# A Spatial Framework for Understanding Population Structure and Admixture

Gideon S. Bradburd[1], Peter L. Ralph[2,¶], Graham M. Coop[3,¤a],
**1 Center for Population Biology, Department of Evolution and Ecology, University of California, Davis, CA, USA**
**2 Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA**
**3 Center for Population Biology, Department of Evolution and Ecology, University of California, Davis, CA, USA**
*** E-mail: gbradburd@ucdavis.edu**

## Abstract

Geographic patterns of genetic variation within modern populations, produced by complex histories of migration, can be difficult to infer and visually summarize. A general consequence of geographically limited dispersal is that samples from nearby locations tend to be more closely related than samples from distant locations, and so genetic covariance often recapitulates geographic proximity. We use genome-wide polymorphism data to build "geogenetic maps", which, when applied to stationary populations, produces a map of the geographic positions of the populations, but with distances distorted to reflect historical rates of gene flow. In the underlying model, allele frequency covariance is a decreasing function of geogenetic distance, and nonlocal gene flow such as admixture can be identified as anomalously strong covariance over long distances. This admixture is explicitly co-estimated and depicted as arrows, from the source of admixture to the recipient, on the geogenetic map. We demonstrate the utility of this method on a circum-Tibetan sampling of the greenish warbler (*Phylloscopus trochiloides*), in which we find evidence for gene flow between the adjacent, terminal populations of the ring species. We also analyze a global sampling of human populations, for which we largely recover the geography of the sampling, with support for significant histories of admixture in many samples. This new tool for understanding and visualizing patterns of population structure is implemented in a Bayesian framework in the program SpaceMix.

## Author Summary

In this paper, we introduce a statistical method for inferring, for a set of sequenced samples, a map in which the distances between population locations reflect genetic, rather than geographic, proximity. Two populations that are sampled at distant locations but that are genetically similar (perhaps one was recently founded by a colonization event from the other) may have inferred locations that are nearby, while two populations that are sampled close together, but that are genetically dissimilar (e.g., are separated by a barrier), may have inferred locations that are farther apart. The result is a "geogenetic" map in which the distances between populations are effective distances, indicative of the way that populations perceive the distances between themselves: the "organism's-eye view" of the world. Added to this, "admixture" can be thought of as the outcome of unusually long-distance gene flow; it results in relatedness between populations that is anomalously high given the distance that separates them. We depict the effect of admixture using arrows, from a source of admixture to its target, on the inferred map. The inferred geogenetic map is an intuitive and information-rich visual summary of patterns of population structure.

# Introduction

There are many different methods to learn how population structure and demographic processes have left their mark on patterns of genetic variation within and between populations. Model-based approaches focus on developing a detailed view of the migrational history of a small number of populations, often assuming one or a small number of large, randomly mating populations (i.e. little or no geographic structure). There has been considerable recent progress in this area, using a variety of summaries such as the allele frequency spectrum [Gutenkunst et al.(2009), Bhaskar et al.(2014), Excoffier et al.(2013)], or approximations to the coalescent applied to sequence data [Paul et al.(2011), Li and Durbin (2011)].

Other approaches are designed only to visualize patterns of genetic relatedness and population structure, without using a particular population genetic model. Such methods can deal with many populations or individuals as the unit of analysis. Examples of this second set of methods include clustering methods [Pritchard et al.(2000), Alexander et al.(2009), Lawson et al.(2012)] and reduced dimensionality representations of the data, such as Principal Components Analysis (PCA; e.g. [Luca et al.(1994), Patterson et al.(2006), Price et al.(2006)]).

A third set of methods that describe relatedness between populations by constructing a "population phylogeny" was pioneered by [Cavalli-Sforza and Edwards (1967)], as were methods to test whether a tree is a good model of population history [Cavalli-Sforza and Piazza (1975)] (see [Felsenstein (1982)] for a review). Tree-based approaches are appealing because trees are easy to visualize and explain, but the underlying assumptions (unstructured populations that split at discrete points in time) rarely hold true.

Recently, there has been a resurgence of interest in these tree-based methods. Some use population trees as a null model to test and quantify the signal of admixture between samples [Reich et al.(2009)]. Others, such as TreeMix [Pickrell and Pritchard (2012)] and MixMapper [Lipson et al.(2013)], visualize population relationships using a directed acyclic graph; for instance, TreeMix connects branches in a population tree with additional edges to explain excess covariance between groups of populations.

There has also been renewed interest in methods for dimensionality reduction for the visualization of patterns of genetic variation [Patterson et al.(2006)], especially PCA (also pioneered by Cavalli-Sforza [Menozzi et al.(1978)]). Examining such low-dimensional visual summaries has become an indispensable step in the analysis of modern genomic datasets of thousands of loci typed in tens or hundreds of samples. Generally, these visualizations are constructed by plotting the first few eigenvectors of the covariance matrix of normalized allele frequencies against each other.

Both PCA and tree-based methods are valuable as genetic inference and visualization tools, but both also suffer from serious limitations. Because gene flow is frequently pervasive, patterns of relatedness between samples may often be only poorly represented by a tree-based model. PCA is more flexible, as it assumes no explicit model of population-genetic processes, simply describing the axes of greatest variance in the average coalescent times between pairs of samples [McVean (2009)]. This allows PCA to describe more geographically continuous relationships: PCA applied to human populations within continents often shows a close correspondence to geographic locations [Novembre et al.(2008), Wang et al.(2012)]. However, the interpretation of PCA is more difficult, as the results can be strongly affected by the size and design of sampling, and the linearity and orthogonality requirements of the PC axes can lead to counterintuitive results [Novembre and Stephens (2008), Francois et al.(2010), Frichot et al.(2012)].

What is desired, then, is a method for inferring and visualizing patterns of population differentiation that can recapitulate complex, non-hierarchical structures, while also admitting simple and intuitive interpretation. Since gene flow and population movements are often constrained by geography, it is natural to base such a method in a geographic framework. There is a rich history of population genetics theory for populations distributed in continuous space [Malécot(1975), Nagylaki (1978), Felsenstein (1975), Barton et al.(2002)], as well as exciting new developments in the field [Petkova et al.(2014)]. The pattern of increasing genetic differentiation with geographic distance was termed "Isolation by Distance" by Wright [Wright (1943)], and is ubiquitous in natural populations [Meirmans (2012)]. Descriptive models of such patterns rely only on the weak assumption that an individual's mating opportunities are spatially limited by dispersal; a large set of models, ranging from equilibrium migration-drift models to non-equilibrium models, such as recent spatial expansions of populations, give rise to the empirical pattern of isolation by distance.

In this paper, we present a statistical framework for studying the spatial distribution of genetic variation and genetic admixture based on a flexible parameterization of the relationship between genetic and geographic distances. Within this framework, the pattern of genetic relatedness between the samples is represented by a map, in which inferred distances between samples are proportional to their genetic differentiation, and long distance relatedness (in excess of that predicted by the map) is modeled as genetic admixture. These 'geogenetic' maps are simple, intuitive, low-dimensional summaries of population structure, and provide a natural framework for the inference and visualization of spatial patterns of genetic variation and the signature of genetic admixture. The implementation of this method, SpaceMix, is available at https://github.com/gbradburd/SpaceMix.

# Results

**Data** The genetic data we model consist of allele counts at $L$ unlinked, bi-allelic single nucleotide polymorphisms (SNPs), sampled across $K$ populations. After arbitrarily choosing an allele to count at each locus, denote the number of counted alleles at locus $\ell$ in population $k$ as $C_{k,\ell}$, and the total number of alleles observed as $S_{k,\ell}$. The sample frequency at locus $\ell$ in population $k$ is $\hat{f}_{k,\ell} = C_{k,\ell}/S_{k,\ell}$. Although we will refer to "populations", each could consist of a single individual ($S_{k,\ell} = 2$ for a diploid). We will depict results as coordinates on a map; however, the method does not require user-specified sampling locations.

We first compute standardized sample allele frequencies at locus $\ell$ in population $k$, by

$$\hat{X}_{k,\ell} = (\hat{f}_{k,\ell} - \bar{f}_\ell)/\sqrt{\bar{f}_\ell(1 - \bar{f}_\ell)}, \tag{1}$$

where $\hat{f}_{k,\ell}$ is the sample allele frequency at locus $\ell$ in population $k$, and $\bar{f}_\ell$ is the average of the $K$ sample allele frequencies, weighted by mean population size. This normalization is widely used [Nicholson et al.(2002), Patterson et al.(2006)]; mean-centering makes the result invariant to choice of which allele to count at each locus, and dividing by $\sqrt{\bar{f}_\ell(1 - \bar{f}_\ell)}$ makes each locus have roughly unit variance in some sense.

We work with the empirical covariance matrix of these standardized sample allele frequencies, calculated across loci, namely, $\widehat{\Omega} = (1/L)\hat{X}\hat{X}^T$. Using the sample mean to mean-center $\hat{X}$ has implications on their covariance structure, discussed in the Methods ("The standardized sample covariance"). For clarity, here we proceed as if $\bar{f}_\ell$ were instead an unobserved, global mean allele frequency at locus $\ell$.

**Spatial Covariance Model** We wish to model the distribution of alleles among populations as the result of a spatial process, in which migration moves genes locally on an unobserved landsape. Migration homogenizes those differences between populations that arise through genetic drift; populations with higher levels of historical or ongoing migration share more of their demographic history, and so have more strongly correlated allele frequencies.

We assume that the standardized sample frequencies are generated independently at each locus by a spatial process, and so have mean zero and a covariance matrix determined by the pairwise geographic distances between samples. To build the geogenetic map, we arbitrarily choose a simple and flexible parametric form for the covariance matrix in which covariance between allele frequencies decays exponentially with a power of their distance [Diggle et al.(1998), Wasser et al.(2004), Bradburd et al.(2013)]: the covariance between standardized population allele frequencies (i.e. $\hat{X}$ values) between populations $i$ and $j$ is assumed to be, for $i \neq j$,

$$F(D_{i,j}) = \frac{1}{\alpha_0}\exp\left(-\left(\alpha_1 D_{i,j}\right)^{\alpha_2}\right), \tag{2}$$

where $D_{i,j}$ is the geogenetic distance between populations $i$ and $j$, $\alpha_0$ controls the within-population variance (or the covariance when distance between points is 0, known as a "sill" in the geospatial literature), $\alpha_1$ controls the rate of the decay of covariance per unit pairwise distance, and $\alpha_2$ determines the shape of that decay. Within population variance may vary across samples due to either noise from a finite sample size or demographic history unique to that sample (e.g., bottlenecks or endogamy). To accommodate this heterogeneity we introduce population-specific variance terms, resulting in the covariance matrix for standardized sample frequencies

$$\Omega_{i,j} = F(D_{i,j}) + \delta_{i,j}\left(\frac{1}{\bar{S}_i} + \eta_i\right), \tag{3}$$

where $\delta_{i,j} = 1$ if $i = j$ and is 0 otherwise, $\eta_k$ is a nonnegative sample-specific variance term (nugget) to account for variance specific to population $k$ that is not accounted for by the spatial model, and $\bar{S}_k$ is the mean sample size across all loci in population $k$, so that $1/\bar{S}_k$ accounts for the variance introduced by sampling within the population.

