A Markov random field model for inferring population structure

Jeffrey Herman

May 4, 2012

Contents

Abstract

Assessment of population structure is a crucial component of conducting accurate genetic association studies. While many methods exist to detect population structure, no single approach is universally preferred. Furthermore, the desire to leverage large genomic data sets in association studies requires tools capable of handling the computational burden. Here, we present an exploratory effort into developing a novel general modeling framework for inferring population structure from genotype data.

1 Introduction

Non-random mating, mutation, genetic drift, geographic isolation, selective pressures, and other factors contribute to *population structure,* the deviation from an idealized panmictic population. *Population stratification*, systematic differences in allele frequencies, can undermine the effectiveness of statistical analyses which rely on the absence of these factors and on the idealization of a sampled population. In particular, it has been shown that population stratification can confound genome-wide association studies (GWAS), which attempt to identify disease-causing genetic variants among large sets of genomic data. Thus, in order to conduct these studies, as well as to learn about the evolutionary history of the human population, the availability of reliable methods to detect population structure is critical.

Presently, numerous software packages capable of measuring population structure exist. However, concerns over the accuracy, efficiency, and applicability of these tools abound; although a few have emerged as most popular, no single program is universally preferred. There is therefore an ongoing need to develop new approaches to measuring population structure from genomic data.

We present an exploratory effort in developing a novel modeling framework for inferring population structure based on Markov random field theory, first developed in the context of statistical mechanics. We show how relaxations of modeling assumptions made in the popular software package for inferring population structure, STRUCTURE[1], can fundamentally increase computational efficiency while maintaining the potential to capture meaningful results. Our work extends the initial efforts of Lian Garton, who employed the Propp-Wilson algorithm for perfect sampling from STRUCTURE's Markov chain Monte Carlo (MCMC) model for small data sets, in that we seek a Markov random field model with strong neighborhood properties and with well-defined monotonicity over configurations[2].

In the remainder of this introduction, we present some detail about how population stratification can affect genome-wide association studies as well as a survey of existing techniques for inferring population structure. In the next section, we give a brief overview of the mathematics behind Markov random fields and MCMC. In section 3 we provided a detailed description of the STRUCTURE program. After this, we present our own Markov random field model for inferring population structure, followed by a brief discussion and conclusion.

1.1 Population Stratification: Contributing to Bias in Case-Control Studies

As noted in [4], genome-wide association studies (GWAS) conducted since the mid 2000s have contributed greatly to our understanding of the genetic causes of a number of human diseases. The most popular design for one of these studies, largely because of its ability to leverage vast amounts of sequencing data and its wide applicability, is the *case-control* study[5]. The goal of this type of study is to identify alleles which are statistically correlated with the presence of a particular disease. Such genetic markers are then hypothesized to be either causal variants of the disease, or at least in linkage disequilibrium (LD) with the true casual variant.

Conducting a case-control study typically involves obtaining a large number (now $100,000$ to millions) of single nucleotide polymorphisms (SNPs) for a sample of individuals both with (case) and without (control) the disease being studied. The analysis of the data involves determining which SNPs are found disproportionally in case versus control subjects with statistical significance. However, such analysis requires the assumption that all correlation between allele frequencies in the data are due to phenotype differences.

A number of factors can contribute to biased results in these studies because they affect the allele frequencies of case versus control subjects disproportionately. In particular, different degrees of relatedness between individuals with different phenotypes can confound these studies if not accounted for. Furthermore, it can often be difficult to assess the relatedness between individuals with an informal ancestry survey, especially when dealing with admixed populations such as African Americans or Latinos. Therefore, it is necessary to employ computational methods to measure the degree of population structure within a sample of individuals before conducting any disease association study.

1.2 Existing Methods for Inferring Population Structure

Over the last decade, a number of software packages have been developed to aid in the computational detection and quantification of population structure from genotype data. Broadly speaking, these come in two flavors.

