StructHDP: automatic inference of number of clusters and population structure from admixed genotype data

Suyash Shringarpure*, Daegun Won and Eric P. Xing*
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

Motivation: Clustering of genotype data is an important way of understanding similarities and differences between populations. A summary of populations through clustering allows us to make inferences about the history of the populations. Many methods have been proposed to perform clustering on multilocus genotype data. However, most of these methods do not directly address the question of how many clusters the data should be divided into and leave that choice to the user.

Methods: We present StructHDP, which is a method for automatically inferring the number of clusters from genotype data in the presence of admixture. Our method is an extension of two existing methods, Structure and Structurama. Using a Hierarchical Dirichlet Process (HDP), we model the presence of admixture of an unknown number of ancestral populations in a given sample of genotype data. We use a Gibbs sampler to perform inference on the resulting model and infer the ancestry proportions and the number of clusters that best explain the data.

Results: To demonstrate our method, we simulated data from an island model using the neutral coalescent. Comparing the results of StructHDP with Structurama shows the utility of combining HDPs with the Structure model. We used StructHDP to analyze a dataset of 155 Taita thrush, Turdus helleri, which has been previously analyzed using Structure and Structurama. StructHDP correctly picks the optimal number of populations to cluster the data. The clustering based on the inferred ancestry proportions also agrees with that inferred using Structure for the optimal number of populations. We also analyzed data from 1048 individuals from the Human Genome Diversity project from 53 world populations. We found that the clusters obtained correspond with major geographical divisions of the world, which is in agreement with previous analyses of the dataset.

Availability: StructHDP is written in C++. The code will be available for download at http://www.sailing.cs.cmu.edu/structhdp.

Contact: suyash@cs.cmu.edu; epxing@cs.cmu.edu

1 INTRODUCTION

The clustering of individuals into populations is a frequently undertaken task in studies of genetic data. Clustering allows the summarization of individuals into groups based on genetic similarity. Such a summary is easy to visualize and understand. It enables us to make simple inferences about the similarities and differences between different groups of individuals. We can also make inferences about evolutionary history of populations. An understanding of the population stratification present in a set of individuals is also required for avoiding false positives in genetic association studies.

Phenomena like migration and admixture make the clustering problem harder, since individuals need not belong to exactly one population cluster, but can have partial membership in multiple clusters. Different parts of an individual’s genome can be inherited from individuals belonging to different populations. The problem of clustering individuals while allowing partial membership in multiple clusters was addressed using a Bayesian model by Pritchard et al. (2000). Extensions to the underlying model that account for other evolutionary processes such as mutation (Shringarpure and Xing, 2009) and recombination (Falush et al., 2003) have also been proposed.

An important question that needs to be addressed when solving the problem of population stratification is deciding how many populations are required to best explain the variation observed in a given set of individuals. The Bayesian models mentioned earlier require the user to specify a number of clusters (also called ancestral populations) into which the individuals are divided. However, this might not always be possible or desirable. To address this problem, an extension of Structure was developed by Pella and Masuda (2006). Based on their method, Huelsenbeck and Andolfatto (2007) developed Structurama. Structurama automatically infers the number of population clusters into which a given dataset should be divided provided individuals only belong to a single population. A commonly used solution to this problem is to use fixed-dimensionality models in combination with an information criterion (Akaike, 1974; Schwarz, 1978) to decide the optimal number of populations.

In this article, we present StructHDP, a method for automatically inferring the number of population clusters present in a group of individuals, while accounting for admixture between ancestral populations. Using the Hierarchical Dirichlet Process framework for clustering developed by Teh et al. (2006), we extend the Structure model so that the number of populations is inferred by the model and need not be specified by the user. This work is also an extension of the Dirichlet process model developed by Pella and Masuda (2006), which has been implemented in Structurama.

We simulated data from an island model using the neutral coalescent to test the performance of our method at recovering the true number of ancestral populations. Comparing the results of StructHDP with Structurama shows the utility of combining HDPs with the Structure model. We used StructHDP to analyze a set of 155 Taita thrush individuals, Turdus helleri. This dataset has been previously analyzed using Structure and Structurama. We found that StructHDP correctly identifies the optimal number of populations to cluster the data. The clustering enforced by the inferred ancestry proportions for individuals also agrees with that inferred using Structure with the appropriate choice of the number of clusters.
of populations $K$. We also analyzed a set of 1048 individuals from the Human Genome Diversity Project (HGDP) using StrucHDP. We found that the clusters inferred coincide with the major geographical divisions present in the data. We also observed that the distance between populations (based on their cluster memberships) is strongly positively correlated with $F_{ST}$ between populations, which suggests that the inferred cluster memberships capture the genetic variation present in the data well.

