

RESEARCH ARTICLE

A Spatial Framework for Understanding Population Structure and Admixture

Gideon S. Bradburd^{1*}, Peter L. Ralph², Graham M. Coop¹

1 Center for Population Biology, Department of Evolution and Ecology, University of California, Davis, California, United States of America, **2** Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California, United States of America

* gbradbud@ucdavis.edu



 OPEN ACCESS

Citation: Bradburd GS, Ralph PL, Coop GM (2016) A Spatial Framework for Understanding Population Structure and Admixture. *PLoS Genet* 12(1): e1005703. doi:10.1371/journal.pgen.1005703

Editor: Montgomery Slatkin, University of California, Berkeley, UNITED STATES

Received: February 10, 2015

Accepted: November 5, 2015

Published: January 15, 2016

Copyright: © 2016 Bradburd et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the following grants: National Science Foundation (nsf.gov) Grant Number 1262645 to PLR and GMC; National Science Foundation (nsf.gov) Grant Numbers 1148897 and 1402725 to GB; National Institute of Health (<http://nih.gov/>) Grant numbers RO1GM83098 and RO1GM107374 to GC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

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 [1–3], or approximations to the coalescent applied to sequence data [4–6].

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 [7–9] and reduced dimensionality representations of the data (e.g. [10–12]).

A third set of methods that describe relatedness between populations by constructing a “population phylogeny” was pioneered by Cavalli-Sforza and Edwards [13], as were methods to test whether a tree is a good model of population history [14] (see [15] 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 [16]. Others, such as TreeMix [17] and MixMapper [18], 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 [11], especially Principal Components Analysis (PCA; also pioneered by Cavalli-Sforza [19]). 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 [20]. This allows PCA to describe more geographically continuous relationships: applied to human populations within continents, it often shows a close correspondence to geographic locations [21, 22]. 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 [23–25].

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 [26–29], as well as exciting new developments in the field [30]. The pattern of increasing genetic differentiation with geographic distance was termed “Isolation by Distance” by Wright [31], and is ubiquitous in natural populations [32]. 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 ℓ in population k as $C_{k,\ell}$, and the total number of alleles observed as $S_{k,\ell}$. The sample frequency at locus ℓ 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 ℓ 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 ℓ in population k , and \bar{f}_ℓ is the average of the K sample allele frequencies, weighted by mean population size. This normalization is widely used [11, 33]; 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 if the amount of drift since a common ancestor is small.

We work with the empirical covariance matrix of these standardized sample allele frequencies, calculated across loci, namely, $\hat{\Omega} = (1/L)\hat{X}\hat{X}^T$. 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}_ℓ were instead an unobserved, global mean allele frequency at locus ℓ .

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 landscape. 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 [34–36]: the covariance between standardized population allele frequencies (i.e. \hat{X} values) between populations i and j is assumed to be, for $i \neq j$,

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

where D_{ij} is the geogenetic distance between populations i and j , α_0 controls the within-population variance (or the covariance when distance between points is 0, known as a “sill” in the geospatial literature), α_1 controls the rate of the decay of covariance per unit pairwise distance, and α_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

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

where $\delta_{ij} = 1$ if $i = j$ and is 0 otherwise, η_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 Ω given in Eq (3). We denote this by

$$P(\hat{\Omega} | \Omega) = \mathcal{W}(L\hat{\Omega} | \Omega, L). \tag{4}$$

Note that if the standardized sample frequencies are Gaussian, then the sample covariance matrix is a sufficient statistic, so that calculating the likelihood of $\hat{\Omega}$ is the same as calculating the probability 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. This modeling approach rests on the assumption that the loci in the dataset are independent, that is, not in linkage disequilibrium (LD). Linkage disequilibrium between loci included in the dataset will have the effect of decreasing the true number of degrees of freedom, effectively making this likelihood calculation

a composite likelihood and artificially increasing confidence in parameter estimation. We discuss possible ways to accommodate linkage disequilibrium further in the Discussion.