The distribution of the sample covariance matrix $\hat{\Omega}$ is not known in general, but the central limit theorem implies that if the number of loci is large, it will be close to Wishart. Therefore, we assume that $\hat{\Omega}$ is Wishart distributed with degrees of freedom equal to the number of loci ($L$) used and mean equal to the parametric form $\Omega$ given in equation (3). We denote this by

$$P(\widehat{\Omega} \mid \Omega) = \mathcal{W}\left(L\widehat{\Omega} \mid \Omega, L\right). \tag{4}$$

Note that if the standardized sample frequencies are Gaussian, then the sample covariance matrix is a sufficient statistic, so that calculating the likelihood of $\widehat{\Omega}$ is the same as calculating the likelihood of the data up to a constant. Handily, it also means that once the sample covariance matrix has been calculated, all other computations do not scale with the number of loci, making the method scalable to genome size datasets.

**Location Inference** Non-equilibrium processes like long distance admixture, colonization, or population expansion events will distort the relationship between covariance and distance across the range, as will barriers to dispersal on the landscape. To accommodate these heterogeneous processes we infer the locations of populations on a map that reflects genetic, rather than geographic, proximity. To generate this map, we treat populations' locations (i.e. coordinates in the geogenetic map) as parameters that we estimate with a Bayesian inference procedure (described in the Methods). These location parameters for each population are denoted by $G$, and determine the matrix of pairwise geogenetic distances between populations, $D(G)$, which together with the parameters $\vec{\alpha}$ and $\eta$ determine the parametric covariance matrix $\Omega$ (given by equation (3)). We acknowledge this dependence by writing $\Omega(\vec{\alpha}, D(G), \eta)$.

The prior distributions on the parameters that control the shape and scale of the decay of covariance with distance ($\vec{\alpha}$ and $\eta$) are given in the Methods. The priors on the geogenetic locations, $G$, are independent across populations; because the observed locations naturally inform the prior for populations locations, we use a very weak prior on population $k$'s location parameter ($G_k$) that is centered around the observed location. This prior on geogenetic locations also encourages the resulting inferred geogenetic map to be anchored in the observed locations and to represent (informally) the minimum distortion to geographic space necessary to satisfy the constraints placed by genetic similarities of populations. In practice, we also compare results to those produced using random locations as the "observed" locations, and can change the variance on the spatial priors to ascertain the effect of the prior on inference.

We then write the posterior probability of the parameters as

$$P\left(G, \vec{\alpha}, \eta \mid \widehat{\Omega}, L\right) \propto P\left(\widehat{\Omega} \mid \Omega(\vec{\alpha}, D(G), \eta)\right) P(\vec{\alpha}) P(G) P(\eta), \tag{5}$$

where $P()$ denotes the various priors, and the constant of proportionality is the normalization constant.

We then use a Markov chain Monte Carlo algorithm to estimate the posterior distribution on the parameters as described in more detail in the Methods.

**Illustrated with Simulations.** We first apply the method to several scenarios simulated using the coalescent simulator *ms* [Hudson (2002)]. Each scenario is simulated using a stepping stone model in which populations are arranged on a grid with symmetric migration to nearest neighbors (eight neigbors, including diagonals) with 10 haploid individuals sampled from every other population at 10,000 unlinked loci (for details on all simulations, see Methods and Supplementary Materials). The basic scenario is shown in Fig. 1a, which is then embellished in various ways. In the SpaceMix analysis of each simulated dataset, we treat population locations as unknown parameters to be estimated as part of the model, and center the priors on each population's location at a random point. The resulting geogenetic maps are produced from the parameters having maximum posterior probability. Since overall translation, rotation, and scale are nuisance parameters, we present inferred locations after a Procrustes transformation (best-fit rotation, translation, and dilation) to match the coordinates used to simulate the data.

**Figure 1** Simulation scenarios and and their corresponding geogenetic maps estimated with SpaceMix. The smaller circles in the simulation scenarios represent unsampled populations. a) configuration of simulated populations on a simple lattice with spatially homogeneous migration rates; b) a lattice with a barrier along the center line of longitude, across which migration rates are reduced by a factor of 5; c) a lattice with recent expansion on the eastern margin; d) the maximum *a posteriori* (MAP) estimate from the posterior distribution of population locations under the scenario in 1a; e) MAP

estimate of population locations under the scenario in 1b; f) MAP estimate of population locations under the scenario in 1c;

The lattice scenarios, illustrated in Figs. 1 and 2, are: homogeneous migration rates across the grid; a longitudinal barrier across the center of the grid; a series of recent expansion events; and an admixture event between opposite corners of the lattice. In the simple lattice scenario with homogeneous migration rates (Figs. 1a and 1d), SpaceMix recovers the lattice structure used to simulate the data (i.e., populations correctly choose their nearest neighbors). After adding a longitudinal barrier to dispersal across which migration rates are reduced by a factor of 5 (Fig. 1b), the two halves of the map are pushed farther away from one another, reflecting the decreased gene flow between them.

In the expansion scenario, in which all populations in the last five columns of the grid have expanded simultaneously in the immediate past from the nearest population in their row (Fig. 1c), the daughter populations of the expansion event cluster with their parent populations, reflecting the higher relatedness (per unit of geographic separation) between them. In all scenarios, populations at the corners of the lattice are pulled in somewhat because these have the least amount of data informing their relative placements. In Fig. S1, we show the relationship between genetic covariance, geographic distance, and inferred geogenetic distance for these simulations.

We next simulated a long-distance admixture event on the same grid, by sampling half of the alleles of each individual in the northeast corner population from the southwest corner population (Fig. 2a). We then ran a SpaceMix analysis in which the locations of these populations were estimated (Fig. 2b). The admixture creates excess covariance over anomalously long distances, which is clearly difficult to accommodate with a two-dimensional geogenetic map. Fig. 2b shows the torturous lengths to which the method goes to fit a good geogenetic map: the admixed population 30 is between population 1, the source of its admixture, and populations 24, 25, and 29, the nearest neighbors to the location of its non-admixed portion. However, this warping of space is difficult to interpret, and would be even more so in empirical data for which a researcher does not know the true demographic history.

**Figure 2** Simulation scenarios and SpaceMix inference. a) a lattice with recent admixture event between population 1 in the southwest corner and population 30 in the northeast corner, so that population 30 is drawing half of its ancestry from population 1; b) the estimate of population locations under this scenario; c) the estimate of population locations and their sources of admixture under this scenario. The 95% credible interval on $w_{30}$ is 0.36–0.40. In panel (c), the width and opacity of the admixture arrows are drawn proportional the admixture proportions.

## Inference of Spatial Admixture

To incorporate recent admixture, we allow each allele sampled in population $k$ to have a probability $w_k$ ($0 \leq w_k \leq 0.5$) of being sampled from location $G_k^*$, which we refer to as population $k$'s source of admixture, and a probability $1 - w_k$ of being sampled from location $G_k$. With no nugget, each allele would be sampled independently, but the nugget introduces correlations between the alleles sampled in each population.

With this addition, the parametric covariance matrix before given by (3) becomes a function of all the pairwise spatial covariances between the locations of populations $i$ and $j$ and the points from which they draw admixture (illustrated in Fig. 3); now, we model the covariance between $\hat{X}_{k,\ell}$ and $\hat{X}_{k,\ell}$, for each $\ell$, as

$$\begin{aligned}\Omega^*_{i,j} &= (1-w_i)(1-w_j)F(D_{i\ ,\ j}\ ) \\ &+ w_i(1-w_j)F(D_{i^*,\ j}\ ) \\ &+ w_j(1-w_i)F(D_{i\ ,\ j^*}) \\ &+ w_i w_j F(D_{i^*,\ j^*}) \\ &+ \delta_{i,j}(\eta_i + 1/\bar{S}_i)\end{aligned} \tag{6}$$

where $D$ is the $2k \times 2k$ matrix of pairwise distances between all inferred locations and sources of admixture, and for readability, we denote, e.g., $F(D(G_i, G^*{}_j))$, as $F(D_{i\ ,\ j^*})$. The spatial covariance, $F(D)$, is as given in equation (2), and we reintroduce the nugget, $\eta_k$, and the sample size effect, $1/\bar{S}_k$, for each population as above in Eqn. (3).

**Figure 3** An illustration of the form of the admixed covariance given in Equation (6). Populations $i$ and $j$ are drawing admixture in proportions $w_i$ and $w_j$ from their respective sources of admixture, $i^*$ and $j^*$, and all pairwise spatial covariances (the $F$'s) are shown. In this cartoon example, population $j$ is drawing more admixture from its source $j^*$ than $i$ is from its source $i^*$ (i.e., $w_j > w_i$).

We proceed in our inference procedure as before, but now with the locations of the sources of admixture and the admixture proportions to infer. The likelihood of the sample covariance matrix is exactly as before in (4), except with $\Omega$ replaced by $\Omega^*$. The posterior probability of these parameters can be expressed as a function of this parametric admixed covariance, $\Omega^*$,

$$P(G, G^*, w, \vec{\alpha}, \eta \mid \widehat{\Omega}, L) \propto P(\widehat{\Omega} \mid \Omega^*)P(\vec{\alpha})P(G)P(G^*)P(w)P(\eta) \tag{7}$$

as specified by the parameters $w$, $G^*$, $\vec{\alpha}$, and $\eta$, and the inferred locations, $G$. We place a weak spatial prior on the sources of admixture, $G^*$ around the centroid of the observed locations. The admixture proportions, $w$, are capped at 0.5, to ensure identifiability, and are heavily weighted towards small values to be conservative with respect to admixture inference. These priors are detailed in Table 2.

The models described above may be used in various combinations. In the simplest model, populations do not choose their own locations, nor are they allowed to draw admixture; the only parameters to be estimated are those of the spatial covariance function given in equation (2), and the population-specific variance terms ($\eta_i$). In the most complex model, population locations, the locations of their sources of admixture, and the proportions of admixture are all estimated jointly in addition to the parameters of the spatial covariance function and the population specific variances. Users may wish to employ the more constrained models (e.g., fixing the locations or admixture proportions for some or all samples) in a model selection framework to test specific hypotheses.

Allowing admixture gives sensible results for the scenario of Fig. 2a: in the resulting map, the only population that draws substantial admixture is the one that is actually admixed, and it draws admixture (95% CI: 0.36 - 0.40) from the correct location (Fig. 2c).

A more subtle simulated admixture scenario, with admixture proportion of 10% across a geographic barrier, is shown Fig. 4a. The resulting SpaceMix map (Fig. 4b), separates the east and west sides of the grid to accommodate the effect of the barrier, and the admixed population (population 23) chooses admixture from very close to its true source (population 13), and in close to the correct amount ($\bar{w}_{(23)} = 0.05$; 95% CI $= 0.02 - 0.08$).

Another difficult scenario is shown in Fig. 4c, where 40% admixture has occurred between two populations immediately adjacent to each other on either side of a barrier. Here, the admixed population 18 is correctly identified as admixed, but it explains its intermediate genetic relationships by taking a location close to its true admixture source (population 13), and drawing admixture (95% CI: 0.04–0.14) from a location on the far margin of the half of the grid on its own side of the barrier. Because there is no sampled intervening population between admixed population 18 and its source of admixture 13, there is nothing to stop 18 from explaining its higher covariance with 13 via its chosen location $G_{(18)}$ rather than via that of its source of admixture $G^{*}_{(18)}$. In each of these scenarios, the estimated admixture proportion is less than that used to simulate the data. This is due to the stringent prior we place against admixture. We discuss these examples further in the Methods.

**Figure 4** Simulation scenarios and inferred population maps for two different admixture scenarios: a) lattice with a barrier and an admixture event (10%) across the barrier to an 'inland' population; b) the inferred population map for the scenario in (a), where the admixed population 23 is the only population drawing non-negligible admixture (95% CI: 0.02-0.08); c) lattice with a barrier and an admixture event (40%) across the barrier to a 'neighbor' population on the border of the barrier; (d) the inferred population map for the scenario in (c), where the admixed population 18 is the only population drawing non-negligible admixture (95% CI: 0.04–0.14).