2 RELATED WORK

Model-based clustering methods have become popular since the Structure model by Pritchard et al. (2000). The method uses a Bayesian model to capture admixture in populations, with model parameters inferred by Markov Chain Monte Carlo (MCMC) sampling. Extensions to the underlying model that account for other evolutionary processes such as mutation (Shringarpure and Xing, 2009) and recombination (Falush et al., 2003) have also been proposed. These methods have the advantage that they use an explicit demographic model of population stratification. However, they are not efficient for large datasets.

Population clustering methods have also been developed in a complementary direction, using low-dimensional projections and eigenanalysis to cluster individuals (Patterson et al., 2006). These methods assume no underlying evolutionary model but have been shown to be good at modeling population structure. They are efficient and also allow the user to compute the statistical significance of the results obtained.

Recently, a set of methods such as Admixture (Alexander et al., 2009) and Frappe (Tang et al., 2005) have been developed that use the likelihood model of Structure but use faster optimization methods for estimating model parameters. Frappe uses an expectation–maximization algorithm to determine individual ancestries. Admixture uses a block relaxation algorithm to speed up the optimization. Both these models are faster than Structure and allow the analysis of large datasets with hundreds of thousands of markers.

In all the models mentioned above, an important choice that must be made by the user is the number of populations ($K$) or clusters that the dataset must be divided into. To help the user make this choice, some methods provide an approximation to the posterior probability of the number of populations given in the data. Alternatively, the user can use information criterion such as the Akaike Information Criterion (AIC) (Akaike, 1974) or the Bayesian Information Criterion (BIC) (Schwarz, 1978) to choose the number of populations that best explain the data. The choice of number of populations is a trade-off between model expressiveness (how well the model captures the variation in the data) and model complexity (how many parameters the model needs to capture the variation).

Huelsenbeck and Andolfatto (2007) developed an extension to Structure, called Structurama, that automatically infers the number of populations that are present in the data. Their method assumes that an individual only belongs to a single cluster and that the number of populations is also a random variable. They use a Dirichlet Process model that can have membership in multiple clusters and the number of populations will automatically be inferred by the model. In the following sections, we describe the model and discuss the results of analyzing some simulated and real data using StrucHDP.

3 APPROACH

Bayesian models for clustering of genotype data use the framework of mixture models to model individuals. In a mixture model, an individual is assumed to be made up of a number of genetic markers. Each of those markers is assumed to originate from one of a finite number of ancestral populations. Ancestral populations are usually defined as a collection of allele frequencies at the markers under consideration. An individual can thus be considered to be a mixture of one or more ancestral populations, which explicitly accounts for admixture between populations. The proportions of different populations within an individual’s genome are usually individual specific and provide a compact summary of the individual’s genome. This is the basis of the Structure model by Pritchard et al. (2000).

We approach the problem of finding the optimal number of ancestral populations by extending the above model to a setting where there are potentially infinite ancestral population components in the mixture. Performing inference then allows us to examine the number of ancestral populations that have a non-zero contribution to the set of individuals under consideration. We use the Hierarchical Dirichlet Process framework (Teh et al., 2006) to model the mixture of infinite ancestral populations.

Consider the problem of clustering the markers within a single individual based on their population of origin. We can assume that the number of populations that contribute to the single individual’s genome is unknown and is a random variable. The Dirichlet process (DP) (Ferguson, 1973) was proposed to solve a problem of this nature, where objects (genetic markers) belong to one of a potentially infinite number of mixture components (ancestral populations). In the case of multiple individuals, we can posit multiple DPs, one for each individual, that will address the problem of not knowing the optimal number of populations. We also require that the ancestral populations inferred for the DPs be the same across all the individuals. Mathematically, this is analogous to ensuring that mixture components are shared across DPs.

The Hierarchical Dirichlet process (HDP) is a framework for clustering of observations when the observations are present in groups. Each group can be modeled using a finite mixture model or a DP. The mixture models or DPs across groups are linked by sharing mixture components. It is useful to think of each group as having its own DP, with the processes linked to each other by the parameters of the HDP. StrucHDP is based on the HDP described by Teh et al. (2006). In the following section, we provide a description and mathematical representation of the HDP model.

4 METHODS

A commonly used analogy for representing HDPs is the Chinese Restaurant Process (CRP). This is an extension of the representation of the DP as a Chinese restaurant with customers. The DP representation and its application to Structurama are described in more detail by Huelsenbeck and Andolfatto (2007). A CRP comprises of a number of Chinese restaurants which share
The dishes served in the restaurant are the mixture components, and sharing HDP so that the generative process looks as follows:

\[\gamma, \theta, \alpha, \beta, \pi, \phi, z_{ji} \]

i.e. the observation is assigned the particular mixture component that is associated with the table. All observations that are assigned to a particular table are considered to originate from the same mixture component, clustering the observations within the group. The same mixture component might also be shared across multiple tables within a group. The method of choosing a table for a new customer is similar to a 'rich gets richer' model which is regulated by the probability of starting a new table. This is the property of the HDP that is responsible for its clustering behavior.

