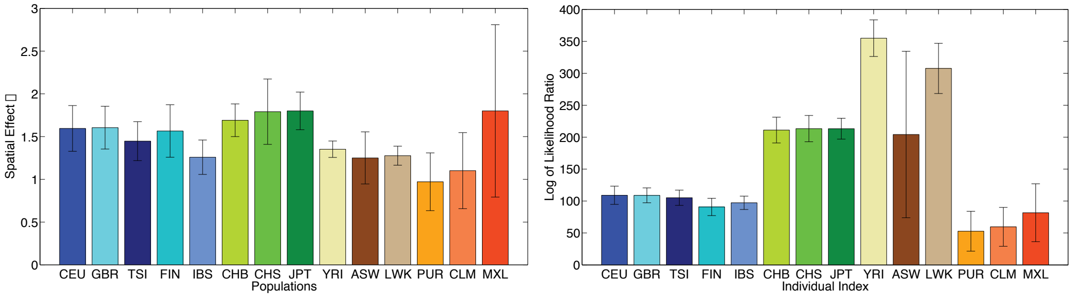


Figure 5.2: Estimated spatial copying effects λ^* across different populations in 1000 Genomes data. Left shows the average λ^* across all individuals in a given population while right displays the log likelihood ratio of the model with λ^* as compared to $\lambda = 0$. The error bars indicate the standard deviations for each population.

5.3 Experimental results

5.3.1 Estimation of spatial copying effect in the 1000 Genomes data

We applied our methods to data generated part of the 1000 Genomes project [CAA10]. A total of 1092 individuals were collected from 14 populations across the European, Asian, African and American continents. For all of our simulations we used $157,827$ SNPs on chromosome 22, where 79.5% of SNPs are rare SNPs (allele frequency < 0.05), and the rest 20.5% are common SNPs; although the original data contained $473,481$ SNPs, for computational efficiency we down sampled to every third SNP. Among the considered SNPs, we assumed that only $2,931$ SNPs present on the Affymetrix 6.0 SNP array are collected and the remaining SNPs will be imputed using our model. This amounts to using 1.86% SNPs to impute the rest 98.14% SNPs. We apply PCA [NJB08] to assign a geographical location to each individual in the dataset. Although we note that the imputation performance can be further improved if denser SNPs are assumed to be typed, we expect the general trends reported below to maintain.

Starting from the $2,931$ SNPs, we estimated the spatial effect parameter λ for each of the $2,184$ haplotypes in the dataset. The average λ values are 1.54, 1.76, 1.30 and 1.32 for European, Asian, African and American populations, respectively (Figure 5.2). Generally, the higher value of λ corresponds to stronger spatial copying effect, which leads to more segments copied from nearby haplotypes. To test the significance of spatial effect, we compared the likelihoods of the data (the 2,184 haplotypes) within the model assuming no spatial effect ($\lambda = 0$) versus the model with spatial effect (λ^* estimated from the data). The log likelihood ratio between spatial haplotype copying model and standard haplotype model is given in Figure 5.2. The likelihood is computed for each haplotype being emitted from the rest of haplotypes. Across all populations we observe that the model with a spatial effect fits the data much better than the model with no spatial assignment. This is expected since we use haplotypes across all continents (except the target) in the reference panel, and it is expected that haplotypes share more continental-specific segments.

5.3.2 Spatial-aware model improves imputation accuracy

Having established that the model with spatial effect fits the data much better than the standard model with no spatial effect, we focused next on haplotype imputation (a standard approach in genome-wide association studies through prephasing [HFS12]). We carry out a leave-one-out procedure to perform the evaluation. In each round, we select one haplotype as a target and use the rest as the reference panel. To remove potential bias, instead of using all haplotypes, we randomly select one haplotype from each individual to use a total of $1,092$ haplotypes (i.e. each round imputes one haplotype from the remaining $1,091$). The imputation results are evaluated using the average per-SNP r^2 correlation coefficients averaged across all leave-one-out rounds for either all haplotypes, or for data within each population.

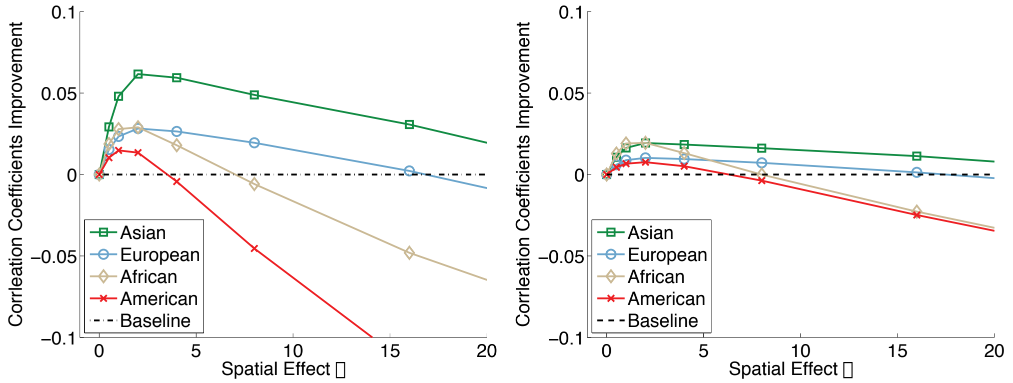


Figure 5.3: Effect of spatial copying parameter λ on imputation accuracy. Left shows results for low-frequency $(1 - 5\%)$ while right displays results for common variants $(> 5\%)$. The maximum accuracy is attained at a $\lambda \approx 2$, close to the maximum likelihood estimate for λ (1.3 to 1.7, see Section 3.1).