First, there are dimensionality-reduction methods, such as principal components analysis (PCA), which reduce the high-dimensional space of hundreds of thousands of measurements down to a handful of orthogonal dimensions in an abstract ancestry space. The program EIGENSOFT is one popular example[6]. Reducing the dimensionality of the genotype data allows fast identification of subpopulations in many cases, and allows visualization of the data. However, it is often difficult to assess the biological significance of the results produced by dimensionality-reduction methods, as their findings do not allow quantification of population stratification on a per-locus basis.

The next category of methods for detecting population structure are model-based approaches, which employ parametric statistical inference. While these approaches tend to be much slower than dimensionality-reduction methods, they

provide distinct advantages. For example, these methods typically allow for the quantification of several phenomena at once, such as population allele frequencies and individual genetic admixture. Additionally, many models allow the natural incorporation of additional sampling information outside of genotype, such as geography or phenotype. Therefore, our focus will be on this second class of methods. In particular, we will examine the popular STRUCTURE program in detail in section 3.

Though STRUCTURE is currently the most widely-used software for population structure due to its age and the extent to which it has been tested, a number of other methods deserve mention. These include software programs based on STRUCTURE itself, such as ADMIXTURE[7] and mSTRUCT[8], which have demonstrated useful speed-ups and modeling improvements over the original, respectively. Additionally, PLINK is gaining popularity due to its efficiency and incorporation of a full suite of association analysis tools[9].

2 Mathematical Background

In this section, we provide an introduction to the mathematics underlying Markov random fields and Markov chain Monte Carlo. We will frame our discussion of STRUCTURE and of our own method in these terms. For a more in depth treatment of these topics, we recommend [10].

2.1 Markov Random Fields and the Gibbs Distribution

2.1.1 random fields

Let V be a finite set of elements $v \in V$, and let Λ be a finite set of labels. Then $X = \{X(v) : v \in V\}$ is *random field* on the set V with probability distribution $\pi(x) \triangleq \mathbb{P}(X = x)$ where the values $x \in \Lambda^V$ are called *configurations*. A configuration x can be expressed as $x = (x(v) : v \in V)$, where $x(v) \in \Lambda$ for all $v \in V$. Furthermore, let $x(V') = (x(v') : v' \in V')$ denote those labels in the configuration of x restricted to a subset $V' \subset V$.

2.1.2 neighborhoods

A *neighborhood system* $N \triangleq \{\mathcal{N}_v : v \in V\}$ on V satisfies, for all $v \in V$,

- $\mathcal{N}_v \subset V$
- $v \notin \mathcal{N}_v$
- $u \in \mathcal{N}_v \Rightarrow v \in \mathcal{N}_u$

We call $\mathcal{N}_v$ the *neighborhood* of v. Intuitively, if we think of V as a set of vertices, then N defines a set of edges E such that $G = (V, E)$ is a simple undirected graph and E contains an edge connecting a pair (u, v) of vertices if and only if u and v are neighbors.

2.1.3 Markov random fields

For a set S with elements $s \in S$, let $S \backslash A$ denote the compliment of A in S. We will use a shorthand notion where $S \backslash \{s\}$ is written as $S \backslash s$ for some element $s \in S$. Then, X is a *Markov random field* (MRF) with respect to a neighborhood system N if, for all $v \in V$,

$$\begin{aligned} \pi^v(x) &\triangleq \mathbb{P}(X(v) = x(v) \mid X(V \backslash v) = x(V \backslash v)) \\ &= \mathbb{P}(X(v) = x(v) \mid X(\mathcal{N}_v) = x(\mathcal{N}_v)) \end{aligned} \tag{1}$$

In other terms, for all $v \in V$, v and $X(V \backslash \{v \cup \mathcal{N}_v\})$ are conditionally independent given $\mathcal{N}_v$. π^v is called the *local characteristic* of X.

2.1.4 Gibbs random fields

Any collection of random variables $X = (X_1, \ldots, X_d)$ with probability distribution $f(x)$, which factors according to

$$f(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \varphi_C(x(C))$$

where $C \in \mathcal{C}(G)$ are the *cliques* of graph G with vertices $V = (V_1, \ldots V_d)$, Z is the *partition function*

$$Z = \sum_{x \in \Lambda^d} \prod_{C \in \mathcal{C}(G)} \varphi_C(x(C))$$

and the φ_Cs define a family of functions over the configurations of $x(C)$, is a *Gibbs random field* with respect to the graph G. The clique functions φ_C define the edge set E of G.