This analysis can be easily extended to the case of genetic data, with every individual considered to be a separate group corresponding to a restaurant. The loci within an individual are the customers in the restaurant, and the ancestral populations are the mixture components or the dishes in the CRF. A minor subtlety that arises in this case is that the set of possible alleles at each locus might be different, which needs to be accounted for in the inference process. This can be accomplished easily with some minor bookkeeping without changing the inference process significantly.

Consider a dataset having \(N\) individuals genotyped at \(M\) loci. The observed allele for individual \(j\) at locus \(i\) is denoted by \(x_{ji}\). For ease of representation, we will ignore the diploid nature of genotype data. In implementation, we shall allow our method to handle data of any fixed ploidy. The HDP can then be used to generate the allele \(x_{ji}\) for the \(j\)-th individual at the \(i\)-th locus as follows:

\[ G_j \sim \text{DP}(\gamma, H_1) \]
\[ G_{ij} \sim \text{DP}(\alpha_{ij}, G_j) \]
\[ \theta_j \sim G_j \]
\[ x_{ji} \sim \theta_j \]

Here, \(H_1\) is the base distribution over alleles at locus \(i\), commonly a Dirichlet distribution. \(\gamma\) and \(\alpha_{ij}\) are parameters of the HDP that control how fast new populations are added to the model. \(G_j\) is an intermediate probability distribution over alleles at locus \(i\) and \(G_j\) is a distribution specific to individual \(j\). The individual-specific distributions \(G_{ij}\) are connected to one another through \(G_j\) and \(\alpha_{ij}\), ensuring the sharing of ancestral populations across individuals. \(G_j\) and \(G_{ij}\) are both generated by DP that use \(\gamma\) and \(\alpha_{ij}\) as parameters. \(\theta_j\) denotes the mixture components. \(x_{ji}\) is a sample from a distribution \(F(\theta_j)\), a multinomial distribution over alleles in our case.

For modeling purposes, it is helpful to modify the representation of the HDP so that the generative process looks as follows:

\[ \beta | \gamma \sim \text{GEM}(\gamma) \]  
\[ \pi | \alpha, \beta \sim \text{DP}(\alpha, \beta) \]  
\[ \phi_0 | H_i \sim H_i \]  
\[ \phi_0 | \pi \sim \pi \]  
\[ x_{ji} | \phi_{kji} \sim \text{F}(\phi_{kji}) \]

where we say that \(\beta = (\beta_{kji})_{k=1}^{\infty}\) if it satisfies the following construction:

\[ \beta_{kji} | \gamma \sim \text{Beta}(1, \gamma) \]
\[ \beta_k = \sum_{j=1}^{M} \beta_{kji} \]

This construction ensures that \(\sum_{k} \beta_k = 1\). The \(\beta\) thus represents the fractional contributions of the potentially infinite populations to the given set of individuals.

In the HDP representation above, \(\phi_{kji}\) represents the allele frequencies of the \(k\)-th population at the \(i\)-th locus. \(\pi\) is a vector that denotes the ancestry proportions (contributions from all populations) for individual \(j\), and its components sum to 1. The indicator variable \(z_{ji}\) denotes which population the observed allele \(x_{ji}\) at locus \(i\) originates from. We will use this notation for representing the HDP model for our problem due to its similarity with the Structure generative process. This representation also shows how the model can account for diploid individuals by changing the step of sampling \(\gamma\) and \(\beta\) to the following:

\[ z_{ji,1} | \pi \sim \pi \]
\[ z_{ji,2} | \pi \sim \pi \]
\[ x_{ji,1} | (\phi_{kji,1}, \phi_{kji,2}) \sim \text{F}(\phi_{kji,1}) \]
\[ x_{ji,2} | (\phi_{kji,1}, \phi_{kji,2}) \sim \text{F}(\phi_{kji,2}) \]

where \(z_{ji,1}\) and \(z_{ji,2}\) now represent the two alleles at locus \(i\) in individual \(j\) and \(z_{ji,1}\) and \(z_{ji,2}\) are their respective population indicator variables. This allows the model to account for mixed ancestries at a single locus as well.

For ease of representation, we will drop the subscript indicating the ploidy in the analysis.

Figure 1 shows the graphical model representation of the StructHDP generative process. In this graphical model representation, the nodes represent random variables which have been described earlier. The edges denote dependencies between the random variables due to the sampling steps in the generative process. The shaded nodes represent the random variables we observe, namely, the alleles observed at each locus.