# Empirical Applications

To demonstrate the applications of this novel method, we analyzed population genomic data from two systems: the greenish warbler ring species complex, and a global sampling of contemporary human populations. Maps showing our sampling in these two systems are given in Fig. 5, and information on the specific samples included is given in the Supplementary Materials, Tables S1 and S2. For all analyses presented below, we centered the priors on location parameters at randomly chosen locations rather than at the observed geographical locations. Each geogenetic map shown here is the maximum a posteriori estimate (over all parameters), transformed by rotation, translation, and scaling to best fit inferred locations ($G$) to the observed latitude and longitudes (a full Procrustes transformation).

**Figure 5** Sampling maps of both empirical systems analyzed. (a) greenish warbler subspecies distributions of all 22 sampled populations (breeding grounds), consisting of 95 individuals and colored by subspecies [Alcaide et al.(2014)]; (b) sampling map for human dataset, consisting of 1,490 individuals from 95 population samples [Hellenthal et al.(2014)].

## Greenish Warblers

The greenish warbler (*Phylloscopus trochiloides*) species complex is broadly distributed in their breeding habitat around the Tibetan plateau, and exhibits gradients around the ring in a range of phenotypes including song, as well as in allele frequencies [Ticehurst (1938),Irwin et al.(2001),Irwin et al.(2005)]. At the northern end of the ring in central Siberia, where the eastern and western arms of population expansion meet, there are discontinuities in call and morphology, as well as reproductive isolation and a genetic discontinuity [Irwin et al.(2001), Irwin et al.(2008)]. It is proposed that the species complex represents a ring species, in which selection and/or drift, acting in the populations as they spread northward on either side of the Tibetan plateau, have led to the evolution of reproductive isolation between the terminal forms.

The question of whether it fits the most strict definition of a ring species focuses on whether gene flow around the plateau has truly been continuous throughout the history of the expansion or if, alternatively, discontinuities in migration around the species complex's range have facilitated periods of differentiation in genotype or phenotype without gene flow [Mayr (1942), Mayr (1970), Coyne (2004)] (see Wake and Schneider [Wake and Schneider (1998)] for discussion). Alcaide *et. al* [Alcaide et al.(2014)] have suggested that the greenish warbler species complex constitutes a 'broken' ring species, in which historical discontinuities in gene flow have facilitated the evolution of reproductive isolation between adjacent forms.

To investigate this question, we applied SpaceMix to the dataset from Alcaide *et. al* [Alcaide et al.(2014)], consisting of 95 individuals sampled at 22 distinct locations and sequenced at 2,334 SNPs, of which 2,247 were bi-allelic and retained for SpaceMix runs. The libraries were prepared using a genotype-by-sequencing protocol and were run on an Illumina HiSeq 2000 with a paired-end sequencing protocol [Alcaide et al.(2014)].

We first ran SpaceMix on the population dataset, with no admixture. The resulting inferred map (Fig. 6a) largely recapitulates the geography of the sampled populations around the ring. The Turkish population (**TU**, *Phylloscopus trochiloides* ssp. *nitidus*) clusters with the populations in the subspecies *ludlowi*, due to its recent expansion, but also chooses a relatively high nugget parameter (see Fig. S2a), reflecting the population history it does not share with its *ludlowi* neighbors. In the North, where the twin waves of expansion around the Tibetan Plateau are hypothesized to meet, the inferred geogenetic distance between populations from opposite sides of the ring was much greater than their observed geographic separation, reflecting the reproductive isolation between these adjacent forms (see Fig. S3).

**Figure 6** Inferred population maps with population labels colored as in Fig. 5a: a) the map inferred with no admixture inference; b) the map inferred with admixture inference.

We then ran the method allowing admixture (Fig. 6b). The only population sample with appreciable admixture is the Stolby sample (**ST**; $w = 0.19$, 95% credible interval: 0.146-0.238; Fig. S4). This sample is known to be composed of an equal mixture of eastern *plumbeitarsus* and western *viridanus* individuals [Alcaide et al.(2014)]. Multiple runs agreed well on the level of admixture of the Stolby sample (see Fig. S5). What does vary across runs is whether the Stolby sample chooses to locate itself by the *viridanus* cluster and draw admixture from near the *plumbeitarsus* cluster or vise versa; however, this is to be expected given the 50/50 nature of the sample's makeup (Fig. S5). The somewhat intermediate position of the Stolby sample, and its non-50/50 admixture proportion, likely partially reflect the influence of the priors (Fig. S6).

We repeated these analyses (with and without admixture) on an individual level (Fig. 7). No individual drew appreciable admixture (see Fig. S7 for admixture proportions), and so we discuss the results with admixture (those without admixture are nearly identical, see Figs. S8, S9, and S10). As with the analysis on multi-sample populations, the results approximately mirror the geography of the individuals.

**Figure 7** Inferred maps for warbler individuals, colored by subspecies in an analysis with admixture inference. a)map inferred with admixture; b) close-up of all non-*nitidus*

samples in the admixture map.

There are, however, a number of obvious departures in the individual geogenetic map from the population map. The most obvious is that the location of a pair of *nitidus* samples (in purple) is very far from the rest of the samples. These individuals appear to be fairly close relatives: in the population-level analysis of Fig. S2a, this increase in shared ancestry was accounted for by a large nugget for the *nitidus* population; but in the individual-level analysis, they cannot share a nugget parameter, and must therefore choose a location close to each other and far from the rest of the samples. The same phenomenon seems to be at work in determining the locations of a pair of individuals, one identified as *P. t. ludlowi* (Lud-MN3), one as *P. t. trochiloides* (Tro-LN11), as they also show an unusually low pairwise sequence divergence (see Fig. S11).

The split between *viridanus* and *plumbeitarsus* individuals (blue and red, respectively), in the north at the contact zone of the two waves of expansion, is clearer now than in the population-based analysis, as individuals from the Stolby population have moved to near their respective clusters. Despite the fact that *viridanus* and *plumbeitarsus* individuals have moved away from each other in our geogenetic map, they are still closer to each other than we would expect if all gene flow between the two was mediated by the southern populations, in which case we would expect the populations to form a line, with *viridanus* at one end and *plumbeitarsus* at the other. This horseshoe, with *viridanus* and *plumbeitarsus* at its tips, is steady within and among runs of the MCMC and choice of position priors (see Figs. S9a-S9c).

Is this biologically meaningful? A similar horseshoe shape appears when a principal components (PC) analysis is conducted and individuals are plotted on the first two PCs (see Fig. S12 and [Alcaide et al.(2014)]). However, as discussed by Novembre and Stephens [Novembre and Stephens (2008)], such patterns in PC analysis can arise for somewhat unintuitive reasons. If populations are simulated under a one dimensional stepping stone model, then plotting individuals on the first two PCs results in a horseshoe (e.g. see Fig. S13b) not because of gene flow connecting the tips, but rather because of the orthogonality requirement of PCs (see [Novembre and Stephens (2008)] for more discussion). In contrast, when SpaceMix is applied to data simulated on a one dimensional array of populations, the placement of samples is consistent with a line (see Figs. S13c, S13d). The proximity of *viridanus* and *plumbeitarsus* in geogenetic space may be due to gene flow between the tips of the horseshoe north of the Tibetan Plateau. This conclusion is in agreement with that of Alcaide *et al.* [Alcaide et al.(2014)], who observed evidence of hybridization between *viridanus* and *plumbeitarsus* using assignment methods.

The SpaceMix map also diverges from the observed map in the distribution of individuals from the subspecies *ludlowi* (in green). These samples were taken from seven sampling locations along the southwest margin of the Tibetan Plateau, but, in the SpaceMix analysis, they partition into two main clusters, one near the *trochiloides* cluster, and one near the *viridanus* cluster. This break between samples from the same subspecies, which is concordant with the findings of Alcaide *et al.* [Alcaide et al.(2014)], makes the *ludlowi* cluster unusual compared to the estimated spatial distributions of the other subspecies (see Fig. S14), and suggests a break in historic or current gene flow.

## Human Populations

Human population structure is a complex product of the forces of migration and drift acting on both local and global scales, patterned by geography [Novembre et al.(2008),Ralph and Coop (2013)], time [Skoglund et al.(2012),Skoglund et al.(2014)], admixture [Hellenthal et al.(2014)], landscape and environment [Beall et al.(2010), Bigham et al.(2010), Bradburd et al.(2013)], and shaped by culture [Reich et al.(2009), Atzmon et al.(2010), Moorjani et al.(2011)]. To visualize the patterns these processes have induced, we create a geogenetic map for a worldwide sample of modern human populations. Of course, human history at these geographic scales has many aspects that are not well captured by static maps with discrete "arrows" of admixture. Nonetheless, we talk about the locations of samples and their sources of admixture as if these are fixed, even though both reflect the compounding of drift and gene flow over many historical processes. We therefore urge caution in the interpretation our results, and view them as a simplistic but rich visualization of patterns of population structure.

We used a random subset of 10,000 SNPs from the dataset of Hellenthal *et al.* [Hellenthal et al.(2014)], comprised of 1,490 individuals from 95 population samples (see Fig. 5b for map of sampling), as well as the latitude and longitude attributed to each sample. We ran two sets of SpaceMix analyses: in the first, we estimated population sample locations, and in the second, we also allowed admixture. We note that few of the putative admixture events that we report have escaped the notice of previous investigators, which is unsurprising given the depth of recent attention on human admixture studies, particularly on the subset of HGDP samples [Rosenberg et al.(2002), Li et al.(2008), Loh et al.(2013), Patterson et al.(2012), Hellenthal et al.(2014)]. Below, when discussing a pattern we see in our analyses, we often cite other authors who have seen or suggested similar patterns. However, what is novel here is the ability to visualize these admixture events in a geographic context, and that these admixture signals stand out against a null model of migration in continuous space (rather than tree-based models).

When we only infer the location of each sample, the map roughly recapitulates the geography of the samples (Fig. 8a), a result that holds nicely when we zoom in on the more heavily sampled area of Eurasia (Fig. 8b). We see that samples both in the Americas and in Oceania lie close to the East Asian samples, but that they form two clusters on opposite sides. The proximity of these groups to the East Asians represents the fact that both groups share an ancestral population in the relatively recent past with East Eurasian populations, but the two expansions occurred independently. As in our simulations (Fig. 1f) population expansions/bottlenecks have distorted the relationship between geographic and geogenetic distance. Geogenetic distances between samples within Africa are much greater than those between any other group (see Fig. S15), and the slope of the relationship between geographic and geogenetic distances between populations on each continent decays with distance from Africa. This pattern is consistent with a history of human colonization events characterized by serial bottlenecks [Harpending and Rogers (2000), Prugnolle et al.(2005), Ramachandran et al.(2005)] following an out-of-Africa expansion, and subsequent expansions into Western Eurasia, East Asia, the Americas, and Oceania (although see Pickrell and Reich [Pickrell and Reich (2014)] for a discussion of other models).

**Figure 8** Map of human samples, inferred without admixture.
(a) complete map; (b) close-up of Eurasian samples.

To investigate possible patterns of admixture further, we ran a SpaceMix analysis with admixture (results shown in Figs. 9 and 10). The biggest change between the geogenetic map of human populations inferred with admixture and that without is the positioning of African samples with respect to the rest of the world. The relatively large geogenetic distances between these groups reflects the fact that Eurasian, North African, Oceanian, and American populations all share relatively large amounts of population history (and hence genetic drift) not shared with the Sub-Saharan African samples.