We first demonstrate the effect of the lambda parameter on imputation accuracy by applying our model using a wide range of lambda parameter values. Compared with the baseline method $(\lambda = 0)$, we observe that a clear improvement is obtained for a value of λ around 2, especially for European and Asian populations (see Figure 5.3). This is consistent with the spatial model fitting those populations (see Figure 5.2). We also observe that the spatial model improves the imputation of rare variants more significantly than common variants, which is expected as the rare variants are more clustered geographically [NWE12]. Moreover, the improvement for Asian and European populations is larger than for African and American populations.

Although we have shown that spatial model improves accuracy, in practice the value of λ is unknown and needs to be estimated from the data. We re-assessed the accuracy of our approach by not setting λ to pre-specified values but by estimating it from the data. The performance of the model using the maximum likelihood λ^* over baseline method is given in Table 5.1. As before, we observe a larger improvements for rare variants than common variants. A plausible explanation

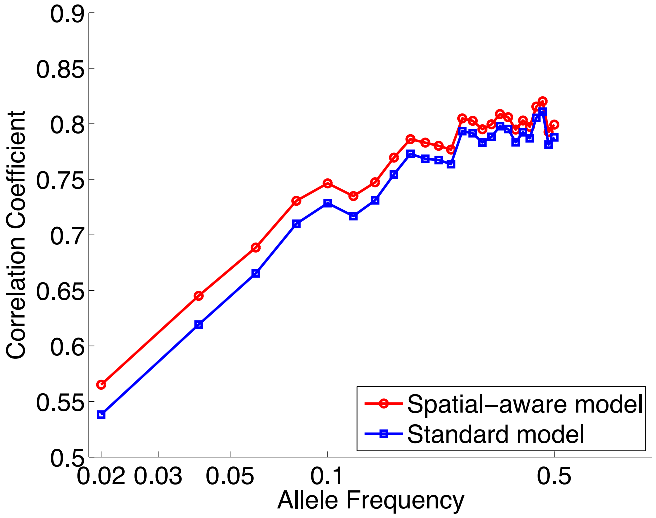


Figure 5.4: Absolute imputation improvement across all spectrum of allele frequencies. Spatial-aware model uses $\lambda*$ inferred from the data.

for this effect is that that rare variants are more clustered in geography [NWE12] than common variants. Overall for all populations, the improvement is highly correlated with allele frequency. The trend is shown in Figure 5.4, where we can see that the improvement is higher for SNPs with lower allele frequency.

Table 5.1: Performance of spatial model compared to the standard model

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

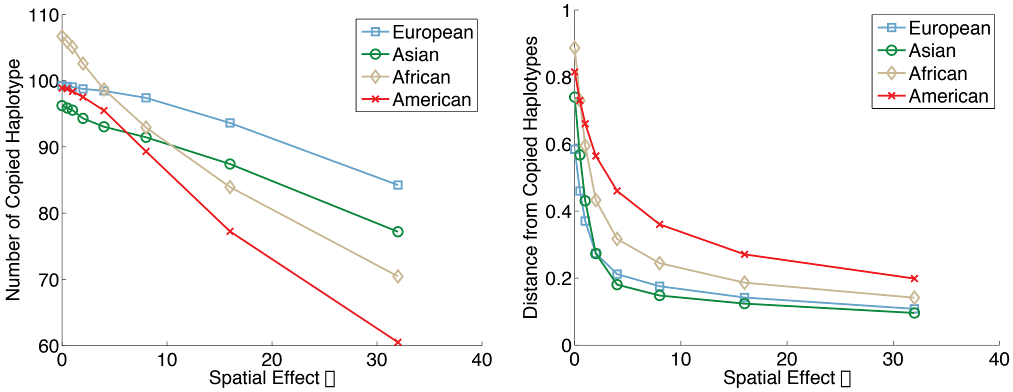


Figure 5.5: Spatial effect on copied haplotypes from reference. Left shows that the number of copied haplotypes decreases while the spatial effect parameter is larger. Right shows that the averaged distance from copied haplotype decreases while the spatial effect parameter is larger.

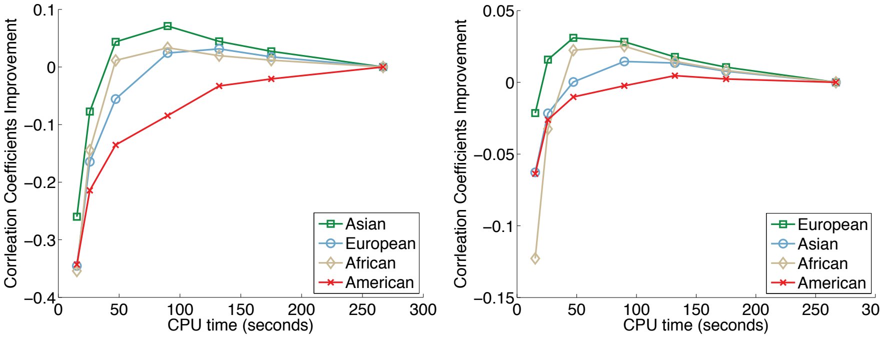


Figure 5.6: Imputation accuracy versus computational time. Left shows low-frequency variants (1-5%) while right shows results over common variants ($> 5\%$).