To allow for more flexibility with the parameter settings, we impose priors on \(\alpha_{ij}\) and \(\gamma\) and the base distributions \(H_i\). We assume that \(\alpha_{ij}\) and \(\gamma\) have Gamma priors with parameters \((\alpha_{ij, aij})\) and \((\gamma, \gamma_0)\), respectively, and that \(H_i\) has a symmetric Dirichlet distribution with parameter \(\pi\). The graphical model with all priors shown can be seen in Figure A11 in the Appendix A.

\[ a_{ij} \sim \text{Gamma}(\alpha_{ij, aij}) \]  
\[ \gamma \sim \text{Gamma}(\gamma, \gamma_0) \]  
\[ H_i \sim \text{Dir}(\pi) \]  

4.1 Inference

For performing inference on the model, we use Gibbs sampling, a MCMC sampling method described for the HDP by Teh et al. (2006). For inference in the CRF representation of the HDP, we create some bookkeeping variables \(m\) that keep count of the number of tables at the restaurant and franchise levels. More details about the distributions used for sampling the variables are presented in the Appendix A.
4.2 Inference steps

Using all the variable updates, the inference process can be described as:

(1) Set the values for the prior parameters \(a_0, a_0, \beta_0, \gamma_0, \lambda\).
(2) Start with random values for all other variables.
(3) Sample \(x\) variables given all other variables.
(4) Sample \(m\) variables given all other variables, using updated value of \(x\).
(5) Sample \(\beta\) given all other variables, using updated values of \(x\) and \(m\).
(6) Sample \(a_0\) using updated values of \(x, m\) and \(\beta\).
(7) Sample \(y\) using updated values of all other variables.
(8) Repeat steps 3–7 until convergence.

The Gibbs sampling update distributions can be derived following the methodology in Teh et al. (2006). Due to space limitations, the details of the Gibbs sampling update distributions and their derivations are deferred to the Appendix A.

4.3 Other inference details

Like all MCMC methods, the sampler is run for a large number of iterations, with some initial iterations discarded as burn-in. Samples from the posterior can then be used to estimate the ancestry proportions \(\pi_j\) for each individual. The posterior distribution for the individual ancestry proportions \(\pi_j\) can be shown to be a Dirichlet distribution:

\[
\pi_j \sim \text{Dir}
\left(\frac{a_{0j}}{\sum_{i=1}^{N_{m}} I[z_{ij} = k]}, \ldots\right)
\]

where \(I\) denotes an indicator function. If the number of populations remains constant across iterations in the sampling (as is often seen to happen in our experiments after a large number of iterations), this estimate can be averaged over multiple samples to get a more accurate estimate of the individual ancestry proportions.

As with the Gibbs sampler used in STRUCTURE, our method could have problems with the identifiability of clusters, if label switching for the clusters was a frequent occurrence. In practice, we find that label switching is infrequent, and is avoided by the use of the restricted growth function (RGF) notation of Stanton and White (1986) in summarizing MCMC results.

5 RESULTS

5.1 Coalescent simulation data

We performed coalescent simulations based on an island model similar to Hulsenbeck and Andolfatto (2007). We used the program ms (Hudson, 2002) to simulate samples under a neutral coalescent model. As an initial evaluation of the performance of StructHDP in recovering the correct number of population clusters, we simulated data from a symmetric equilibrium island model with 4 demes of equal size, with the mutation rate \(\theta = 4N_e \mu = 0.5\) and migration rate \(M = 4N_e m = \{1, 2, 4\}\). In each case, 100 diploid individuals were sampled with an equal number being sampled from each deme. Fifty replicates were created for each parameter setting.

We analyzed the data using StructHDP, STRUCTURAMA and ADMIXTURE. For StructHDP, the priors for both concentration parameters were set to \((0.5,0.5)\) and the parameter for the Dirichlet distribution of \(H\) was set to 0.5. The StructHDP Gibbs sampler was run for 25,000 iterations, with the first 12,500 iterations discarded as burn-in. To thin the Markov chain, samples were taken every 25 iterations. We computed the expected value of the number of populations, \(K\), using the sampled values of \(K\) from the Gibbs sampler. The expected value of \(K\), \(E(K|X)\) can then be compared against the true value of the number of demes, \(K_T = 4\), across multiple replicates, to get an error measure that is given by \(E(K|X) - K_T\) (Hulsenbeck and Andolfatto, 2007). For STRUCTURAMA, the experiments for each parameter setting were performed with different priors on the expected number of populations in Hulsenbeck and Andolfatto (2007). For comparison purposes, we chose the best result, i.e. the prior setting that gave the least error. Model selection with ADMIXTURE can be done in three different ways by choosing either the AIC, BIC or the cross-validation error as the measure of model fit. We present results for all three measures.

Table 1 shows the results of the simulation. We can see that the error in recovering \(K\) is much smaller for StructHDP than for STRUCTURAMA and for ADMIXTURE, except when the migration rate is small. The underlying assumption of the DP model of STRUCTURAMA is that there is no admixture and individuals only belong to a single ancestral population. As a result, in a simulation setting with less admixture due to migration, the number of recovered populations for STRUCTURAMA is almost perfect. As the amount of admixture increases, the error in the number of recovered populations increases. On the other hand, StructHDP explicitly accounts for admixture in the model. Therefore, it recovers the true number of demes in the island model with low error for all parameter values. In terms of \(F_{ST}\), we can say that as the \(F_{ST}\) between the demes decreases (as migration increases), the accuracy of STRUCTURAMA drops while that of StructHDP is nearly unaffected.