Relative to the geogenetic map inferred without admixture, the inclusion of admixture allows samples that fall intermediate between Sub-Saharan Africa and North Africa and the Middle East to move closer to one or the other, which, in turn, allows each of those major clusters to move relatively farther apart. The Ethiopian and Ethiopian Jewish samples move to be closer to the Sub-Saharan samples than the of the North African samples, but draw substantial amounts of admixture ($\sim 40\%$) from close to where the Egyptian sample has positioned itself in the the Middle East cluster, as do the Sandawe [Hodgson et al.(2014), Pickrell et al.(2012)]. The SanKhomani draw admixture from near Syria, which may reflect multiple distinct geographic sources of admixture as discussed by [Hellenthal et al.(2014)] and [Pickrell et al.(2014)]. Interestingly the Bantu South African sample, though it moves to join the other Bantu samples, draws admixture from close to the San populations. This is consistent with previous signals of the expansion of Bantu-speaking peoples into southern Africa [Pickrell et al.(2012), Schlebusch et al.(2012), Pickrell et al.(2014), Hellenthal et al.(2014)]. The inferred sample-specific drift parameters (the 'nuggets') are similar between runs with and without admixture (Fig. S16).

**Figure 9** Map of human samples, inferred with admixture. (a) complete map; (b) close-up of Eurasian samples. Italicized labels denote locations of admixture sources, with darkness proportional to the amount of admixture.

**Figure 10** Mean admixture proportions (and 95% CIs) for each population sample.

The majority of North African samples (Egyptian, Tunisian, Morocan, Mozabite) join the Middle Eastern samples (positioned in rough accord with their sampling location along North Africa), and draw admixture from near the Ethiopian samples. All of the Middle Eastern samples draw admixture from close to the geogenetic location of the Ethiopian samples and where most of the North African samples draw admixture from, representing the complex history of North African–Middle Eastern gene flow [Henn et al.(2012), Hellenthal et al.(2014)].

A number of other population samples draw admixture from Africa. The Sindhi, Makrani, and Brahui draw admixture from close to the location of the Bantu samples [Hellenthal et al.(2014)], and the Balochi and Kalash draw admixture from some distance away from African population samples. Of the European samples, the Spanish and both East and West Sicilian samples draw small amounts of admixture from close to the Ethiopian samples, presumably reflecting a North African ancestry component [Moorjani et al.(2011), Botigué et al.(2013)].

The other significant signal of admixture is between East and West Eurasia, a signal documented by many authors [Rosenberg et al.(2002), Li et al.(2008), Xu and Jin (2008), Hellenthal et al.(2014)]. The majority of samples maintain their relative positions within each of these groups; however, several of the populations whose inferred locations are intermediate between eastern and western Eurasia (in the SpaceMix analysis without admixture) now move towards one side and draw admixture from the other. The Uzbekistani and Hazara samples move to be closer to the East Asian samples, while drawing a substantial admixture proportion from close to where the Georgian and Armenian samples have located themselves, while conversely the Uygur sample moves to be close to the Burusho, Kalash, and Pathan samples. The Tu sample (with a geogenetic location in East Asia) draws a small amount of ancestry from close to where the Uygur have positioned themselves. The Chuvash move close to Russian and

Lithuanian samples, drawing admixture from close to the Yakut; the Turkish sample also draws a smaller amount of admixture from there. There are several other East-West connections: the Russian and Adygei samples have admixture from a location “north” of the East Asian samples, and the Cambodia sample draws admixture from close to the Eygptian sample [Pickrell and Pritchard (2012),Hellenthal et al.(2014)].

There are also a number of samples that draw admixture from locations that are not immediately interpretable. For example, the Hadza and Bantu Kenyan samples draw admixture from somewhat close to India, and the Xibo and Yakut from close to “northwest” of Europe. The Pathan samples draw admixture from a location far from any other samples' locations, but close to where the India samples also draws admixture from. The Myanmar and the Burusho samples both draw admixture far from the locations estimated for other samples as well.

There are a number of possible explanations for these results. As we only allow a single admixture arrow for each sample, populations with multiple, geographically distinct sources of admixture may be choosing admixture locations that average over those sources. This may be the case for the Hadza and Bantu Keynan samples [Hellenthal et al.(2014)]. A second possibility is that the relatively steep prior on admixture proportion forces samples to choose lower proportions of admixture from locations that overshoot their true sources; this may explain the Xibo and Yakut admixture locations. A final explanation is that good proxies for the sources of admixture may not be included in our sampling, either because of of the limited geographic sampling of current day populations, or because of old admixture events from populations from which there are not other more directly descending modern populations. The admixture into the Indian and Pathan samples (whose admixture source also clusters with the Indian Jew samples in some MCMC runs) may be an example of this; Reich *et al.* [Reich et al.(2009)] and Moorjani *et al.* [Moorjani et al.(2013)] have hypothesized that many populations from the Indian subcontinent may be descended from an admixture event involving an ancestral Southern Indian population not otherwise represented in this dataset.

In Figs. S17 and S18, we show the results of other independent MCMC analyses on these data. The broad-scale patterns and results discussed above are consistent across these runs. However, as is to be expected, there is significant heterogeneity in the exact layout of sample and admixture locations. For example, there is some play, among MCMC runs, in the internal orientation of the African locations with respect to Eurasia. Samples that draw a significant amount of admixture, such as the central Asian populations (Uygur, Hazara and Uzbekistani), switch their location with that of their source of admixture (as was also seen across MCMC runs in the warbler data analysis). Similarly the Ethiopian and Ethiopian Jew samples choose locations, in some MCMC runs, close to the other North African samples, and draw admixture from near the Sub-Saharan samples (as do the other North African samples).

## Discussion

In this paper we have presented a statistical framework for modeling the geography of population structure from genomic sequencing data. We have demonstrated that the method, SpaceMix, is able to accurately present patterns of population structure in a variety of simulated scenarios, which included the effects of spatially heterogeneous migration, population expansion, and population admixture. In empirical applications of SpaceMix, we have largely recovered previously estimated population relationships in a circum-Tibetan sample of greenish warblers and in a global sample of human populations, while also providing a novel way to depict these relationships. The geogenetic maps SpaceMix generates serve as simple, intuitive, and information-rich summaries of patterns of population structure. SpaceMix combines the advantages of other methods for inferring and illustrating patterns of population structure, using model-based inference to infer population relationships (like TreeMix [Pickrell and Pritchard (2012)], and MixMapper [Lipson et al.(2013)]), and producing powerful visualizations of genetic structure on a map (like PCA [Patterson et al.(2006)] and SPA [Yang et al.(2014)]).

The patterns of genetic variation observed in modern populations are the product of a complex history of demographic processes. We choose to model those patterns as the outcome of a spatial process with geographically determined migration, and we have included statistical elements to accommodate deviations from spatial expectations. However, the true history of a sample of real individuals is vastly more complex than any low-dimensional summary, and, as with any summary of population genetic data, SpaceMix results should be interpreted with this in mind. Furthermore, our "admixture" events are shorthands for demographic relationships that occurred over possibly substantial lengths of time and regions of the globe; approximating this by a single arrow between two points on a map is certainly an oversimplification. Aspects of population history that are better described as a population phylogeny may be difficult to interpret using SpaceMix, and may be better suited to visualization with hierarchical clustering-based methods [Pritchard et al.(2000)] or TreeMix/MixMapper-like methods [Pickrell and Pritchard (2012),Lipson et al.(2013)]. There is obviously no one best approach to studying and visualizing population structure; investigators should employ a range of appropriate methods to identify those that provide useful insight.

SpaceMix offers much of the flexibility of PCA – like PCA, it is well suited to describing population structure in a continuous fashion – but it also has a number of advantages over PCA. When isolation by distance holds, the first (one or) two PCs often correspond reasonably well to some simple rotation of latitude and longitude; however, these first two PCs explain a relatively small part of the total variance of the data. Furthermore, because PCs are linear functions of the genotypes, sometimes many PCs must be used to depict patterns produced by simple isolation by distance [Novembre and Stephens (2008)]. These higher order PCs can be hard to interpret in empirical data (see discussion in the warbler section). The recently introduced SPA approach [Yang et al.(2012)], which also assumes allele frequencies are monotonically increasing in a given direction, may suffer from the same problem (although we note that PCA and SPA both have significant speed advantages over SpaceMix).

In comparison, if isolation by distance holds, then the two dimensions in which SpaceMix infers geogenetic positions for the samples will suffice to capture the geographic patterns of genetic differentiation (to the extent to which the parametric form of the covariance is flexible enough to capture the empirical decay of covariance with distance). The application of SpaceMix to humans nicely illustrates the utility of our approach: the first two PCs of this dataset resemble a boomerang (Fig. S19), with its arms corresponding to the Africa/Non-Africa split and the spread of populations across Eurasia. In contrast, while the SpaceMix geogenetic map is dominated by the genetic drift induced by migration out of Africa, it also captures much more detail than is contained in the first two PCs (e.g., Fig. 9b). This comparison is also nicely illustrated by the example in Fig. S13.

An advantage of PCA is that it can explain more complex patterns of population structure by allowing up to $K$ different axes. Although SpaceMix can easily be extended to more than two dimensions, simply by allowing $G_i$ to describe the location of a sample in $d$ dimensions, the interpretation and visualization of these higher dimensions would prove difficult, and so for the moment we stick with two dimensions. On the other hand, SpaceMix can describe in two dimensions patterns that PCA, due to the constraints of linearity, would need more to describe.

Another strong advantage of SpaceMix over current methods is the introduction of admixture arrows. Although PCA can be interpreted in light of simple admixture events [McVean (2009)], and recent methods can locate the recent, spatially admixed ancestry of out-of-sample individuals [Yang et al.(2012), Yang et al.(2014)], neither approach explicitly models admixture between multiple geographically distant locations, as SpaceMix does. Assignment methods are designed to deal with many admixed samples [Pritchard et al.(2000)], but they have no null spatial model for testing admixture. We feel that an isolation by distance null model is often more appropriate for testing for admixture, especially when there is geographically dense sampling. SpaceMix offers a useful tool to understand and visualize spatial patterns of genetic relatedness when many samples are admixed.

As currently implemented, SpaceMix allows each population to have only a single source of admixture, but some modern populations draw substantial proportions of their ancestry from more than two geographically distant regions. In such cases the inferred source of admixture may fall between the true locations of the parental populations. Although it is statistically and computationally feasible to allow each population to choose more than one source of admixture, we were concerned about both the identifiability and the interpretability of such a model, and have not implemented it. However, there may be empirical datasets in which such a modeling scheme is required to effectively map patterns of population structure. In addition, we have assumed that only single populations are admixed, when in fact it is likely that particular admixture events may affect multiple samples.

One concern is that the multiple admixed samples (from a single admixture event) may simply choose to cluster close to each other, and not need to draw admixture from elsewhere due to the fact that their frequencies are well described by their proximity to other admixed populations. Along these lines, it is noticeable that many of our European samples draw little admixture from elsewhere (also noted by [Hellenthal et al.(2014)] using a different approach), despite evidence of substantial ancient admixture [Lazaridis et al.(2014)]. This may reflect the fact that all of the European samples are affected by the admixture events, and are relatively over-represented in our sample. However, this may also simply reflect the fact that the admixture is ancient, and that the ancient populations that took part in these events are not well represented by our extant sampling. Reassuringly, we see multiple cases where similarly admixed populations (Central Asians, Middle Eastern, and North African) populations are separately identified as admixed. This suggests that geogenetic clustering (in lieu of drawing admixture) of populations that share similar histories of admixture is not a huge concern (at least in some cases). The method could in theory be modified to allow geogenetically proximal populations to draw from the same admixture event; however, this may be difficult to make fully automated.