2.1.5 Going between Markov and Gibbs

We note that if X is a Gibbs random field with respect to G, the X is also a Markov random field with respect to G. Furthermore, according to the *Hammersley-Clifford* theorem, if X is a Markov random field with respect to G with $f(x) > 0 \forall_{x \in \Lambda^d}$ then X is a Gibbs random field with respect to G.

2.2 Sampling: Markov Chain Monte Carlo

We now give a brief description of how to leverage the properties of Markov random fields to construct techniques to sample from the complicated probability distributions which they define. Both STRUCTURE and our method for inferring population structure employ this approach to perform sampling. For a more complete introduction to Markov chain Monte Carlo, we recommend [11].

2.2.1 Markov chains

Consider a sequence of random variables $\Theta = (\theta_1, \theta_2, \ldots)$, $\theta_i \in \Omega$ sampled from a conditional probability distribution

$$\mathbb{P}(\theta_t \mid \theta_1, \ldots, \theta_{t-1}) = \mathbb{P}(\theta_t \mid \theta_{t-1})$$

then we say that this sample comes from a first-order *Markov chain* defined by the transition matrix $T_{ij} \triangleq \mathbb{P}(\theta_t = j \mid \theta_{t-1} = i)$. We note the following definitions of properties of a Markov chain.

homogeneity If T remains the same for all t, then we say that Θ form a homogeneous Markov chain.

irreducibility Informally, if starting at any state i we have a probability greater than 0 of ever reaching any other state j, $j \neq i$, then our Markov chain is irreducible.

aperiodicity If $T_{ii} > 0$ for some i, then our Markov chain is aperiodic.

If all of the properties defined above hold for a given Markov chain, then there exists a unique *stationary distribution* π, such that

$$\sum_{i \in \Omega} \pi_i T_{ij} = \pi_j \forall_{j \in \Omega}$$

Furthermore,

$$\mathbb{P}(\theta_t = i) \to \pi_i \forall_{i \in \Omega}$$

as $t \to \infty$. Thus, if we wish to generate approximate samples from a target distribution π, we simply construct a homogenous, irreducible, aperiodic Markov chain with stationary distribution π, initialize the chain at a random initial state θ_1, and run the chain for "a long time", until we are satisfied that $\theta_t, \theta_{t+1}, \ldots$ represent a sample approximately from π. This general approach for approximate sampling is known as *Markov chain Monte Carlo.*

2.2.2 Gibbs sampling

Gibbs sampling is an MCMC technique for generating approximate samples from the stationary distribution of a Markov chain defined by the joint probability of a Markov random field. We will focus on generating samples from the joint distribution of a collection of random variables $X = (X_1, \ldots, X_d)$ which are a Markov random field with respect to a graph G. This technique involves the following steps

1. Choose an initial state $\theta_1 = x_1$
2. sample $\theta_t \sim \mathbb{P}(\cdot \mid \theta_{t-1})$

(a) choose a vertex $v \in V$ in the vertex set of G uniformly at random.

(b) sample a new value for $x(v)$ based on its local specification (see equation1).

3. repeat 2 many times, and then treat $\theta_{t+c}, \theta_{t+c+1}, \ldots$ as approximate samples from $f(x)$.

3 STRUCTURE

In this section, we present the STRUCTURE algorithm as described in the 2000 paper by Pritchard, et. al. STRUCTURE uses a Bayesian approach to cluster individuals into populations based on genotype data. Their model and algorithm jointly perform inference on the allele frequency profiles of each population and the population or populations to which each individual in the data set belongs.

3.1 Model

We now give an overview of the models employed in STRUCTURE for absolute classification and classification allowing for genetic admixture (partial membership to subpopulations).