5.3.3 Selection of a personalized reference panel for imputation to increase performance

Inspired by the significant spatial haplotype copying effect in experiments, we hypothesized that imputation efficiency can be improved by only using a personalized reference panel composed only from geographically close haplotypes [PAG10, HMS11]. First, we expect that most of the reference haplotypes are not contributing haplotype segments to target haplotype. In Figure 5.5, we observe that the number of copied haplotypes decreases with higher λ (e.g. an average of 100 haplotypes are used in the copy process of a new target among 1091 reference haplotypes). On the other hand, in Figure 5.5, we plot the distance of those useful reference haplotypes from the target haplotype, weighted by the posterior. We observe there is a significant decrease of haplotype copying distance for higher λ value. It strongly suggests that the haplotype copying model can be significantly sped up by only keeping a small number of nearby haplotypes as reference panel. To assess this scenario, we re-imputed the target data using gradual decreasing sizes for the reference panel (1091, 800, 600, 400, 200, 100 and 50) where we only keep the most nearby haplotypes in geographical space. The relation between imputation correlation coefficients and computational CPU time is shown in Figure 5.6. We observe that the computational time can be improved linearly in the size of reference panel but the imputation performance is also improved even using less number of reference haplotypes. For rare variants, the best imputation performance is obtained at 400 haplotypes and for common variants, the best imputation performance is obtained at 200 haplotypes.

5.3.4 Localization of individuals on a map

Finally, we explored whether we can use our approach to infer the location on the map of a new individual given data of individuals with known locations. We localized individual haplotypes using spatial-aware copying model with optimal λ value estimated before assuming known locations for the rest of the haplotype data. That is, in each round, we apply spatial-aware model to infer the optimal x^* for one individual using all other other individuals as reference panel (PCA was used to infer locations for the reference panel). We observe that spatial-aware model is able to well identify individual locations, in terms of the clear separating of different continents (see Figure 5.7). We observe a high correlation coefficient between the PCA and our inferred geographical ($r = 0.87$), thus showing that our

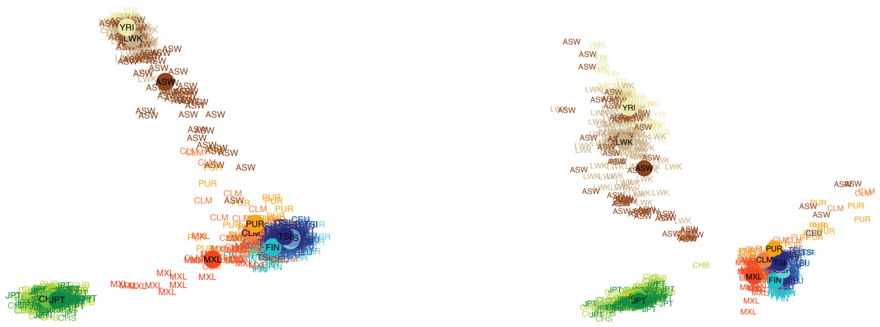


Figure 5.7: Left shows results of PCA on chromosome 22 of the 1000 Genomes data while right shows results of our leave one out procedure to localize 1000 Genomes individuals.

approach can potentially be used to localize individuals on a map given training data with known locations (see Figure 5.7) .

5.4 Discussion

The haplotype copying model plays an important role in a wide variety of genetic applications. A major drawback is that the model assumes that all haplotypes in the reference panel equally contribute a priori to the observed haplotype. In this paper, we have proposed a spatial-aware haplotype copying model that takes the spatial effects into account. We have also presented a highly efficient algorithm to estimate the spatial effect parameter before using the proposed model. We applied the proposed model to the 1000 genomes data set for several applications. First, we estimate the likelihood ratio between the spatial-aware model and spatial-unaware model, and a significant improvement is observed. Second, we test the application of imputation using spatial-aware model and obtain significant improvement over standard model. Finally, we apply this model to localize individuals and the results indicate high accuracy can be obtained.

CHAPTER 6

Conclusion and Future Work

6.1 Summary and Conclusion

In this dissertation, I have presented several probabilistic models and inference algorithms for analysis of human genetic data. The spatial ancestry analysis explicitly models the allele frequency in space and utilizes this model to place individuals in a two dimensional map or three dimensional sphere. We show that our method for localization of samples in space is slightly more accurate than principal component analysis, and importantly, unlike principle component analysis, it can be used to localize individuals of mixed ancestry in space. The spatial ancestry analysis for admixed individuals is a generalization for localization of admixed individuals and their genome blocks. An Hidden Markov model and expectation maximization algorithm is devised for efficient inference. The experimental results show that this generalization significantly improves the localization accuracy for admixed individuals. It is also the first method that enables us to localize the ancestral genome blocks into a continuum map. Moreover, I presented an efficient sampling based method for leveraging reads that span multiple single nucleotide polymorphisms for haplotype inference using sequencing data. It allows us to significantly improve standard haplotype phasing and genotype imputation methods. Finally, inspired by the spatial structure of genetic data, I presented a spatial-aware haplotype copying model, which assumes a priorly up weights the contribution of haplotypes closer in geographical distance to the copying process. We show that this novel model is able to improve both the accuracy and efficiency of genotype imputation.