ADMIXTURE performs worse than both StructHDP and STRUCTURAMA in recovering the true number of populations. This may be due to the small number of markers that are used in the simulation study.

5.2 Real data analysis

5.2.1 Taita thrush data

We used our method to analyze a dataset of \(N = 155\) Taita thrushes, Turdus helleri (Galbusera et al., 2000). Each individual was genotyped at \(M = 7\) microsatellite loci. Individuals were sampled at four locations in southeast Kenya [Chawia (17 individuals), Ngangao (54), Mboholo (80) and Yale (4)]. The thrush data were previously analyzed in Hulsenbeck and Andolfatto (2007); Pritchard et al. (2000) so we use it to verify the correctness of StructHDP.

We ran StructHDP for 25,000 iterations, with the first 12,500 iterations as burn-in. Samples were taken every 25 iterations to thin the Markov chain. The priors for both concentration parameters were
S.Shringarpure et al.

Fig. 2. Posterior distribution for number of populations, \( \Pr(K|X) \) for the thrush data.

Fig. 3. A single sample of the ancestry proportions for the thrush data. The black lines separate the individuals according to their geographic labels. The analysis did not use any geographical information.

Fig. 4. The ancestry proportions for the thrush data from a single \textit{Structure} run for \( K = 3 \) set to (0.5,0.5) and the parameter for the Dirichlet distribution of \( H \) was set to 0.5.

We find that our method converges to \( K = 3 \) populations in a few thousand iterations. The posterior distribution for \( K \) is shown in Figure 2. From the posterior, we can see that \( K = 3 \) is the most likely value for \( K \). Figure 3 shows a single sample for the ancestry proportions of the thrush data. The clusters agree with geographical labels well except for a few individuals. We also see that the four Yale individuals fall into the same cluster as the Ngangao individuals. All these findings agree with those of Pritchard et al. (2000) when \textit{Structure} is initialized with \( K = 3 \) clusters. Figure 4 shows the results of \textit{Structure} analysis of the thrush data with \( K = 3 \). In their analysis, Pritchard et al. also found that \( K = 3 \) explains the data best. Their conclusion was based on an \textit{ad hoc} approximation to \( \Pr(K|X) \), the posterior likelihood of \( K \) given the data \( X \) while StructHDP automatically infers this from the data.

5.2.2 Human genome diversity project The Human Genome Diversity Project dataset we analyze consists of 1048 individuals from 53 world populations genotyped at 783 microsatellite loci. Along with genotype information, the individuals are also labeled with the geographical divisions to which they belong. Using \textit{Structure}, Rosenberg et al. (2002) have previously analyzed the genotype data and found that the population clusters correspond to major geographical divisions of the world. We used StructHDP to reanalyze this data (without making use of the geographical information). The sampler was run for 20,000 iterations with the first 10,000 iterations discarded as burn-in. Samples were taken every 25 iterations to thin the Markov chain.

To determine the optimal number of ancestral populations, we examined the posterior distribution of the number of populations (\( K \)). Figure 5 shows the posterior distribution. We find the posterior distribution has a single mode at \( K = 4 \) and non-zero probability mass for values of \( K \) up to 8. For further analyses, we use the maximum-likelihood sample from the MCMC sampling, which has four ancestral population components.

The contributions of the four ancestral populations to an individual’s genome can be represented using a 4D vector whose components sum to 1. All these vectors (referred to as ancestry proportions) lie within a tetrahedron in 3D space. Each of the four vertices of the tetrahedron represents an ancestral population. To visualize the clustering, we plotted the ancestry proportions for the 1048 individuals in three dimensions along with the tetrahedron in which the vectors lie. In this representation, the distance of a vector from the vertices of the tetrahedron indicates the amount of admixture present in an individual’s genome. The further away from a vertex the vector is (and the closer it is to the center of the tetrahedron), the more the admixture present in the individual’s genome.