3.1.1 modeling assumptions

Let $\boldsymbol{X}$ denote the observed genotypes of a data set of individuals, and let the unobserved random variables $\boldsymbol{Z}$ and $\boldsymbol{P}$ be the populations of origin of each individual and the allele frequency profile of each population, respectively. STRUCTURE assumes Hardy-Weinberg equilibrium and complete linkage equilibrium between loci within each population. Under these assumptions, each individual's genotype is made up of independent samples given the appropriate allele frequency distributions for each locus. Fixing $\boldsymbol{Z}$ and $\boldsymbol{P}$ completely specifies $\mathbb{P}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{P})$.

STRUCTURE uses a Bayesian framework to perform inference on $\boldsymbol{Z}$ and $\boldsymbol{P}$ in the following way. First, note that

$$\mathbb{P}(\boldsymbol{Z}, \boldsymbol{P} \mid \boldsymbol{X}) \propto \mathbb{P}(\boldsymbol{Z})\,\mathbb{P}(\boldsymbol{P})\,\mathbb{P}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{P})$$

As we describe in the next section, STRUCTURE uses a Markov chain Monte Carlo approach to obtain joint samples $\left(\boldsymbol{Z}^i, \boldsymbol{P}^i\right)$ from the this posterior distribution.

3.1.2 without admixture

In the STRUCTURE model without admixture, the goal is to assign each individual to a unique population of origin. To make the notation introduced above a bit more concrete, let $x_l^{(i,a)} \in \{1, \ldots, J_l\}$ be the observed allele copy a from the genotype of individual i at locus l, where J_l denotes the number

of possible alleles at locus l. Furthermore, let $z^i \in \{1, \ldots, K\}$ be the population of origin of individual i, and let p_{klj} be the frequency of allele j at locus l for population k. As mentioned in the previous subsection, the assumptions of Hardy-Weinberg equilibrium and linkage disequilibrium allow the probabilities $\mathbb{P}\left(x_l^{(i,a)} = j \mid Z, P\right)$ to be independent for each l and a. Furthermore, according to the assumption that individual alleles are random samples from the appropriate allele frequency profile

$$\mathbb{P}\left(x_l^{(i,a)} = j \mid Z, P\right) = p_{z^i lj} \tag{2}$$

A priori, STRUCTURE assumes that each individual is equally likely to have originated from each of the K populations. In other words, for all i

$$\mathbb{P}\left(z^i = k\right) = \frac{1}{k} \tag{3}$$

3.1.3 with admixture

The STRUCTURE model allowing for admixture includes an additional set of parameters, Q, which represent the degree of admixture in each individual. q_k^i is the proportion of individual i's genome which originated from population k. Z now has an extra dimension for each allele copy of each individual. $z_l^{(i,a)}$ is the population of origin of the allele copy $x_l^{(i,a)}$. Note that there is a one-to-one correspondence now between the elements of Z and X, and also that the distribution of Z now depends on Q. Equations 2 and 3 become, respectively,

$$\mathbb{P}\left(x_l^{(i,a)} = j \mid Z, P, Q\right) = p_{z_l^{(i,a)} lj}$$

$$\mathbb{P}\left(z_l^{(i,a)} = k \mid X, Q\right) = q_k^i$$

3.2 Sampling Methods

STRUCTURE uses an MCMC method to perform inference on P and Q by generating samples from the posterior probability distribution

$$\mathbb{P}\left(\boldsymbol{Z}, \boldsymbol{P}, \boldsymbol{Q} \mid \boldsymbol{X}\right) \propto \mathbb{P}\left(\boldsymbol{Z}\right) \mathbb{P}\left(\boldsymbol{P}\right) \mathbb{P}\left(\boldsymbol{Q}\right) \mathbb{P}\left(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{P}, \boldsymbol{Q}\right)$$

Analogous to our description of Gibbs sampling in the previous section, STRUCTURE performs the following steps in sampling