I expect that the proposed methods above are able to provide more insights of genetic data when more and more population genetic data become available. The identified different spatial structure and ancestral localizations will improve our knowledge about human origins and evolutionary history. It will also help us identify genetic variants associated with human diseases, and ultimately contribute to human health.

6.2 Future Works

Rare variants compose more than 99% of human genetic variants. It is of significant interests in current genetic research, compared to common variants composing less than 1% of human genetic variants. However, I do not explicitly model rare variants in the original spatial ancestry analysis. Thus, modeling spatial distribution of rare variants and improving the ancestry localization accuracy with rare variants has been a new challenge after the spatial ancestry analysis. Apparently, using a logistic gradient is not a sufficient approximate, apart from common variants. I am looking for a generalization model from spatial ancestry analysis, which takes both common and rare variants into accounts.

Another interesting future direction is prediction of multiple phenotypes using genetic data. Ancestral localization in spatial ancestry analysis can be considered as two phenotypes, latitude and longitude. Thus, we are actually able to predict more dimensional phenotypes using the similar model as spatial ancestry analysis. It will be extremely interesting to develop a method for predicting body mass index, height, disease risk and even a computer generated human face for a given individual's genetic data.

Finally, detection of structure variants using the HARSH model is of potential interests for the future research. We can encode genome insertion and deletion using binary variables, as well as single nucleotide polymorphism. Then, a similar model as HARSH can be applied for in-del calling and phasing. It will help us collect more accurate genetic data from current sequencing technologies.

References

[AI12] Derek Aguiar and Sorin Istrail. "HapCompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data." *Journal of Computational Biology*, **19**(6):577–590, Jun 2012.

[ANL09] David H Alexander, John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals." *Genome Research*, **19**(9):1655–1664, 2009.

[BB08] Vikas Bansal and Vineet Bafna. "HapCUT: an efficient and accurate algorithm for the haplotype assembly problem." *Bioinformatics*, **24**(16):i153–159, August 2008.

[BB09] Brian L. Browning and Sharon R. Browning. "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals." *American Journal of Human Genetics*, **84**(2):210–223, 2009.

[BB10] Sharon R. Browning and Brian L. Browning. "High-resolution detection of identity by descent in unrelated individuals." *Am J Hum Genet*, **86**(4):526–539, Apr 2010.

[BHA08] Vikas Bansal, Aaron L. Halpern, Nelson Axelrod, and Vineet Bafna. "An MCMC algorithm for haplotype assembly from whole-genome sequence data." *Genome Research*, **18**(8):1336–1346, August 2008.

[BPS12] Yael Baran, Bogdan Pasaniuc, Sriram Sankararaman, Dara G. Torgerson, Christopher Gignoux, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G. Ford, Pedro C. Avila, Jose Rodriguez-Santana, Esteban Gonzlez Burchard, and Eran Halperin. "Fast and accurate inference of local ancestry in Latino populations." *Bioinformatics*, **28**(10):1359–1367, 2012.

[BQC13] Yael Baran, Ins Quintela, ngel Carracedo, Bogdan Pasaniuc, and Eran Halperin. "Enhanced localization of genetic samples through linkage-disequilibrium correction." *American Journal of Human Genetics*, **92**(6):882–894, 2013.

[Bro06] Sharon R. Browning. "Multilocus association mapping using variable-length Markov chains." *Am J Hum Genet*, **78**(6):903–913, Jun 2006.

[BSP04] 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." *American Journal of Human Genetics*, **74**:1111–1120, 2004.

[BVK10] Katarzyna Bryc, Christopher Velez, Tatiana Karafet, Andres Moreno-Estrada, Andy Reynolds, Adam Auton, Michael Hammer, Carlos D. Bustamante, and Harry 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.

[BY09] Brian L. Browning and Zhaoxia Yu. "Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies." *American Journal of Human Genetics*, **85**(6):847–861, Dec 2009.

[CAA10] 1000 Genomes Project Consortium, Gonalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. "A map of human genome variation from population-scale sequencing." *Nature*, **467**(7319):1061–1073, Oct 2010.

[CKW13] Charles C Chung, Peter A Kanetsky, Zhaoming Wang, Michelle A T Hildebrandt, Roelof Koster, Rolf I Skotheim, Christian P Kratz, Clare Turnbull, Victoria K Cortessis, Anne C Bakken, D. Timothy Bishop, Michael B Cook, R. Loren Erickson, Sophie D Foss, Kevin B Jacobs, Larissa A Korde, Sigrid M Kraggerud, Ragnhild A Lothe, Jennifer T Loud, Nazneen Rahman, Eila C Skinner, Duncan C Thomas, Xifeng Wu, Meredith Yeager, Fredrick R Schumacher, Mark H Greene, Stephen M Schwartz, Katherine A McGlynn, Stephen J Chanock, and Katherine L Nathanson. "Meta-analysis identifies four new loci associated with testicular germ cell tumor." *Nature Genetics*, **45**(6):680–685, Jun 2013.