In this paper we have focused on the covariance among alleles at the same locus, but linkage disequilibrium (LD; covariance of alleles among loci) holds rich information about the timing and source of admixture events [Chakraborty and Weiss (1988), Moorjani et al.(2013), Hellenthal et al.(2014), Gravel (2012)] as well as information about isolation by distance [Ralph and Coop (2013)]. Just as population graph approaches have been extended to incorporate information from LD [Loh et al.(2013)], a spatial covariance approach could be informed by LD. A null model inspired by models of LD under isolation by distance models [De and Durrett (2007), Barton et al.(2013)] could be fitted, allowing the covariance among alleles to decay with their geographic distance and the recombination distance between the loci. In such a framework, sources and time-scales of admixture could be identified through unusually long-distance LD between geographically separated populations.

The landscape of allele frequencies on which the location of populations that were the source of population's admixture are estimated is entirely informed by the placement of other modern samples, even though the admixture events may have occurred many generations ago. This immediately leads to the caveat that, instead of "location of the parental population," we should refer to the "location of the closest descendants of the parental population." The increased sequencing of ancient DNA (see Pickrell and Reich [Pickrell and Reich (2014)] for a recent review) promises an interesting way forward on that front, and it will also be exciting to learn where ancient individuals fall on modern maps, as well as how the inclusion of ancient individuals changes the configuration of those maps [Skoglund et al.(2014)]. The inclusion of ancient DNA samples in the analyzed sample offers a way to get better representation of the ancestral populations from which the ancestors of modern samples received their admixture. However, it is also possible to model genetic drift as a spatiotemporal process, in which covariance in allele frequencies decays with distance in both space and in time. We are currently exploring using ancient DNA samples as 'fossil calibrations' on allele frequency landscapes at points in the past, so that modern day samples may draw admixture from coordinates estimated in spacetime.

# Methods

Here we describe in more detail the algorithm we use to estimate the posterior distribution defined by (7) of the population locations, $G$, their sources of admixture, $G^*$, their admixture proportions, $w$, their independent drift parameters, $\eta$, and the parameters of the model of isolation by distance, $\vec{\alpha}$. First, we give the exact form of the covariance matrix we use, and then describe the Markov chain Monte Carlo algorithm that samples parameter values from the posterior distribution.

## The standardized sample covariance

As motivation, consider several randomly mating (Wright-Fisher) populations that all split from an ancestral population in which a neutral allele is present at frequency $\epsilon_\ell$, and then subsequently exchange migrants. Since the allele is neutral, the mean change in its frequency in each population after $t$ generations is zero, and if $t$ is much smaller than the population size (so the frequencies remain close to $\epsilon_\ell$), the variance is proportional to $\sqrt{\epsilon_\ell(1-\epsilon_\ell)}$. Conveniently, additional variance introduced by binomial sampling of alleles is also proportional to $\sqrt{\epsilon_\ell(1-\epsilon_\ell)}$. It would then be natural to consider the covariance matrix of

$$X_{k,\ell} = \frac{\hat{f}_{k,\ell} - \epsilon_\ell}{\sqrt{\epsilon_\ell(1-\epsilon_\ell)}}, \tag{8}$$

since these standardized allele frequencies would be independent if the loci are unlinked, and would have mean zero and variance independent of the sample sizes or allele frequencies. The central limit theorem would then imply that in the limit of a large number of loci, the sample covariance matrix $X^T X$ is Wishart with degrees of freedom equal to the number of loci and mean determined by the pattern of migration.

Although the conditions are not strictly met, these theoretical considerations indicate that such a normalization may be a reasonable thing to do, even after substituting the empirical mean allele frequency $\bar{f}_\ell$ in place of $\epsilon_\ell$, which is what we do to define $\hat{X}_{k,\ell}$ in equation (1). Recall that the sample allele frequency at locus $\ell$ in population $k$ is given by $\hat{f}_{k,\ell} = C_{k,\ell}/S_{k,\ell}$, where $C_{k,\ell}$ is the number of (arbitrarily chosen) counted alleles, and $S_{k,\ell}$ is the total number of sampled alleles. As sample size may vary across loci, we first calculate $\bar{S}_k$, the mean sample size in population $k$, as

$\bar{S}_k = \frac{1}{L}\sum_{\ell=1}^{L} S_{k,\ell}$. We then compute the global mean allele frequency at locus $\ell$ as

$$\bar{f}_\ell = \frac{1}{\sum_K S_{k,\ell}} \sum_K \hat{f}_{k,\ell} S_{k,\ell}. \tag{9}$$

If sample size were constant across all loci in each population, this would be equivalent to defining the variance-normalized sample frequencies

$$\hat{Y}_{k,\ell} = \frac{\hat{f}_{k,\ell}}{\sqrt{\bar{f}_\ell(1-\bar{f}_\ell)}} \tag{10}$$

and writing $\hat{X}_\ell = TY_\ell$ where $T$ is the mean centering matrix whose elements are given by

$$T_{ij} = \delta_{i,j} - \frac{\bar{S}_j}{\sum\limits_{k=1}^{K} \bar{S}_j}, \tag{11}$$

where $\delta_{i,j} = 1$ if $i = j$ and is 0 otherwise. If the covariance matrix of $Y$ is $\Omega^*$, then the covariance matrix of $\hat{X}_\ell$ would be $T^T\Omega^*T$. Since allowing $T$ to vary by locus would be computationally infeasible, we make one final assumption, that the covariance matrix of the standardized frequencies $\hat{X}_\ell$ at each locus is given by $T^T\Omega^*T$. This makes it inadvisable to include loci for which there are large differences in sample sizes across populations. This mean centering acts to to reduce the covariance among populations in $\hat{X}_\ell$ compared to $\hat{f}_\ell$, and can induce negative covariance between more unrelated populations (as, across loci, they are often on opposite sides of the mean).

Additionally, the covariance matrix of the standardized frequencies has rank $K-1$ rather than $K$, and so the corresponding Wishart distribution is singular. To circumvent this problem we compute the likelihood of a $(K-1)$-dimensional projection of the data. Any projection would do; we choose a projection matrix $\Psi$ by dropping the last column of the orthogonal matrix in the QR decomposition of $T$, and compute the likelihood of the empirical covariance matrix of allele frequencies $\widehat{\Omega} = \hat{X}^T\hat{X}$ as

$$P(\widehat{\Omega} \mid \Omega^*) = \mathcal{W}\left(L\Psi^T X^T X\Psi \mid \Psi^T\Omega^*\Psi, L\right). \tag{12}$$

## Markov chain Monte Carlo Inference Procedure

The inference algorithm described here may be used to estimate the parameters with any of these held fixed, for instance: (1) population locations are fixed, and they do not draw any admixture; (2) populations may choose their own locations, but not admixture; (3) populations may draw admixture, but their own locations are fixed; or (4) populations may both choose their own locations and draw admixture. The free parameters for each of options are given in Table 1.

Below, we outline the inference procedure for the most parameter-rich model (inference on both population locations, their sources of admixture, and the proportions in which they draw admixture, in addition to inference of the parameters of the spatial covariance function). A table of all parameters, their descriptions, and their priors is given in Table 2.

We now specify in detail the Markov chain Monte Carlo algorithm we use to sample from the posterior distribution on the parameters, for Bayesian inference.

We assume that the user has specified the following data:

- the allelic count data, $C$, from $K$ population over $L$ variant loci, where $C_{k,\ell}$ gives the number of observations of a given allele at locus $\ell$ in population $k$.

| Model | # of Free Parameters | Parameters |
|---|---|---|
| stationary population locations, no admixture | $K+3$ | $\alpha_0, \alpha_1, \alpha_2, \eta$ |
| inferred population locations, no admixture | $2K+3$ | $\alpha_0, \alpha_1, \alpha_2, \eta, G$ |
| stationary population locations, inferred admixture | $2K+3$ | $\alpha_0, \alpha_1, \alpha_2, \eta, G^*, w$ |
| inferred population locations, inferred admixture | $3K+3$ | $\alpha_0, \alpha_1, \alpha_2, \eta, G, G^*, w$ |

**Table 1.** List of models that may be specified using SpaceMix, along with the number and identity of free parameters in each.

| Parameter | Description | Prior |
|---|---|---|
| $\boldsymbol{\alpha_0}$ | controls the sill of the covariance matrix | $\alpha_0 \sim Exp(\lambda = 1/100)$ |
| $\boldsymbol{\alpha_1}$ | controls the rate of the decay of covariance with distance | $\alpha_1 \sim Exp(\lambda = 1)$ |
| $\boldsymbol{\alpha_2}$ | controls the shape of the decay of covariance with distance | $\alpha_2 \sim U(0.1, 2)$ |
| $\boldsymbol{\eta_k}$ | the nugget in population $k$ (population specific drift parameter) | $\eta_k \sim Exp(\lambda = 1)$ |
| $\boldsymbol{G_k}$ | the geogenetic location of population $k$ | $G_k \sim \mathcal{N}(\mu = G_k^{(obs)}, \sigma = \frac{1}{2}\bar{D}(G^{(obs)}))$ |
| $\boldsymbol{w_k}$ | the proportion of admixture in population $k$ | $2w_k \sim \beta(\alpha = 1, \beta = 100)$ |
| $\boldsymbol{G_k^*}$ | the geogenetic location of the source of admixture in population $k$ | $G_k^* \sim \mathcal{N}(\mu = \bar{G}^{(obs)}, \sigma = 2\bar{D}(G^{(obs)}))$ |

**Table 2.** List of parameters used in the SpaceMix models, along with their descriptions and priors. $\bar{D}(G^{(obs)})$ is the mean of the pairwise distances between observed locations $G^{(obs)}$.

- the sample size data, $S$, from $K$ population over $L$ variant loci, where $S_{k,\ell}$ gives the total number of alleles typed at locus $\ell$ in population $k$.

It is not necessary, but a user may also specify

- the geographic sampling locations, $G^{(obs)}$, from each of the $K$ populations, where $G_k^{(obs)}$ gives the longitude and latitude of the $k^{\text{th}}$ sampled individual(s).

The geographic location data may be missing, or generated randomly, for some or all of the samples; if so, the spatial priors on estimated population locations, $G$, and their sources of admixture, $G^*$ will not be tethered to the true map.

**Initiating the MCMC** We then calculate the standardized sample covariance matrix $\widehat{\Omega}$ as described in the section "The standardized sample covariance" above, as well as $\bar{S_k}$, the mean sample size across loci for each population. Armed with the standardized sample covariance, the geographic sampling locations, and the inverse mean sample sizes across samples ($\widehat{\Omega}$, $G^{(obs)}$, $1/\bar{S_k}$), we embark upon the analysis.

To initiate the chain, we specify starting values for each parameter. We draw initial values for $\alpha_0$, $\alpha_1$, $\alpha_2$, $\eta$, and $w$ randomly from their priors. We initiate $G$ at user-specified geographic sampling locations and $G^*$ at randomly drawn, uniformly distributed values of latitude and longitude in the observed range of both axes.

**Overview of MCMC procedure** We use a Metropolis-Hastings update algorithm. In each iteration of the MCMC, one of our current set of parameters $\Theta = \{\alpha_0, \alpha_1, \alpha_2, \eta, w, G, G^*\}$ is randomly chosen to be updated by proposing a new value. In the cases of $\{\eta, w, G, G^*\}$, where each population has its own parameter, a single population, $k$ is randomly selected and only its parameter value (e.g. $\eta_k$) is chosen to be updated. Below we outline the proposal distributions for each parameter. This gives us a proposed update to our set of parameters $\Theta'$, which differs from $\Theta$ at only one entry.