1. Update $\boldsymbol{p}_{kl}$ for all k, l according to

$$\boldsymbol{p}_{kl} \mid X, Z \sim \mathcal{D}\left(\lambda_1 + n_{kl1}, \ldots, \lambda_J + n_{klJ}\right)$$

where

$$n_{klj} = \#\left\{(i,a) : x_l^{(i,a)} = j, z_l^{(i,a)} = k\right\}$$

2. Update $\boldsymbol{q}^{(i)}$ for all i according to

$$\boldsymbol{q}^{(i)} \mid X, Z \sim \mathcal{D}\left(\alpha + m_1^{(i)}, \ldots, \alpha + m_K^{(i)}\right)$$

where

$$m_k^{(i)} = \#\left\{(l,a) : z_l^{(i,a)} = k\right\}$$

3. For each $z_l^{(i,a)} \in Z$, sample a new population of origin based on

$$\mathbb{P}\left(z_l^{(i,a)} = k \mid X, P, Q\right) = \beta q_k^{(i)} p_{klx_l^{(i,a)}}$$

where β is a normalizing factor

$$\beta = \sum_{k \in K} q_k^{(i)} p_{klx_l^{(i,a)}}$$

4 A Markov Random Field Model for Population Structure

We now present our Markov random field model for inferring population structure. We note that our model allows for the sampling of each $z_l^{(i,a)} \in Z$ independently, conditioned only to the neighborhood system $\mathcal{N}$, defined below

$$\mathcal{N}_l^{(i,a)} = \left\{x_l^{(i',1)}, x_l^{(i',2)}, z_l^{(i',1)}, z_l^{(i',2)} : i' \neq i\right\} \cup \left\{x_{l'}^{(i,1)}, x_{l'}^{(i,2)}, z_{l'}^{(i,1)}, z_{l'}^{(i,2)} : l' \neq l\right\} \tag{4}$$

Our model also gets rid of the latent variables P, Q in structure, replacing them with the following quantities

$$\tilde{p}_{klj} \mid \mathcal{N}_l^{(i,a)} = \mathbb{E}\left[\mathcal{D}\left(\lambda_1 + \tilde{n}_{kl1}, \ldots, \lambda_J + \tilde{n}_{klJ}\right)_j\right] = \frac{\lambda_j + \tilde{n}_{klj}}{\sum_{j'=1}^{J} \lambda_{j'} + \tilde{n}_{klj'}}$$

where

$$\tilde{n}_{klj} = \#\left\{(i,a) : x_l^{(i,a)} = j, z_l^{(i,a)} = k, x_l^{(i,a)}, z_l^{(i,a)} \in \mathcal{N}_l^{(i,a)}\right\}$$

and similarly,

$$\tilde{q}_k^{(i)} \mid \mathcal{N}_l^{(i,a)} = \mathbb{E}\left[\mathcal{D}\left(\alpha + \tilde{m}_1^{(i)}, \ldots, \alpha + \tilde{m}_K^{(i)}\right)_k\right] = \frac{\alpha + \tilde{m}_k^{(i)}}{K\alpha + \sum_{k'=1}^{K} \tilde{m}_{k'}^{(i)}}$$

where

$$\tilde{m}_k^{(i)} = \#\left\{(l,a) : z_l^{(i,a)} = k, z_l^{(i,a)} \in \mathcal{N}_l^{(i,a)}\right\}$$

Note that $\tilde{p}_{klj}$ and $\tilde{q}_k^{(i)}$ are approximations of their counterpart quantities in the STRUCTURE program. This model allows a Gibbs sampling technique for inference in which an allele copy can be chosen uniformly at random, and updated according to a probability distribution defined only by the populations of origins of the allele copies in its neighborhood, defined by equation 4.

5 Results

We evaluated our model qualitatively on two data sets for which STRUCTURE has also been tested. The first data set is comprised of a series of dominant markers, detected using amplified fragment length polymorphisms (AFLPs), from a number of whitefish. Half of these fish were known to have a dwarf phenotype, and the other half were normal. The results of both my model and of STRUCTURE are shown in Figure 1. Observe that both our model and STRUCTURE detected admixture in roughly the same individuals. However, the membership of each individual seems to be more skewed in our model as compared with STRUCTURE.

The second data set on which we verified our model is a sample of 2810 single nucleotide polymorphisms (SNPs) from individuals from 53 human populations across 6 continents[15]. Outputs from our model and STRUCTURE are shown in Figure 2.