[CPN09] Graham Coop, Joseph K. Pickrell, John Novembre, Sridhar Kudaravalli, Jun Li, Devin Absher, Richard M. Myers, Luigi L. Cavalli-Sforza, Marcus W. Feldman, and Jonathan K. Pritchard. "The role of geography in human adaptation." *PLoS Genetics*, **5**:e1000500+, 2009.

[CWD10] Graham Coop, David Witonsky, Anna Di Rienzo, and Jonathan K. Pritchard. "Using environmental correlations to identify loci underlying local adaptation." *Genetics*, **185**:1411–23, 2010.

[DHM10] Jorge Duitama, Thomas Huebsch, Gayle McEwen, Eun-Kyung Suk, and Margret R. Hoehe. "ReFHap: a reliable and fast algorithm for single individual haplotyping." In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pp. 160–169, New York, NY, USA, 2010.

[DLR77] Arthur Dempster, Nan Laird, and Donald Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B*, **39**(1):1–38, 1977.

[DMH12] J. Duitama, G. K. McEwen, T. Huebsch, S. Palczewski, S. Schulz, K. Verstrepen, E.-K. Suk, and M. R. Hoehe. "Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques." *Nucleic Acids Research*, **40**(5):2041–2053, 2012.

[DMZ12] Olivier Delaneau, Jonathan Marchini, and Jean-Franois Zagury. "A linear complexity phasing method for thousands of genomes." *Nature Methods*, **9**(2):179–181, Feb 2012.

[DRS01] Mark J. Daly, John D. Rioux, Stephen F. Schaffner, Thomas J. Hudson, and Eric S. Lande. "High-resolution haplotype structure in the human genome." *Nature Genetics*, **29**(6):229–232, Jun 2001.

[Dur10] Durbin, R. et. al. "A map of human genome variation from population-scale sequencing." *Nature*, **467**(7319):1061–1073, October 2010.

[dYP05] Paul I W. de Bakker, Roman Yelensky, Itsik Pe'er, Stacey B. Gabriel, Mark J. Daly, and David Altshuler. "Efficiency and power in genetic association studies." *Nature Genetics*, **37**(11):1217–1223, Nov 2005.

[EPF02] Wolfgang Enard, Molly Przeworski, Simon E. Fisher, Cecilia S. L. Lai, Victor Wiebe, Takashi Kitano, Anthony P. Monaco, and Svante Pääbo. "Molecular evolution of $FOXP2$, a gene involved in speech and language." *Nature*, **418**:869–72, 2002.

[ER08] Laurent Excoffier and Nicolas Ray. "Surfing during population expansions promotes genetic revolutions and structuration." *Trends in Ecology & Evolution*, **23**:347–351, 2008.

[FD01] P. Fearnhead and P. Donnelly. "Estimating recombination rates from population genetic data." *Genetics*, **159**(3):1299–1318, Nov 2001.

[FSP03] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies." *Genetics*, **164**(4):1567–1587, 2003.

[GBH03] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, et al. "The International HapMap Project." *Nature*, **426**(6968):789–796, Dec 2003.

[GG84] Stuart Geman and Donald Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6):721–741, 1984.

[HCP10] Dan He, Arthur Choi, Knot Pipatsrisawat, Adnan Darwiche, and Eleazar Eskin. "Optimal algorithms for haplotype assembly from whole-genome sequence data." *Bioinformatics*, **26**(12):i183–i190, 2010.

[HCZ01] J. P. Hugot, M. Chamaillard, H. Zouali, S. Lesage, J. P. Cézard, J. Belaiche, S. Almer, C. Tysk, C. A. O'Morain, M. Gassull, V. Binder, Y. Finkel, A. Cortot, R. Modigliani, P. Laurent-Puig, C. Gower-Rousseau, J. Macry, J. F. Colombel, M. Sahbatou, and G. Thomas. "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." *Nature*, **411**(6837):599–603, May 2001.

[HDM09] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." *PLoS Genetics*, **5**(6):e1000529, Jun 2009.

[HE13] Dan He and Eleazar Eskin. "Hap-seqX: Expedite Algorithm for Haplotype Phasing with Imputation using Sequence Data." *Gene*, **518**(1):2–6, 2013.

[HFS12] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonalo R Abecasis. "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." *Nature Genetics*, **44**(8):955–959, Aug 2012.

[HHE12] Dan He, Buhm Han, and Eleazar Eskin. "Hap-seq: An Optimal Algorithm for Haplotype Phasing with Imputation Using Sequencing Data." In *Proceedings of the 16th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pp. 64–78, 2012.

[HKE09] Buhm Han, Hyun M. Kang, and Eleazar Eskin. "Rapid and Accurate Multiple Testing Correction and Power Estimation for Millions of Correlated Markers." *PLoS Genet*, **5**(4):e1000456+, April 2009.