Figure 6 shows the resulting plot for the 1048 individuals in the HGDP dataset. In the plot, each individual is represented by a small sphere. For ease of interpretation, the individual spheres are colored based on the geographical division they belong to. In the populations we examine, the divisions are Africa, Americas, Central and South Asia, East Asia, Europe, Middle East and Oceania. These are represented by seven different colors. From the figure, we can see that individuals from a single continent cluster together in the same region of the tetrahedron. Some individual genomes are derived from a single ancestral population and lie at the vertices of the tetrahedron. Some other individuals, particularly those belonging to the Middle Eastern, Central Asian and South Asian populations, show a lot of admixture.
To analyze these results further, we plotted the ancestry proportions of the 1048 individuals as a bar graph, where every individual is represented by a thin bar with four components which sum to 1. Figure 7 shows the resulting bar graph. We can see that the clusters obtained correspond to the major geographical divisions of the world and the ancestral populations can be roughly described as ancestral African (denoted by green color), ancestral American-East Asian (blue), ancestral European (yellow) and ancestral Oceanian (red). From the ancestry proportions, we can see that the modern East Asian populations and American populations are similar, with the modern East Asian populations having a larger contribution from the ancestral population corresponding to Europe. Modern Asian populations also show some Oceanic ancestry (from the ancestral population denoted by red color). Modern Central and South Asian populations show an admixture of European and East Asian ancestral populations. The Middle Eastern populations show contributions from the ancestral African population and the ancestral European population. Modern Oceanic populations are an admixture of an ancestral Oceanic population with an ancestral East Asian population. All these observations are in agreement with previous analyses of the data by Rosenberg et al. (2002) and other studies of regional populations. We should note that the clusters inferred by StructHDP are not identical to the ones observed by Rosenberg et al. for \( K = 4 \). Rosenberg et al. observe that East Asia separates out into a separate cluster for \( K = 4 \) while Oceania separates from the rest of the data only for values of \( K \) larger than 4.

To analyze the similarity and differences within and between continental divisions, we computed the mean ancestry proportions for the seven continental divisions by averaging the ancestry proportions for all individuals belonging to each continental division. We then constructed a distance matrix by computing the Euclidean distance between the 4D vectors representing each continental division. Figure 8 shows the resulting distance matrix. From the figure, we can see that the distance matrix has a block structure. Modern American and East Asian populations are similar to each other and show little separation. We also see that modern European, Central South Asian and Middle Eastern populations are close to each other. Within these three divisions, we see that Europeans and Middle Eastern populations group together while the Central South Asians are further apart.

We hypothesized that if the inferred ancestry proportions capture the genetic variation between and across populations, then the pairwise Euclidean distance computed earlier should be correlated with genetic distance. To test this hypothesis, we computed the pairwise \( F_{ST} \) distance between the seven continental divisions of the data. To test for correlation between the pairwise Euclidean distance matrix and the pairwise \( F_{ST} \) distance matrix, we used a Mantel test. A Mantel test tests the alternate hypothesis of correlation between two matrices against the null hypothesis of no correlation.
by permuting the rows and columns of one of the matrices and observing the distribution of the correlation statistic. The Mantel test on the Euclidean and $F_{st}$ distance matrices shows that the correlation between the two distance matrices is 0.57 ($P=0.0025$ with 10 000 replicates). The distribution of the observed and simulated Mantel correlation statistic is shown in Figure 9. Thus, we can see that the Euclidean distance and $F_{st}$ distance are strongly positively correlated, which supports the inferred population structure.

To compare our results on the HGDP data with other methods, we analyzed the data using Structurama. However, due to computational reasons, we were unable to run Structurama on the full data at optimal settings. Therefore, we analyzed a subset of the data that included only 100 loci per individual. We found that the posterior distribution of $K$ inferred by Structurama has non-zero mass only at $K=5$. Figure 10 shows the inferred ancestry proportions based on the mean partition from Structurama. We can see that Structurama also clusters the European, Middle Eastern and Central South Asian populations into a single cluster. However, since it does not allow partial membership, the individuals in different clusters have zero similarity. It is, therefore, unable to model the partial similarity between populations from different geographical divisions, e.g. the Central Asian populations and European populations.

Fig. 9. The distribution of the Mantel correlation statistic for the pairwise Euclidean distance matrix and the pairwise $F_{st}$ distance matrix. The stem indicates the observed value of the statistic. The result is significant, with the associated $P=0.0025$.

Fig. 10. The ancestry proportions for the 1048 individuals from the Human Genome Diversity Project inferred by Structurama. Each thin line denotes the ancestry proportions for a single individual. Different colors correspond to different ancestral populations. Dark black lines separate individuals from different major geographical divisions. The geographical divisions are indicated by the labels on top of the graph.

6 DISCUSSION

We have presented StructHDP, a method for automatically inferring the number of population clusters present in a group of individuals while accounting for admixture between populations. At the same time, it also infers individual ancestry estimates under a Structure-like model. We demonstrated the effectiveness of our method on data simulated from an island model. We also analyzed the Taït thrush dataset and demonstrated that StructHDP chooses the number of clusters that best explain the data. Our analysis of the HGDP dataset shows that our method is able to cluster populations even when the individuals in the dataset are admixed. The ancestry proportions inferred for populations can be used to compute a distance measure between populations. We found that the Euclidean distance between populations has a strong positive correlation with the $F_{st}$ distance between populations. The ancestry proportions, therefore, provide a useful low-dimensional representation of populations.