6 Conclusion

In summary, we have described a modification of the STRUCTURE program for inferring population structure based on Markov random field approach. Our model is theoretically simpler, more flexible, and faster for sampling and other computations. We have shown that our model uses a simplifying approximation while still

7 Future Work

We hope to conduct further model verification in the future by gathering more quantitative results. Additionally, we would like to implement more of the modeling layers of the STRUCTURE program in terms of our Markov random field framework.

References

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

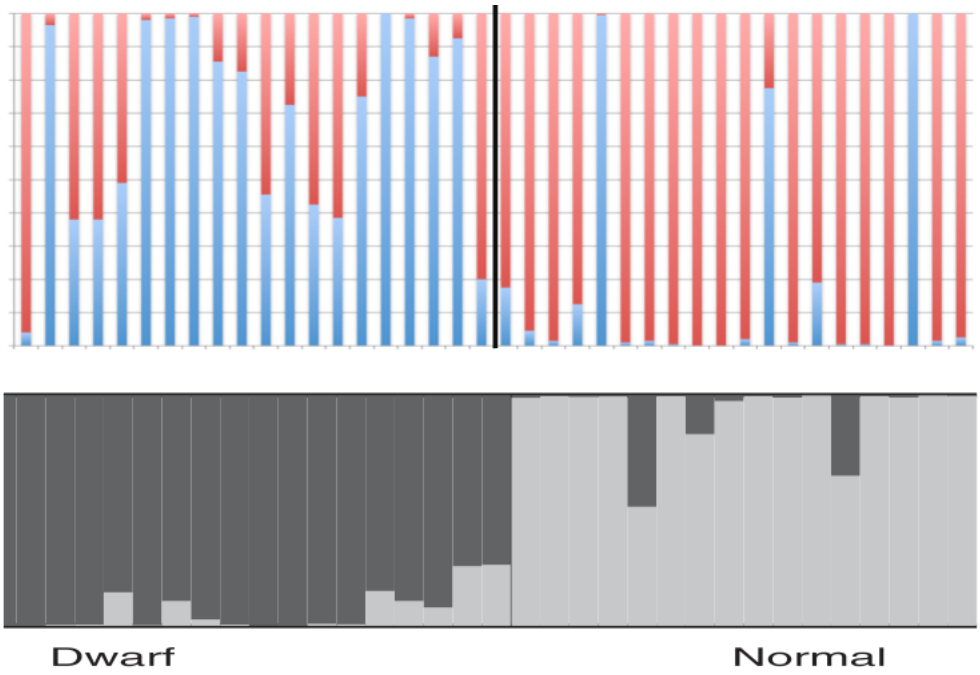

Figure 1: Stacked bar plots indicating inferred admixture of each individual ($K = 2$). (top) output of my model (bottom) STRUCTUREs output. Individuals are aligned between the top and bottom plots.