[HMS11] Bryan Howie, Jonathan Marchini, and Matthew Stephens. "Genotype imputation with thousands of genomes." *G3: Genes, Genomes, Genetics*, **1**(6):457–470, 2011.

[HTP11] 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.

[HW09] Kent E. Holsinger and Bruce S. Weir. "Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}." *Nature Reviews Genetics*, **10**:639–650, 2009.

[Int05] International HapMap Consortium. "A haplotype map of the human genome." *Nature*, **437**(7063):1299–1320, October 2005.

[JSS08] Mattias Jakobsson, Sonja W. Scholz, Paul Scheet, Raphael J. Gibbs, Jenna M. Vanliere, Hon-Chung Fung, Zachary A. Szpiech, James H. Degnan, Kai Wang, Rita Guerreiro, Jose M. Bras, Jennifer C. Schymick, Dena G. Hernandez, Bryan J. Traynor, Javier Simon-Sanchez, Mar Matarin, Angela Britton, Joyce van de Leemput, Ian Rafferty, Maja Bucan, Howard M. Cann, John A. Hardy, Noah A. Rosenberg, and Andrew B. Singleton. "Genotype, haplotype and copy-number variation in worldwide human populations." *Nature*, **451**:998–1003, 2008.

[JSS12] 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.

[KMA11] Jacob O Kitzman, Alexandra P Mackenzie, Andrew Adey, Joseph B Hiatt, Rupali P Patwardhan, Peter H Sudmant, Sarah B Ng, Can Alkan, Ruolan Qiu, Evan E Eichler, and Jay Shendure. "Haplotype-resolved genome sequencing of a Gujarati Indian individual." *Nat Biotechnol*, **29**(1):59–63, Jan 2011.

[Kru99] L. Kruglyak. "Prospects for whole-genome linkage disequilibrium mapping of common disease genes." *Nature Genetics*, **22**(2):139–144, Jun 1999.

[KZE10] Hyun Min Kang, Noah A. Zaitlen, and Eleazar Eskin. "EMINIM: an adaptive and memory-efficient algorithm for genotype imputation." *J Comput Biol*, **17**(3):547–560, Mar 2010.

[LAT08] Jun Z. Li, Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, Gregory S. Barsh, Marcus Feldman, Luigi L. Cavalli-Sforza, and Richard M. Myers. "Worldwide human relationships inferred from genome-wide patterns of variation." *Science*, **319**:1100–1104, 2008.

[Laz01] L. C. Lazzeroni. "A chronology of fine-scale gene mapping by linkage disequilibrium." *Stat Methods Med Res*, **10**(1):57–76, Feb 2001.

[LBC09] Kirk E. Lohmueller, Carlos D. Bustamante, and Andrew G. Clark. "Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data." *Genetics*, **182**(1):217–231, May 2009.

[Liu08] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, January 2008.

[LK73] R. C Lewontin and Jesse Krakauer. "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms." *Genetics*, **74**:175–195, 1973.

[LLN08] Oscar Lao, Timothy T. Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, Jaume Bertranpetit, Laurence A. Bindoff, and David Comas. "Correlation between genetic and geographic structure in Europe." *Current Biology*, **18**:1241–1248, 2008.

[LLW13] Eric Yi Liu, Mingyao Li, Wei Wang, and Yun Li. "MaCH-Admix: Genotype Imputation for Admixed Populations." *Genetic epidemiology*, **37**(1):25–37, 2013.

[LML00] C X Liu, S Musco, N M Lisitsina, S Y Yaklichkin, and N A Lisitsyn. "Genomic organization of a new candidate tumor suppressor gene, *LRP2B*." *Genomics*, **69**:271–274, 2000.

[LMN09] Quan Long, Daniel MacArthur, Zemin Ning, and Chris Tyler-Smith. "HI: haplotype improver using paired-end short reads." *Bioinformatics*, **25**(18):2436–2437, 2009.

[LS03] Na Li and Matthew Stephens. "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data." *Genetics*, **165**(4):2213–2233, Dec 2003.

[LSK11] Yun Li, Carlo Sidore, Hyun Min Kang, Michael Boehnke, and Gonalo R. Abecasis. "Low-coverage sequencing: implications for design of complex trait association studies." *Genome Res*, **21**(6):940–951, Jun 2011.

[LWD10] Yun Li, Cristen J. Willer, Jun Ding, Paul Scheet, and Gonalo R. Abecasis. "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." *Genet Epidemiol*, **34**(8):816–834, Dec 2010.

[Mar08] Elaine R Mardis. "The impact of next-generation sequencing technology on genetics." *Trends Genet*, **24**(3):133–141, Mar 2008.

[McV09] Gil McVean. "A genealogical interpretation of principal components analysis." *PLoS Genetics*, **5**:e1000686+, 2009.

[MG03] Simon R. Myers and Robert C. Griffiths. "Bounds on the minimum number of recombination events in a sample history." *Genetics*, **163**(1):375–394, Jan 2003.

[MHM07] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. "A new multipoint method for genome-wide association studies by imputation of genotypes." *Nature Genetics*, **39**(7):906–913, Jul 2007.