The set of locations of populations and their sources of admixture specify a pairwise geographic distance matrix $D$, which, given the current $\vec{\alpha}$ and $\eta$ parameters, gives the admixed covariance matrix described in (6), $\Omega^*$. The likelihood of the two sets of parameters $\Theta$ and $\Theta'$, calculated with (12) and the priors of Table 2, combine to give the Metropolis-Hastings ratio, $R$, the probability of accepting the proposed parameter values $\Theta'$:

$$R = \min\left(1, \frac{P(\widehat{\Omega} \mid \Omega^*(\Theta'))}{P(\widehat{\Omega} \mid \Omega^*(\Theta))} \frac{P(\Theta')}{P(\Theta)}\right), \tag{13}$$

Note that all of our moves, described below, are symmetric, so the Hastings ratio is always 1. If we accept our proposed move, $\Theta$ is replaced by $\Theta'$ and this is recorded, otherwise $\Theta'$ is discarded and we remain at $\Theta$.

**Updates for $\vec{\alpha}$, $\eta$, and $w$** We propose updates to the values of the $\vec{\alpha}$, $\eta$, and $w$ parameters via a symmetric normal density with mean zero and variance given by a tuning parameter specific to that parameter. For example, $\alpha_0' \sim \alpha_0 + \delta$, where $\delta \sim \mathcal{N}(0, \sigma^2_{\alpha_0})$ and $\sigma_{\alpha_0}$ is the tuning parameter for $\alpha_0$. For $\eta$ and $w$, each of which consists of $K$ parameters, each parameter receives its own independent tuning parameter. If the proposed move takes the parameter outside the range of its prior, the move is rejected and we do not move in that iteration of the MCMC.

**Updates for geographic coordinates $G$ and $G^*$** Updates to the location parameters, $G$ and $G^*$, are somewhat more complicated due to the curvature of the Earth. Implementing updates via a symmetric normal density on estimated latitude and longitude directly would have the drawback of a) being naive to the continuity of the spherical manifold and b) vary the actual distance of the proposed move as a function of the current lat/long parameter values (e.g., a 1° change in longitude at the equator is a larger distance than at the North Pole).

Instead, we propose a bearing and a distance traveled, and, given these two quantities and a starting position, calculate the latitude and longitude of the proposed update to the location. For example, in an update to the location of population $i$, $G_i$, we propose a distance traveled $\Delta_{G_i}$, where, e.g., $\Delta_{G_i} \sim |\mathcal{N}(0, \sigma_{G_i})|$, and a bearing, $\gamma$, where $\gamma \sim U(0, 2\pi)$. Then we use the following equations to calculate the latitude and longitude of the proposed location:

$$\begin{aligned} \text{proposed latitude} = \arcsin(&\sin(\text{current latitude}) \times \\ &\cos(\Delta) \times \cos(\text{current latitude}) \times \\ &\sin(\Delta) \times \cos(\gamma)) \end{aligned} \tag{14}$$

and

$$\begin{aligned} &\text{proposed longitude} = \text{current longitude} \\ &\quad - \arctan\left(\frac{\sin(\gamma)\sin(\Delta)\cos(\text{current latitude})}{\cos(\Delta) - \sin(\text{current latitude})\sin(\text{proposed latitude})}\right), \end{aligned} \tag{15}$$

where latitude and longitude are given in radians and are taken $\mod 2\pi$. As with $\eta$ and $w$, each population's location and admixture source location parameters have their own tuning parameters.

**Adaptive Metropolis-within-Gibbs proposal mechanism** We use an adaptive Metropolis-within-Gibbs proposal mechanism on each parameter [Roberts and Rosenthal (2009), Rosenthal (2011)]. This mechanism attempts to maintain an acceptance proportion for each parameter as close as possible to 0.44, which is optimal for one-dimensional proposal mechanisms [Roberts et al.(1997), Roberts and Rosenthal (2001)]. We implement this mechanism by creating, for each variable $i$, an associated variable $\zeta_i$, which gives the log of the standard deviation of the normal distribution from which parameter value updates are proposed. As outlined above, in the cases of $\{\eta$, $w$, $G$, $G^*\}$, for which each population receives a free parameter, each population gets its own value of $\zeta$.

When we start our MCMC, $\zeta_i$ for all parameters is initiated at a value of 0, which gives a proposal distribution variance of 1. We then proceed to track the acceptance rate, $r_i$ for each parameter in windows of 50 MCMC iterations, and, after the $n$th set of 50 iterations, we adjust $\zeta_i$ by an "adaption amount", which is added to $\zeta_i$ if the acceptance proportion in the $n$th set of 50 iterations ($r_i^{(n)}$) has been above 0.44, and subtracted from $\zeta_i$ if not. The magnitude of the adaption amount is a decreasing function of the index $n$, so that updates to $\zeta_i$ proceed as follows:

$$\zeta_i^{n+1} = \begin{cases} \zeta_i^n + \min(\min(0.01, n^{-\frac{1}{2}}), 20), & \text{if } r_i^{(n)} \ > \ 0.44 \\ \zeta_i^n - \min(0.01, n^{-\frac{1}{2}}), & \text{if } r_i^{(n)} \ < \ 0.44 \end{cases} \tag{16}$$

We choose to cap the maximum adaption amount at 20 (which is the equivalent of capping the variance of the proposal distribution at $4.85 \times 10^8$) to avoid proposal distributions that offer absurdly large or small updates. This procedure, also referred to as "auto-tuning", results in acceptance rate plots like those shown in Fig. S20, and in more efficient mixing and decreased autocorrelation time of parameter estimates in the MCMC.

## Simulations

We ran our simulations using a coalescent framework in the program *ms* (Hudson). Briefly, we simulated populations on a lattice, with nearest neighbor (separated by a distance of 1) migration rate $m_{i,j}$, as well as migration on the diagonal of the unit square at rate $m_{i,j}/\sqrt{2}$. For each locus in the dataset, we used the **-s** option to specify a single segregating site, and then we simulated 10,000 loci independently, which were subsequently conglomerated into a single dataset for each scenario. For all simulations, except the "Populations on a line" scenario (Fig. S13), we sampled only every other population, and, from each population, we sampled 10 haplotypes (corresponding to 5 diploid individuals). In the "Populations on a line" scenario, we simulated no intervening populations, such that every population was sampled.

To simulate a barrier event, we divided the migration rate between neighbors separated by the longitudinal barrier by a factor of 5. To simulate an expansion event, we used the **-ej** option to move all lineages from each daughter population to its parent population at a very recent point in the past. For admixture events, we used the **-es** and **-ej** options to first (again, going backward in time) split the admixed population into itself and a new subpopulation of index $k+1$, and second, to move all lineages in the ($k^{\text{th}}+1$) into the source of admixture. Forward in time, this procedure corresponds to cloning the population that is the source of admixture, then merging it, in some admixture proportion, with the (now) admixed population. The command line arguments used to call ms for a single locus for each simulation are included in the Appendix.

## Intuition on identifiability of admixture parameters

A natural concern is whether all of the parameters we infer are separately identifiable, most notably whether population locations, admixture locations, and proportions can be estimated. That is, if a population has received some admixture from another population, what is to stop it from simply moving toward that population in geogenetic space to satisfy its increased resemblance to that population, rather than choosing admixture from that location? We do not provide a formal proof, but here build and illustrate some relevant intuition.

Admixture is identifiable in our model because there are covariance relationships among populations that cannot simply be satisfied by shifting population locations around (as demonstrated by the tortured nature of Fig. 2b). Consider the simple spatial admixture scenario shown in Fig. 11. Populations A–D are arrayed along a line, but there is recent admixture from D into B (such that 40% of the alleles assigned to B are sampled from location D). The lines show the expected covariance under isolation by distance that each population (A, C, or D, as indicated by line color) has with a putative population at a given distance. The dots show the admixed covariance between B and the three other populations, as well as B's variance with itself (B-B) as specified by equation (6), with no nugget or sampling effect.

**Figure 11** Lines show the covariances populations A, C, and D would have with population B as a function of B's location, with no admixture, under the parametric form of equation (2). The colored dots above 'B' show the covariances observed with B at that location if B has 40% admixture from D. There is no single spatial location with unadmixed covariances remotely similar to these.

Due to its admixture from D, B has lower covariance with A than expected given its distance, somewhat higher covariance with C, and much higher covariance with D. In addition, the variance of B is lower than that of the other three populations, which each have variance 1: the value of the covariance when the distance is zero. This lower variance results from the fact that the frequencies at $B$ represent a mixture of the frequency at $D$ and the frequency at $B$ before the admixture.

Now, using this example scenario, let us return to the concern posed above: that admixture location and population location are not identifiable. For the sake of simplicity, assume that we hold the locations of A, C, and D constant, as well as the decay of covariance with distance (as could be the case if A-D are part of a larger analysis). The covariance relationships of $B$ to the other populations cannot be simply satisfied by moving $B$ towards D, as B would then have a covariance with C that is higher, and a covariance with A that is lower, than that we actually observe.

Introducing admixture into the model allows it to satisfy all of these conditions: it can draw ancestry from D but keep part of its resemblance to A, it avoids B having to move closer to C, and it explains B's low variance. Even in the absence of a sample from population C, B is better described as a linear mixture of a population close to A and D. However, there are specific scenarios in which a limited sampling scheme (both in size and location), can lead to tradeoffs in the likelihood between estimated population location and that of its source of admixture.

The analyses depicted in Figs. 2c, 4b, and 4c, give examples of these tradeoffs. In each, the inferred admixture proportions in the admixed populations are less than those used to simulate the data, and the admixed populations are able to explain the high covariance they have with their sources of admixture via their inferred location, rather than just via their inferred source of admixture and admixture proportion. The reason they choose to do explain their anomalous covariance with their inferred location, rather than with their admixture source, is that we place a very harsh prior against admixture inference (Table 2). The prior is designed to make inference conservative with respect to admixture, but it has the side effect of skewing the posterior probability toward lower admixture proportions.

## Empirical Applications

Below, we describe the specifics of our analyses of the greenish warbler dataset and the global human dataset. The analysis procedure for each dataset is given here:

For each analysis,

1. Five independent chains were run for $5 \times 10^6$ MCMC iterations each in which populations were allowed to choose their own locations (but no admixture). Population locations were initiated at the origin (i.e. at iteration 1 of the MCMC, $G_i = (0, 0)$), and all other parameters were drawn randomly from their priors at the start of each chain.

2. The chain with the highest posterior probability at the end of the analysis was selected and identified as the "Best Short Run".

3. A chain was initiated from the parameter values in the last iteration of the Best Short Run. Because inference of admixture proportion and location was not allowed in the five initial runs, admixture proportions were initiated at 0 and admixture locations, $G^*$ were initiated at the origin. This chain (the "Long Run") was run for $10^8$ iterations, and sampled every $10^5$ iterations for a total of 1000 draws from the posterior.

For each dataset, we ran two analyses using the observed population locations as the prior on $G$. Then, to assess the potential influence of the spatial prior on population locations, we ran one analysis in which the observed locations were replaced with random, uniformly distributed locations between the observed minima and maxima of latitude and longitude. For the warbler dataset, we repeated this analysis procedure, treating each sequenced individual as its own population. For clarity and ease of interpretation, we present a full Procrustes superimposition of the inferred population locations ($G$) and their sources of admixture ($G^*$), using the observed latitude and longitude of the populations/individuals ($G$) to give a reference position and orientation. As results were generally consistent across multiple runs for each dataset regardless of the prior employed, we (unless stated otherwise) present only the results from the 'random' prior analyses.