Our method uses an HDP to model the admixture of an unknown number of ancestral populations present in individual genomes in a given dataset. We use an MCMC sampling algorithm, Gibbs sampling, to estimate the model parameters. The number of ancestral populations that best explain the data is one of the parameters of our model. The collapsed Gibbs sampler we implemented according to Teh et al. (2006) marginalizes the uncertainty in the population allele frequencies, thus eliminating a possible source of error in the inference. Our experiments suggest that the HDP is not sensitive to the priors on the parameters $\alpha$ and $\gamma$ since we sample them in the algorithm. The results are more sensitive to the choice of $\lambda$ for the base distributions. A large value of $\lambda$ tends to produce populations with uniform (high-entropy) allele frequency distributions while a small value of $\lambda$ produces populations with allele frequency distributions highly skewed in favor of very few alleles (low-entropy). A complete sensitivity analysis, however, is beyond the scope of this article. It would also be instructive to perform simulations with more realistic and complex demographic models to understand the limitations of StructHDP.

The model as described here can handle both single nucleotide polymorphism (SNP) and microsatellite markers. However, one of the limitations of our method is the computational time required for the Gibbs sampling. This means that while our method can handle datasets of a few thousand markers and individuals, it cannot be efficiently used on large datasets of hundreds of thousands of markers. However, as our simulations show, even with few loci, the method performs well at recovering the number of populations required to explain the data best. Teh et al. (2008) have described a way of implementing collapsed variational inference for HDPs.
Applying the variational inference algorithm to StructHDP would improve its speed significantly.

In this work, we have shown how the basic admixture model can be extended to allow automatic inference of the number of populations. Just as extensions to the Structure model that account for recombination (Falush et al., 2003) and mutation (Shringarpure and Xing, 2009) have been developed, we can also extend StructHDP to model other evolutionary processes.

Genetic datasets are often accompanied by geographical information about the genotyped individuals. In some cases, there is a single geographical label associated with each individual, while in others, there are labels at different resolutions (for example, region, nation, continent). It has been shown that geographical distance correlates well with genetic distance between populations (Cavalli-Sforza et al., 1994; Novembre et al., 2008; Ramachandran et al., 2005). Therefore, the amount of sharing of ancestral population components between modern population groups is likely to depend on their geographical distance.

In its current form, StructHDP does not make use of geographical information in the inference process. Teh et al. (2009) describe how an HDP can be extended to include multiple levels of hierarchy and be generalized to a tree-like hierarchy. Use of the hierarchical geographical labels and distances.

In this work, we have shown how the basic admixture model that account for recombination (Falush et al., 2003) and mutation (Shringarpure and Xing, 2009) have been developed, we can also extend StructHDP to model other evolutionary processes.

APPENDIX A—INFERENCES AND ESTIMATION

The population allele frequencies at locus $i$ are assumed to be $\{\phi_1, \ldots, \phi_K\}$ where $K$ is infinity and only a finite number of the populations are used in the dataset. The prior over the allele frequencies $\phi_k$ is $H_i$. In the restaurant analogy, we use $x_{ji}$ to denote the table for customer $xy$, $n_{jk}$ to denote the number of customers in restaurant $j$ at table $t$ eating dish $k$, and $n_{jk}$ denotes the number of tables in restaurant $j$ serving dish $k$. Marginal counts are represented with dots. So, $n_{jk}$ denotes the number of customers in restaurant $j$ at table $t$, and $m$ represents the total number of tables in the franchise.

Let $x = (x_{ji}; \forall j, i)$, $x_{ji} = (x_{ji}; \forall j, i)$ and $x_{ji} = (x_{ji}; \forall j, i)$, where $x_{ji}$ is the number of observations in group $j$ at table $t$, and $m$ represents the total number of tables in the franchise. Let $x = (x_{ji}; \forall j, i)$, $x_{ji} = (x_{ji}; \forall j, i)$, and $x_{ji} = (x_{ji}; \forall j, i)$, where $x_{ji}$ is the number of observations in group $j$ associated with table leaving out observation $x_{ji}$.

An important quantity we will use often in sampling is the conditional density of $x_{ji}$ under mixture component $k$ given all data except $x_{ji}$. This can be computed as

$$f_{ji|\phi_k}(x_{ji}) = \frac{f(x_{ji}|\phi_k) \prod_{i=1}^m f(x_{ji}|\phi_k) d\phi_k}{f(x_{ji}|\phi_k) \prod_{i=1}^m f(x_{ji}|\phi_k) d\phi_k}$$

(A12)

Here, we are marginalizing out the effects of the uncertainty in the allele frequencies $\phi_k$. For our purposes, $f(x_{ji}|\phi_k)$ is a multinomial distribution and $h_{ji}(x_{ji})$ is a symmetric Dirichlet distribution with parameters $\lambda$, on the simplex of dimension $p$ if we observe $P$ different alleles at locus $i$. Therefore, the numerator and denominator conditional distributions are

**Fig. A11.** Graphical model representation of StructHDP with all priors represented. The shaded circle indicates the observed alleles. The dataset has $N$ individuals, each genotyped at $M$ loci. For ease of representation, we do not show the pleiotropism of the individual in the graphical model. The diamonds indicate parameters that are supplied by the user.