[NBK08] Matthew R Nelson, Katarzyna Bryc, Karen S King, Amit Indap, Adam R Boyko, John Novembre, Linda P Briley, Yuka Maruyama, Dawn M Waterworth, Grard Waeber, Peter Vollenweider, Jorge R Oksenberg, Stephen L Hauser, Heide A Stirnadel, Jaspal S Kooner, John C Chambers, Brendan Jones, Vincent Mooser, Carlos D Bustamante, Allen D Roses, Daniel K Burns, Margaret G Ehm, and Eric H Lai. "The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research." *Am J Hum Genet*, **83**(3):347–358, Sep 2008.

[NCP11] Amidou N'Diaye, Gary K Chen, Cameron D Palmer, Bing Ge, Bamidele Tayo, Rasika A Mathias, Jingzhong Ding, Michael A Nalls, Adebowale Adeyemo, Véronique Adoue, et al. "Identification, replication, and fine-mapping of loci associated with adult height in individuals of african ancestry." *PLoS genetics*, **7**(10):e1002298, 2011.

[ND09] John Novembre and Anna Di Rienzo. "Spatial patterns of variation due to natural selection in humans." *Nature Review Genetics*, **10**:745:755, 2009.

[NJB08] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltn Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, and Carlos D. Bustamante. "Genes mirror geography within Europe." *Nature*, **456**(7218):98–101, 2008.

[NS08] John Novembre and Matthew Stephens. "Interpreting principal component analyses of spatial population genetic variation." *Nature Genetics*, **40**:646–649, 2008.

[NW00] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization.* Springer, New York, USA, 2000.

[NWE12] Matthew R. Nelson, Daniel Wegmann, Margaret G. Ehm, Darren Kessner, Pamela St. Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, Liling Warren, Jennifer Aponte, Matthew Zawistowski, Xiao Liu, Hao Zhang, Yong Zhang, Jun Li, Yun Li, Li Li, Peter Woollard, Simon Topp, Matthew D. Hall, Keith Nangle, Jun Wang, Gonçalo Abecasis, Lon R. Cardon, Sebastian Zöllner, John C. Whittaker, Stephanie L. Chissoe, John Novembre, and Vincent Mooser. "An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People." *Science*, **337**(6090):100–104, July 2012.

[PAG10] Bogdan Paaniuc, Ram Avinery, Tom Gur, Christine F. Skibola, Paige M. Bracci, and Eran Halperin. "A generic coalescent-based framework for the selection of a reference panel for imputation." *Genetic Epidemiology*, **34**(8):773–782, 2010.

[PCL13] Minoli A Perera, Larisa H Cavallari, Nita A Limdi, Eric R Gamazon, Anuar Konkashbaev, Roxana Daneshjou, Anna Pluzhnikov, Dana C Crawford, Jelai Wang, Nianjun Liu, et al. "Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study." *The Lancet*, **382**(9894):790–796, 2013.

[PCN09] Joseph K. Pickrell, Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z. Li, Devin Absher, Balaji S. Srinivasan, Gregory S. Barsh, Richard M. Myers, Marcus W. Feldman, and Jonathan K. Pritchard. "Signals of recent positive selection in a worldwide sample of human populations." *Genome Research*, **19**:826–837, 2009.

[PHJ10] John E. Pool, Ines Hellmann, Jeffrey D. Jensen, and Rasmus Nielsen. "Population genetic inference from genomic sequence variation." *Genome Res*, **20**(3):291–300, Mar 2010.

[PPP06] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. "Principal components analysis corrects for stratification in genome-wide association studies." *Nature genetics*, **38**(8):904–909, 2006.

[PRM12] Bogdan Pasaniuc, Nadin Rohland, Paul J. McLaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, Benjamin M. Neale, Mark J. Daly, Pamela Sklar, Patrick F. Sullivan, Sarah Bergen, Jennifer L. Moran, Christina M. Hultman, Paul Lichtenstein, Patrik Magnusson, Shaun M. Purcell, David W. Haas, Liming Liang, Shamil Sunyaev, Nick Patterson, Paul I. W. de Bakker, David Reich, and Alkes L. Price. "Extremely low-coverage sequencing and imputation increases power for genome-wide association studies." *Nature Genetics*, **44**(6):631–635, 2012.

[PSD00] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics*, **155**:945–959, 2000.

[PSK09] Bogdan Pasaniuc, Sriram Sankararaman, Gad Kimmel, and Eran Halperin. "Inference of locus-specific ancestry in closely related populations." *Bioinformatics*, **25**(12):i213–i221, Jun 2009.

[PST13] Bogdan Pasaniuc, Sriram Sankararaman, Dara G. Torgerson, Christopher Gignoux, Noah Zaitlen, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G. Ford, Pedro C. Avila, Jose Rodriguez-Santana, Gary K. Chen, Loic Le Marchand, Brian Henderson, David Reich, Christopher A. Haiman, Esteban Gonzlez Burchard, and Eran Halperin. "Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation." *Bioinformatics*, **29**:1407–1415, 2013.

[PTP09] A.L. Price, A. Tandon, N. Patterson, K.C. Barnes, N. Rafaels, I. Ruczinski, T.H. Beaty, R. Mathias, D. Reich, and S. Myers. "Sensitive detection of chromosomal segments of distinct ancestry in admixed populations." *PLoS genetics*, **5**(6):e1000519, 2009.