Finally, we compared the SpaceMix map to a map derived from a Principal Components Analysis (Patterson and Reich 2006). For this analysis, we calculated the eigendecomposition of the mean-centered allelic covariance matrix, then plotted individuals' coordinates on the first two eigenvectors (e.g. Novembre et al 2008). For consistency of presentation, we show the full Procrustes superimposition of the PC coordinate space around the geographic sampling locations.

# Supporting Information

## simulated covariance decays

**Decays in covariance for four different simulation scenarios (from top to bottom: simple lattice; lattice with barrier; lattice with expansion; lattice with admixture).** Left column: sample covariance plotted against observed pairwise distance. Right column: sample covariance plotted against inferred geogenetic distance.

## warbler population nugget

**Credible intervals on estimated warbler population nugget parameters.** a) analysis without admixture; b) analysis with admixture.

## warbler population geographic/geogenetic distance comparison

**Comparing geographic to geogenetic pairwise distance between warbler populations: a) observed population coordinates; b) pairwise geographic (great-circle) distance between populations compared to that between their geogenetic locations.** The highlighted points show distances between populations from the *plumbeitarsus* and *viridanus* subspecies. Notice that, regardless of their observed distance, their geogenetic separations are roughly constant, and much larger than the geographic distance between them.

## warbler population admixture proportions

**Credible intervals on estimated warbler population admixture proportion parameters.**

## warbler population geogenetic map comparison

**Comparison of inferred maps from three independent analyses.** (a,b) Results from analysis using observed locations as priors on population locations. (c) Results from analysis using random, uniformly distributed locations within the observed range of latitude and longitude as priors on population locations.

## Admixture proportion and population location likelihood surfaces

**Likelihood surfaces for different placements of population ST between *plumbeitarsus* and *viridanus* clusters: (a) log likelihood surface; (b) posterior probability surface, incorporating the priors.** The maximum a posteriori estimate (MAP) is shown as a star.

## warbler individual admixture proportions

**Credible intervals on estimated warbler individual admixture proportion parameters.**

## warbler individual geogenetic map comparison

**Inferred maps for warbler individuals, colored by subspecies under analyses with and without admixture inference.** a) map inferred without admixture; (b) close-up of all non-*nitidus* samples in non-admixture map; c) map inferred with admixture; d) close-up of all non-*nitidus* samples in the admixture map.

## Posterior distributions on locations of warbler individuals

**Maps of the posterior distributions on population locations in three separate SpaceMix analyses on the warbler individual dataset.**

## warbler individual nuggets

**Credible intervals on estimated warbler individual nugget parameters.** a) analysis without admixture; b) analysis with admixture.

## warbler individual pairwise-$\pi$

**Mean pairwise sequence divergence at polymorphic sites calculated between all pairs of individuals from different subspecies, and colored by the subspecies to which each individual in the comparison is drawn.** Note that individuals Tro-LN11 and Lud-MN3 have sequence divergence that is unusually low relative to that of other comparisons between individuals from the same two subspecies.

## warbler individual PCA map

**The map of warbler individuals derived from a Principal Components analysis.** The PC coordinates have undergone a full Procrustes transformation around the actual sampling coordinates.

## simulation scenario of populations on a line

**Simulation scenario of populations on a line, contrasting PCA-based inference and SpaceMix inference.** a) Scenario used to simulate data in a spatial coalescent framework with nearest-neighbor migration; b) PCA map of allele frequencies, plotting PC axis 1 against PC axis 2, forming a ‘U’ shape; c) Posterior distribution of SpaceMix location inference, forming a rough line; d) snapshot of the MAP draw from the posterior, again showing a rough line.

## Human sample geographic/geogenetic distance comparison

**Comparison of geographic pairwise distance to geogenetic pairwise distance between human populations, colored by continent from which populations were sampled (i.e., two populations sampled from Africa are green).** Eurasia is divided into Western Eurasia and East Asia.

## Human sample nuggets

**Credible intervals on estimated human sample nugget parameters. a) analysis without admixture; a) analysis with admixture.**

## SpaceMix output from analysis of human samples

**Map of human populations from a different SpaceMix analysis ("Real_Prior1" – inferred with admixture), using real geographic coordinates as population location priors.** a) complete map; b) close-up of Eurasian populations.

## SpaceMix output from analysis of human samples

**Map of human populations from another SpaceMix analysis ("Real_Prior2", inferred with admixture), using real geographic coordinates as population location priors.** a) complete map; b) close-up of Eurasian populations.

## PCA map of human samples

**PCA map of human samples used in SpaceMix analyses.** The PC coordinates have undergone a full Procrustes transformation around the actual sampling coordinates (shown in the inset map).

## Example parameter acceptance rates

**Example parameter acceptance proportions for the $\alpha_2$ parameter and the nugget parameter, $\eta$, using the adaptive Metropolis-within-Gibbs proposal mechanism.**

## Warbler Sample Metadata

**Subspecies and geographic meta-data for greenish warbler individuals included in analysis.**

## Human Sample Metadata

**sample size and geographic meta-data for human samples included in analysis**

# Acknowledgments

We would like to gratefully acknowledge the Coop Lab, Yaniv Brandvain, Razib Khan, Dan Runcie, Brad Shaffer, Marjorie Weber, and Will Wetzel for helpful discussions. We also thank Miguel Alcaide, Cristian Capelli, Garrett Hellenthal, and Darren Irwin for sharing data. This work was supported in part by the National Science Foundation under award number NSF #1262645 (DBI) to PR and GC, the National Institute of General Medical Sciences of the National Institutes of Health under award numbers NIH RO1GM83098 and RO1GM107374 to GC, and the National Science Foundation under award numbers NSF # 1148897 and # 1402725 to GB.

## References

Alcaide et al.(2014). Miguel Alcaide, Elizabeth S. C. Scordato, Trevor D. Price, and Darren E. Irwin. Genomic divergence in a ring species complex. *Nature*, 511 (7507), July 2014.

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

Atzmon et al.(2010). Gil Atzmon, Li Hao, Itsik Pe'er, Christopher Velez, Alexander Pearlman, Pier Francesco Palamara, Bernice Morrow, Eitan Friedman, Carole Oddoux, Edward Burns, and Harry Ostrer. Abraham's children in the genome era: Major jewish diaspora populations comprise distinct genetic clusters with shared middle eastern ancestry. *The American Journal of Human Genetics*, 86(6): 850 – 859, 2010.

Barton et al.(2013). N.H. Barton, A.M. Etheridge, J. Kelleher, and A. Véber. Inference in two dimensions: Allele frequencies versus lengths of shared sequence blocks. *Theoretical Population Biology*, 87(0):105 – 119, 2013. Coalescent Theory.

Barton et al.(2002). Nick H. Barton, Frantz Depaulis, and Alison M. Etheridge. Neutral evolution in spatially continuous populations. *Theoretical Population Biology*, 61(1):31–48, February 2002.

Beall et al.(2010). Cynthia M. Beall, Gianpiero L. Cavalleri, Libin Deng, Robert C. Elston, Yang Gao, Jo Knight, Chaohua Li, Jiang Chuan Li, Yu Liang, Mark McCormack, Hugh E. Montgomery, Hao Pan, Peter A. Robbins, Kevin V. Shianna, Siu Cheung Tam, Ngodrop Tsering, Krishna R. Veeramah, Wei Wang, Puchung Wangdui, Michael E. Weale, Yaomin Xu, Zhe Xu, Ling Yang, M. Justin Zaman, Changqing Zeng, Li Zhang, Xianglong Zhang, Pingcuo Zhaxi, and Yong Tang Zheng. Natural selection on EPAS1 (HIF2?) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences*, 107(25):11459–11464, 2010.

Bhaskar et al.(2014). Anand Bhaskar, Y.X. Rachel Wang, and Yun S. Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *bioRxiv*, 2014.

Bigham et al.(2010). Abigail Bigham, Marc Bauchet, Dalila Pinto, Xianyun Mao, Joshua M. Akey, Rui Mei, Stephen W. Scherer, Colleen G. Julian, Megan J. Wilson, David López Herráez, Tom Brutsaert, Esteban J. Parra, Lorna G. Moore, and Mark D. Shriver. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet*, 6(9):e1001116, 09 2010.

Botigué et al.(2013). Laura R. Botigué, Brenna M. Henn, Simon Gravel, Brian K. Maples, Christopher R. Gignoux, Erik Corona, Gil Atzmon, Edward Burns, Harry Ostrer, Carlos Flores, Jaume Bertranpetit, David Comas, and Carlos D. Bustamante. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proceedings of the National Academy of Sciences of the United States of America*, 110(29):11791–11796, July 2013.

Bradburd et al.(2013). Gideon S. Bradburd, Peter L. Ralph, and Graham M. Coop. Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, 67(11):3258–3273, 2013.

Cavalli-Sforza and Edwards (1967). L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis: Models and estimation procedures. *Evolution*, 21(3):pp. 550–570, 1967.

Cavalli-Sforza and Piazza (1975). L. L. Cavalli-Sforza and A. Piazza. Analysis of evolution: Evolutionary rates, independence and treeness. *Theoretical Population Biology*, 8(2):127 – 165, 1975.

Chakraborty and Weiss (1988). R Chakraborty and K M Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences of the United States of America*, 85(23):9119–9123, December 1988.

Coyne (2004). H. Allen Orr Jerry A. Coyne. *Speciation*. Sinauer Associates, Sunderland, Mass, 2004.

De and Durrett (2007). Arkendra De and Richard Durrett. Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. *Genetics*, 176(2):969–981, 2007.

Diggle et al.(1998). P J Diggle, J A Tawn, and R A Moyeed. Model-based geostatistics. *Jounal of the Royal Statistical Society. Series C (Applied Statistics)*, 47(3):299–350, 1998.

Excoffier et al.(2013). Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sanchez, Vitor C. Sousa, and Matthieu Foll. Robust demographic inference from genomic and SNP data. *PLoS Genet*, 9(10):e1003905, 10 2013.

Felsenstein (1975). Joseph Felsenstein. A pain in the torus: Some difficulties with models of isolation by distance. *The American Naturalist*, 109(967):359–368, 1975.

Felsenstein (1982). Joseph Felsenstein. How can we infer geography and history from gene frequencies? *Journal of Theoretical Biology*, 96(1):9 – 20, 1982.

Francois et al.(2010). Olivier Francois, Mathias Currat, Nicolas Ray, Eunjung Han, Laurent Excoffier, and John Novembre. Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution*, 27(6):1257–1268, 2010.

Frichot et al.(2012). Eric Frichot, Sean D Schoville, Guillaume Bouchard, and Olivier Francois. Correcting principal component maps for effects of spatial autocorrelation in population genetic data. *Frontiers in Genetics*, 3(254), 2012.

Gravel (2012). Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, June 2012.

Gutenkunst et al.(2009). Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*, 5(10):e1000695, 10 2009.

Harpending and Rogers (2000). H. Harpending and A. Rogers. Genetic perspectives on human origins and differentiation. *Annu Rev Genomics Hum Genet*, 1: 361–385, 2000.

Hellenthal et al.(2014). G. Hellenthal, G. B. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, Feb 2014.

Henn et al.(2012). Brenna M. Henn, Laura R. Botigué, Simon Gravel, Wei Wang, Abra Brisbin, Jake K. Byrnes, Karima Fadhlaoui-Zid, Pierre A. Zalloua, Andres Moreno-Estrada, Jaume Bertranpetit, Carlos D. Bustamante, and David Comas. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet*, 8(1):e1002397, January 2012.