**REFERENCES**

Akaike, H. (1974) A new look at the statistical model identification. IEEE Trans. Automatic Control, 19, 716–723.

Alexandre, D.H. et al. (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res., 19, 1655–1664.

Cavalli-Sforza, L.L. et al. (1994) The History and Geography of Human Genes. Princeton University Press, Princeton, NJ.

Escofet M.D. and West, M. (1995) Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc., 90, 377.

Falush D. et al. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 164, 1567–1587.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. Ann. Stat., 1, 209–230.

Gabourel, P. et al. (2000) Genetic variability and gene flow in the globally critically endangered Taita thrush. Conserv. Genet., 1, 45–55.

Hudson, R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics, 18, 337–338.

Hurelbrunck, J. and Andolfatto, P. (2007) Inference of population structure under a Dirichlet process prior. Genetics, 175, 1877–1892.

Novembre, J. et al. (2008) Genes mirror geography within Europe. Nature, 456, 98–101.

Paterson, N. et al. (2006) Population structure and eigenanalysis. PlasGenet, 2, e1950.

Pelizzola, M. and Marsili, M. (2006) The Gibbs and split-merge sampler for population mutation analysis from genetic data with incomplete baselines. Can. J. Fish. Aquat. Sci., 63, 576–586.

Pritchard, J.K. et al. (2000) Inference of population structure from multilocus genotype data. Genetics, 155, 945–955.

Ramachandran, S. et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. USA, 102, 15942–15947.
in terms of the Gamma function:

\[ f(x_j|\phi_k) = \prod_{p} \frac{x_{ij}^{\phi_{jk} - 1}}{\Gamma(\phi_{jk})} \]

(A13)

Using this in Equation (A12) gives us,

\[ f_{m|z}^{(z)}(\phi_k) = \frac{B(h_1 + \sum_{j\in\mathcal{S}_m} x_{ij})}{B(h_1 + \sum_{j\in\mathcal{S}} x_{ij})} \]

where \( B(\cdot) \) is the multinomial beta function, which can be written in terms of the Gamma function:

\[ B(a_1, \ldots, a_p) = \frac{\prod_{i=1}^p \Gamma(a_p)}{\Gamma(\sum_{i=1}^p a_p)} \]

Sampling for the population indicator variables \( z \) is given by

\[ p(z_{ij} = 1 | \mathbf{z}, \mathbf{\beta}) = \frac{\Gamma(\phi_{jk} + a_0) }{\Gamma(\sum_{j=1}^K \phi_{jk} + \sum_{j\neq j} a_0)} \]

where \( a_0 \) is previously used

To sample \( m \), we use a result derived in Teh et al. (2006),

\[ p(m|\mathbf{z}, \mathbf{\beta}) = \frac{\Gamma(\phi_{jk} + a_0)}{\Gamma(\phi_{jk} + a_0 + m_0)} \]

where \( \beta_{m,n} \) are unsigned Sterling numbers of the first kind.

Sampling for \( \beta \) is given by

\[ (\beta_1, \ldots, \beta_k, \beta_0) \sim \text{Dir} (m_1, \ldots, m_K, \gamma) \]

**CONCENTRATION PARAMETER UPDATES**

For updating the concentration parameter \( a_0 \), we use the method described by Teh et al. (2006), using a sampling scheme of auxiliary variables. For \( N \) individuals, define auxiliary variables, \( \mathbf{w} = (w_j)_{j=1}^N \) and \( s = (s_j)_{j=1}^N \), where each \( w_j \in [0, 1] \) and each \( s_j \) is a binary variable in \([0, 1]\). Then we have the following sampling scheme

\[ q(a_0 | w, s) \sim \text{Gamma} \left( a + m + 1 - \sum_j q_j, b + 1 - \sum_j \log(p_j) \right) \]

\[ q(w_j | a_0) \sim \text{Beta} (a_0 + 1, n_j) \]

\[ q(s_j | a_0) \sim \text{Binomial} (1, n_j / a_0 / (1 + n_j / a_0)) \]

To update \( a_0 \), we iterate these three steps until the value of \( a_0 \) converges. Convergence is usually quick and takes about 20–30 iterations.

For updating \( \gamma \), we use the method described in Escobar and West (1995), using an auxiliary variable \( \eta \). Assume \( \gamma \) has a gamma prior \( \text{Gamma} (a, b) \).

We have

\[ q(\gamma | \eta, K) \sim \pi_\gamma \text{Gamma}(a + k, 1/(b - \log(\eta))) + (1 - \pi_\gamma) \text{Gamma}(a + K - 1, 1/(b - \log(\eta))) \]

where the mixture weights are given by

\[ \pi_\gamma = \frac{a + k - 1}{m_0 (b - \log(\eta))} \]

Secondly, we have

\[ q(\eta | \gamma, K) \sim \text{Beta}(\gamma + 1, m_{\cdot}) \]

Alternating these updates until the value of \( \gamma \) converges provides a method for updating \( \gamma \).