[PZR10] A.L. Price, N.A. Zaitlen, D. Reich, and N. Patterson. "New approaches to population stratification in genome-wide association studies." *Nature Reviews Genetics*, **11**(7):459–463, 2010.

[RDS01] J. D. Rioux, M. J. Daly, M. S. Silverberg, K. Lindblad, H. Steinhart, Z. Cohen, T. Delmonte, K. Kocher, K. Miller, S. Guschwan, E. J. Kulbokas, S. O'Leary, E. Winchester, K. Dewar, T. Green, V. Stone, C. Chow, A. Cohen, D. Langelier, G. Lapointe, D. Gaudet, J. Faith, N. Branco, S. B. Bull, R. S. McLeod, A. M. Griffiths, A. Bitton, G. R. Greenberg, E. S. Lander, K. A. Siminovitch, and T. J. Hudson. "Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease." *Nat Genet*, **29**(2):223–228, Oct 2001.

[RS07] Arindam Roychoudhury and Matthew Stephens. "Fast and accurate estimation of the population-scaled mutation rate, theta, from microsatellite genotype data." *Genetics*, **176**(2):1363–1366, Jun 2007.

[Sch08] Stephan C Schuster. "Next-generation sequencing transforms today's biology." *Nature Methods*, **5**(1):16–18, Jan 2008.

[SMV04] Jay Shendure, Robi D Mitra, Chris Varma, and George M Church. "Advanced sequencing technologies: methods and goals." *Nat Rev Genet*, **5**(5):335–344, May 2004.

[SMW13] Sharon A Savage, Lisa Mirabello, Zhaoming Wang, Julie M Gastier-Foster, Richard Gorlick, Chand Khanna, Adrienne M Flanagan, Roberto Tirabosco, Irene L Andrulis, Jay S Wunder, Nalan Gokgoz, Ana Patio-Garcia, Luis Sierrasesmaga, Fernando Lecanda, Nilgn Kurucu, Inci Ergurhan Ilhan, Neriman Sari, Massimo Serra, Claudia Hattinger, Piero Picci, Logan G Spector, Donald A Barkauskas, Neyssa Marina, Silvia Regina Caminada de Toledo, Antonio S Petrilli, Maria Fernanda Amary, Dina Halai, David M Thomas, Chester Douglass, Paul S Meltzer, Kevin Jacobs, Charles C Chung, Sonja I Berndt, Mark P Purdue, Neil E Caporaso, Margaret Tucker, Nathaniel Rothman, Maria Teresa Landi, Debra T Silverman, Peter Kraft, David J Hunter, Nuria Malats, Manolis Kogevinas, Sholom Wacholder, Rebecca Troisi, Lee Helman, Joseph F Fraumeni, Meredith Yeager, Robert N Hoover, and Stephen J Chanock. "Genome-wide association study identifies two susceptibility loci for osteosarcoma." *Nature Genetics*, **45**(7):799–803, Jul 2013.

[SPP11] M.F. Seldin, B. Pasaniuc, and A.L. Price. "New approaches to disease mapping in admixed populations." *Nature Reviews Genetics*, **12**(8):523–528, 2011.

[SRH02] Pardis C. Sabeti, David E. Reich, John M. Higgins, Haninah Z P. Levine, Daniel J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, Jill V. Platko, Nick J. Patterson, Gavin J. McDonald, Hans C. Ackerman, Sarah J. Campbell, David Altshuler, Richard Cooper, Dominic Kwiatkowski, Ryk Ward, and Eric S. Lander. "Detecting recent positive selection in the human genome from haplotype structure." *Nature*, **419**(6909):832–837, Oct 2002.

[Vit06] A. Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm." *IEEE Transaction on Information Theory*, **13**(2):260–269, 2006.

[VKW06] Benjamin F. Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K. Pritchard. "A map of recent positive selection in the human genome." *PLoS Biology*, **4**:e72+, 2006.

[WKV11] Daniel Wegmann, Darren E Kessner, Krishna R Veeramah, Rasika A Mathias, Dan L Nicolae, Lisa R Yanek, Yan V Sun, Dara G Torgerson, Nicholas Rafaels, Thomas Mosley, Lewis C Becker, Ingo Ruczinski, Terri H Beaty, Sharon L R Kardia, Deborah A Meyers, Kathleen C Barnes, Diane M Becker, Nelson B Freimer, and John Novembre. "Recombination rates in admixed individuals identified by ancestry-based inference." *Nature Genetics*, **43**(9):847–853, Sep 2011.

[WMB07] Samuel K Wasser, Celia Mailand, Rebecca Booth, Benezeth Mutayoba, Emily Kisamo, Bill Clark, and Matthew Stephens. "Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban." *Proceedings of the National Academy of Sciences*, **104**(10):4228–4233, 2007.

[XWJ12] Minzhu Xie, Jianxin Wang, and Tao Jiang. "A fast and accurate algorithm for single individual haplotyping." *BMC Systems Biology*, **6**(Suppl 2):S8, 2012.

[YNE12] 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, Jun 2012.