Hodgson et al.(2014). Jason A. Hodgson, Connie J. Mulligan, Ali Al-Meeri, and Ryan L. Raaum. Early back-to-Africa migration into the Horn of Africa. *PLoS Genet*, 10(6):e1004393, June 2014.

Hudson (2002). Richard R Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

Irwin et al.(2008). D. E. Irwin, M. P. Thimgan, and J. H. Irwin. Call divergence is correlated with geographic and genetic distance in greenish warblers (*Phylloscopus trochiloides*): a strong role for stochasticity in signal evolution? *Journal of Evolutionary Biology*, 21(2):435–448, 2008.

Irwin et al.(2001). Darren E. Irwin, Staffan Bensch, and Trevor D. Price. Speciation in a ring. *Nature*, 409(6818):333–337, January 2001.

Irwin et al.(2005). Darren E. Irwin, Bensch Staffan, H Irwin Jessica, and D Price, Trevor. Speciation by distance in a ring species. *Science*, 307(5708):414–416, January 2005.

Lawson et al.(2012). Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453, 01 2012.

Lazaridis et al.(2014). Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, Mark Lipson, Bonnie Berger, Christos Economou, Ruth Bollongino, Qiaomei Fu, Kirsten I. Bos, Susanne Nordenfelt, Heng Li, Cesare de Filippo, Kay Prüfer, Susanna Sawyer, Cosimo Posth, Wolfgang Haak, Fredrik Hallgren, Elin Fornander, Nadin Rohland, Dominique Delsate, Michael Francken, Jean-Michel Guinet, Joachim Wahl, George Ayodo, Hamza A. Babiker, Graciela Bailliet, Elena Balanovska, Oleg Balanovsky, Ramiro Barrantes, Gabriel Bedoya, Haim Ben-Ami, Judit Bene, Fouad Berrada, Claudio M. Bravi, Francesca Brisighelli, George B. J. Busby, Francesco Cali, Mikhail Churnosov, David E. C. Cole, Daniel Corach, Larissa Damba, George van Driem, Stanislav Dryomov, Jean-Michel Dugoujon, Sardana A. Fedorova, Irene Gallego Romero, Marina Gubina, Michael Hammer, Brenna M. Henn, Tor Hervig, Ugur Hodoglugil, Aashish R. Jha, Sena Karachanak-Yankova, Rita Khusainova, Elza Khusnutdinova, Rick Kittles, Toomas Kivisild, William Klitz, Vaidutis Kučinskas, Alena Kushniarevich, Leila Laredj, Sergey Litvinov, Theologos Loukidis, Robert W. Mahley, Béla Melegh, Ene Metspalu, Julio Molina, Joanna Mountain, Klemetti Näkkäläjärvi, Desislava Nesheva, Thomas Nyambo, Ludmila Osipova, Jüri Parik, Fedor Platonov, Olga Posukh, Valentino Romano, Francisco

Rothhammer, Igor Rudan, Ruslan Ruizbakiev, Hovhannes Sahakyan, Antti Sajantila, Antonio Salas, Elena B. Starikovskaya, Ayele Tarekegn, Draga Toncheva, Shahlo Turdikulova, Ingrida Uktveryte, Olga Utevska, René Vasquez, Mercedes Villena, Mikhail Voevoda, Cheryl A. Winkler, Levon Yepiskoposyan, Pierre Zalloua, Tatijana Zemunik, Alan Cooper, Cristian Capelli, Mark G. Thomas, Andres Ruiz-Linares, Sarah A. Tishkoff, Lalji Singh, Kumarasamy Thangaraj, Richard Villems, David Comas, Rem Sukernik, Mait Metspalu, Matthias Meyer, Evan E. Eichler, Joachim Burger, Montgomery Slatkin, Svante Pääbo, Janet Kelso, David Reich, and Johannes Krause. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518): 409–413, September 2014.

Li and Durbin (2011). Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, July 2011. doi: 10.1038/nature10231.

Li et al.(2008). 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 (New York, N.Y.)*, 319(5866):1100–1104, February 2008.

Lipson et al.(2013). Mark Lipson, Po-Ru Loh, Alex Levin, David Reich, Nick Patterson, and Bonnie Berger. Efficient moment-based inference of admixture parameters and sources of gene flow. *Molecular Biology and Evolution*, 30(8): 1788–1802, August 2013.

Loh et al.(2013). P. R. Loh, M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich, and B. Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4):1233–1254, Apr 2013.

Luca et al.(1994). Luca, Paolo Menozzi, and Alberto Piazza. *The History and Geography of Human Genes.* Princeton University Press, Princeton, NJ, 1994.

Malécot(1975). Gustave Malécot. Heterozygosity and relationship in regularly subdivided populations. *Theoretical Population Biology*, 8(2):212–241, 1975.

Mayr (1942). Ernst Mayr. *Systematics and the origin of species, from the viewpoint of a zoologist.* Harvard University Press, October 1942.

Mayr (1970). Ernst Mayr. *Populations, species, and evolution; an abridgment of Animal species and evolution.* Belknap Press of Harvard University Press Cambridge, Mass, 1970.

McVean (2009). Gil McVean. A genealogical interpretation of principal components analysis. *PLoS Genet*, 5(10):e1000686, October 2009.

Meirmans (2012). Patrick G. Meirmans. The trouble with isolation by distance. *Molecular Ecology*, 21(12):2839–2846, 2012.

Menozzi et al.(1978). P Menozzi, A Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–792, September 1978.

Moorjani et al.(2011). Priya Moorjani, Nick Patterson, Joel N. Hirschhorn, Alon Keinan, Li Hao, Gil Atzmon, Edward Burns, Harry Ostrer, Alkes L. Price, and David Reich. The history of African gene flow into southern Europeans, Levantines, and Jews. *PLoS Genet*, 7(4):e1001373, April 2011.

Moorjani et al.(2013). Priya Moorjani, Kumarasamy Thangaraj, Nick Patterson, Mark Lipson, Po-Ru Loh, Periyasamy Govindaraj, Bonnie Berger, David Reich, and Lalji Singh. Genetic evidence for recent population mixture in India. *American Journal of Human Genetics*, 93(3):422–438, September 2013.

Nagylaki (1978). Thomas Nagylaki. A diffusion model for geographically structured populations. *Journal of Mathematical Biology*, 6(4):375–382, 1978.

Nicholson et al.(2002). George Nicholson, Albert V. Smith, Frosti Jonsson, Omar Gustafsson, Kari Stefansson, and Peter Donnelly. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal Of The Royal Statistical Society Series B*, 64(4):695–715, 2002.

Novembre and Stephens (2008). John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, May 2008.

Novembre et al.(2008). John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán 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, November 2008.

Patterson et al.(2006). Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 12 2006.

Patterson et al.(2012). Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, November 2012.

Paul et al.(2011). Joshua S. Paul, Matthias Steinrucken, and Yun S. Song. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics*, 2011.

Petkova et al.(2014). Desislava Petkova, John Novembre, and Matthew Stephens. Visualizing spatial population structure with estimated effective migration surfaces. *bioRxiv*, 2014.

Pickrell and Pritchard (2012). J. K. Pickrell and J. K. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.*, 8(11):e1002967, 2012.

Pickrell and Reich (2014). J. K. Pickrell and D. Reich. Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet.*, 30(9): 377–389, Sep 2014.

Pickrell et al.(2012). J. K. Pickrell, N. Patterson, C. Barbieri, F. Berthold, L. Gerlach, T. Guldemann, B. Kure, S. W. Mpoloka, H. Nakagawa, C. Naumann, M. Lipson, P. R. Loh, J. Lachance, J. Mountain, C. D. Bustamante, B. Berger, S. A. Tishkoff, B. M. Henn, M. Stoneking, D. Reich, and B. Pakendorf. The genetic prehistory of southern Africa. *Nat Commun*, 3:1143, 2012.

Pickrell et al.(2014). J. K. Pickrell, N. Patterson, P. R. Loh, M. Lipson, B. Berger, M. Stoneking, B. Pakendorf, and D. Reich. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. U.S.A.*, 111(7):2632–2637, Feb 2014.

Price et al.(2006). 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. *Nat Genet*, 38(8):904–909, August 2006.

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

Prugnolle et al.(2005). Franck Prugnolle, Andrea Manica, and François Balloux. Geography predicts neutral genetic diversity of human populations. *Current biology : CB*, 15(5):R159–R160, March 2005.

Ralph and Coop (2013). Peter Ralph and Graham Coop. The geography of recent genetic ancestry across Europe. *PLoS Biol*, 11(5):e1001555, 05 2013.

Ramachandran et al.(2005). S. Ramachandran, O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U.S.A.*, 102(44): 15942–15947, Nov 2005.

Reich et al.(2009). David Reich, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price, and Lalji Singh. Reconstructing Indian population history. *Nature*, 461 (7263):489–494, September 2009.

Roberts et al.(1997). G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7:110–120, 1997.

Roberts and Rosenthal (2001). Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4): 351–367, 2001.

Roberts and Rosenthal (2009). Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

Rosenberg et al.(2002). 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 (New York, N.Y.)*, 298(5602):2381–2385, December 2002.

Rosenthal (2011). Jeffrey S. Rosenthal. Optimal proposal distributions and adaptive MCMC. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, chapter 4. Chapman & Hall, CRC, Florida, USA, first edition, 2011.

Schlebusch et al.(2012). Carina M. Schlebusch, Pontus Skoglund, Per Sjödin, Lucie M. Gattepaille, Dena Hernandez, Flora Jay, Sen Li, Michael De Jongh, Andrew Singleton, Michael G. B. Blum, Himla Soodyall, and Mattias Jakobsson. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science (New York, N.Y.)*, 338(6105):374–379, October 2012.

Skoglund et al.(2012). Pontus Skoglund, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M. Thomas P. Gilbert, Anders Götherström, and Mattias Jakobsson. Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. *Science*, 336(6080):466–469, 2012.

Skoglund et al.(2014). Pontus Skoglund, Per Sjödin, Tobias Skoglund, Martin Lascoux, and Mattias Jakobsson. Investigating population history using temporal genetic differentiation. *Molecular Biology and Evolution*, 31(9):2516–2527, September 2014.

Ticehurst (1938). C.B. Ticehurst. *A Systematic Review of the Genus Phylloscopus*. Trustees of the British Museum, London, 1938.

Wake and Schneider (1998). David B. Wake and Christopher J. Schneider. Taxonomy of the plethodontid salamander genus ensatina. *Herpetologica*, 54(2): pp. 279–298, 1998.

Wang et al.(2012). Chaolong Wang, Sebastian Zöllner, and Noah A. Rosenberg. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet*, 8(8):e1002886, 08 2012.

Wasser et al.(2004). Samuel K Wasser, Andrew M Shedlock, Kenine Comstock, Elaine Ostrander, Benezeth Mutayoba, and Matthew Stephens. Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *PNAS*, 101(41):14847–52, October 2004.

Wright (1943). Sewall Wright. Isolation by distance. *Genetics*, 28(2):114–138, 1943.

Xu and Jin (2008). Shuhua Xu and Li Jin. A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *The American Journal of Human Genetics*, 83(3):322–336, September 2008.

Yang et al.(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, May 2012.

Yang et al.(2014). Wen-Yun Yang, Alexander Platt, Charleston Wen-Kai Chiang, Eleazar Eskin, John Novembre, and Bogdan Pasaniuc. Spatial localization of recent ancestors for admixed individuals. *G3: Genes|Genomes|Genetics*, 4(12): 2505–2518, December 2014.