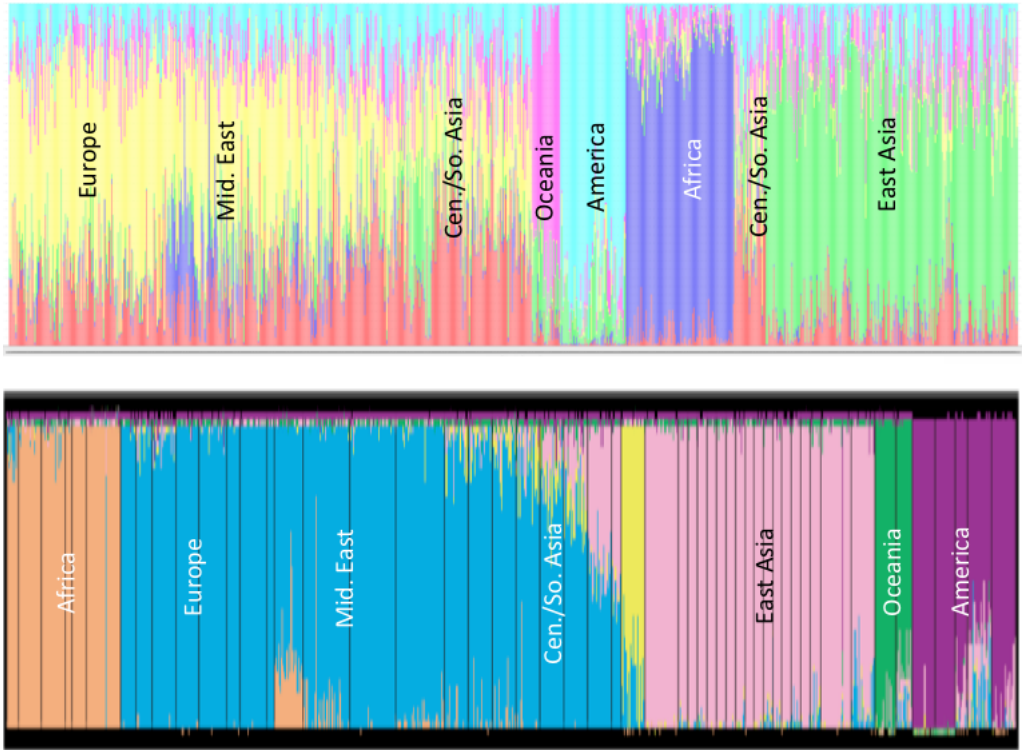

Figure 2: Stacked bar plots indicating inferred admixture of each individual ($K = 6$). (top) output of my model (bottom) STRUCTUREs output. Geographic region of origin for each group of individuals is included in both plots.

[2] Lian Garton. An Investigation of Population Subdivision Methods in Disease Associations with a Focus on Markov Chain Monte Carlo. Undergraduate Honors Thesis in Computer Science, Department of Computer Science, Brown University. May 2, 2008.

[3] Chao Tian, Peter K. Gregersen, and Michael F. Seldin. Accounting for ancestry: population substructure and genome-wide association studies. Human Molecular Genetics, 2008, Vol. 17, Review Issue 2 R143–R150 doi:10.1093/hmg/ddn268

[4] Peter M. Visscher, Matthew A. Brown, Mark I. McCarthy, and Jian Yang. Five years of GWAS discovery. The American Journal of Human Genetics 90, 7–24, January 13, 2012 doi 10.1016/j.ajhg.2011.11.029

[5] Hemant K. Tiwari, Jill Barnholtz-Sloan, Nathan Wineinger, Miguel A. Padilla, Laura K. Vaughana, David B. Allison. Review and Evaluation of Methods Correcting for Population Stratification with a Focus on Underlying Statistical Principles. Human Heredity 2008;66:67–86 doi: 10.1159/000119107

[6] N. Patterson, A. Price, D. Reich (2006) Population structure and eigenanalysis. PLoS Genet 2:e190.

[7] David H. Alexander, John Novembre and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 19: 1655-1664 doi:10.1101/gr.094052.109

[8] S. Shringarpure and E. Xing. Mstruct: Inference of Population Structure in Light of both Genetic Admixing and Allele Mutations. Genetics, vol. 108, pp. 575-593, 2009.

[9] S. Purcell, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559–575 (2007).

[10] Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. New York: Springer-Verlag, 1999.

[11] Häggström, Olle. Finite Markov Chains and Algorithmic Applications. Cambridge: Cambridge University Press, 2002.

[12] Sunil K. Kopparapu, Uday B. Desai. Bayesian Approach to Image Interpretation. Nowell, Massachusettes: Kluwer Academic, 2001.

[13] Daniel Falush, Matthew Stephens, and Jonathan K. Pritchard. Inference of Population Structure Using Multilocous Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics 164: 1567-1587, August 2003

[14] Daniel Falush, Matthew Stephens, and Jonathan K. Pritchard. Inference of Population Structure Using Multilocus Genotype Data: Dominant Markers and Null Alleles. Blackwell Publishing Ltd, 2007.

[15] Noah A. Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman. Genetic Structure of Human Populations. Science 298 (5602), 2381-2385: 20 December 2002. doi:10.1126/science.1078311