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 η determine the parametric covariance matrix Ω (given by Eq (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 η) 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

$$P(G, \vec{\alpha}, \eta \mid \hat{\Omega}, L) \propto P(\hat{\Omega} \mid \Omega(\vec{\alpha}, D(G), \eta))P(\vec{\alpha})P(G)P(\eta), \quad (5)$$

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.

Simulations

We first apply the method to several scenarios simulated using the coalescent simulator `ms` [37]. Each scenario is simulated using a stepping stone model in which populations are arranged on a grid with symmetric migration to nearest neighbors (eight neighbors, 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 generated using the parameters that have 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. The axes of the resultant maps are presented as Northings and Eastings, as population locations in this geogenetic space no longer conform to the latitude or longitude of the original sampling locations. In S1 Fig, we show the

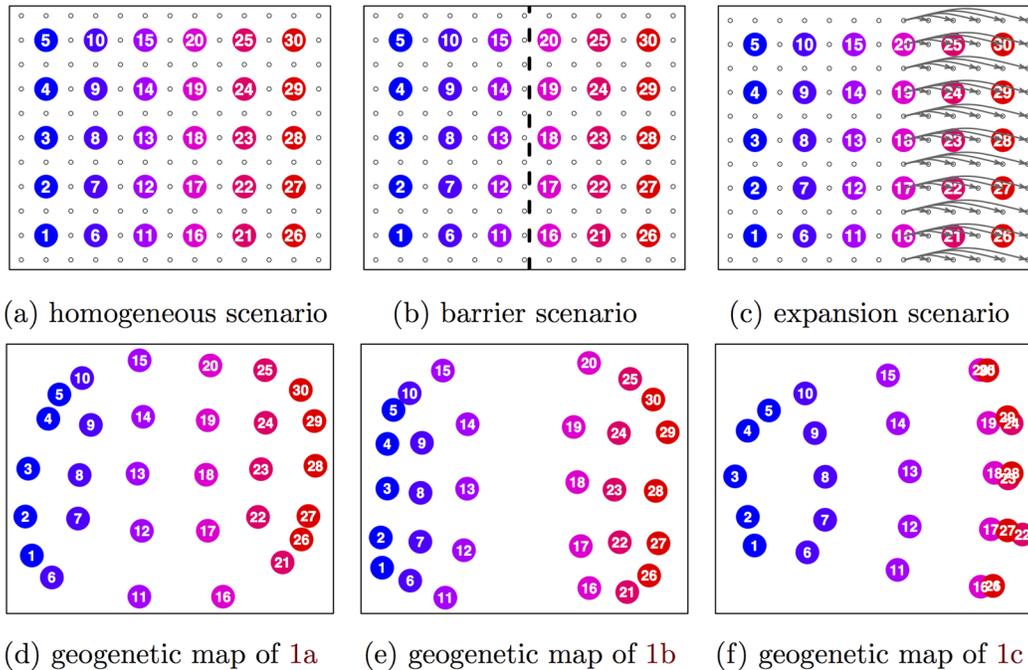


Fig 1. Simulation scenarios and their corresponding geogenetic maps estimated with SpaceMix. The smaller circles in the simulation scenarios represent unsampled populations. **a)** the configuration of simulated populations on a simple lattice with spatially homogeneous migration rates (a plot showing the first two Principal Component axes for this simulation is given in a); **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.

doi:10.1371/journal.pgen.1005703.g001

relationship between genetic covariance, geographic distance, and inferred geogenetic distance for these simulations.

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 (Fig 1a and 1d), SpaceMix recovers the lattice structure used to simulate the data (i.e., populations correctly find 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, and because, without nearest-neighbor migration from farther outside the lattice, they are in fact more closely related to their neighbors.

We also examined the effects of uneven sampling on inference by subsampling a 9×9 grid into a variety of subsets that had successfully greater ‘uneven-ness’ of sampling, and comparing PCA and SpaceMix on these unevenly subsampled datasets. The results of these analyses are shown in S3–S9 Figs. As sampling becomes more uneven, the maps produced by plotting

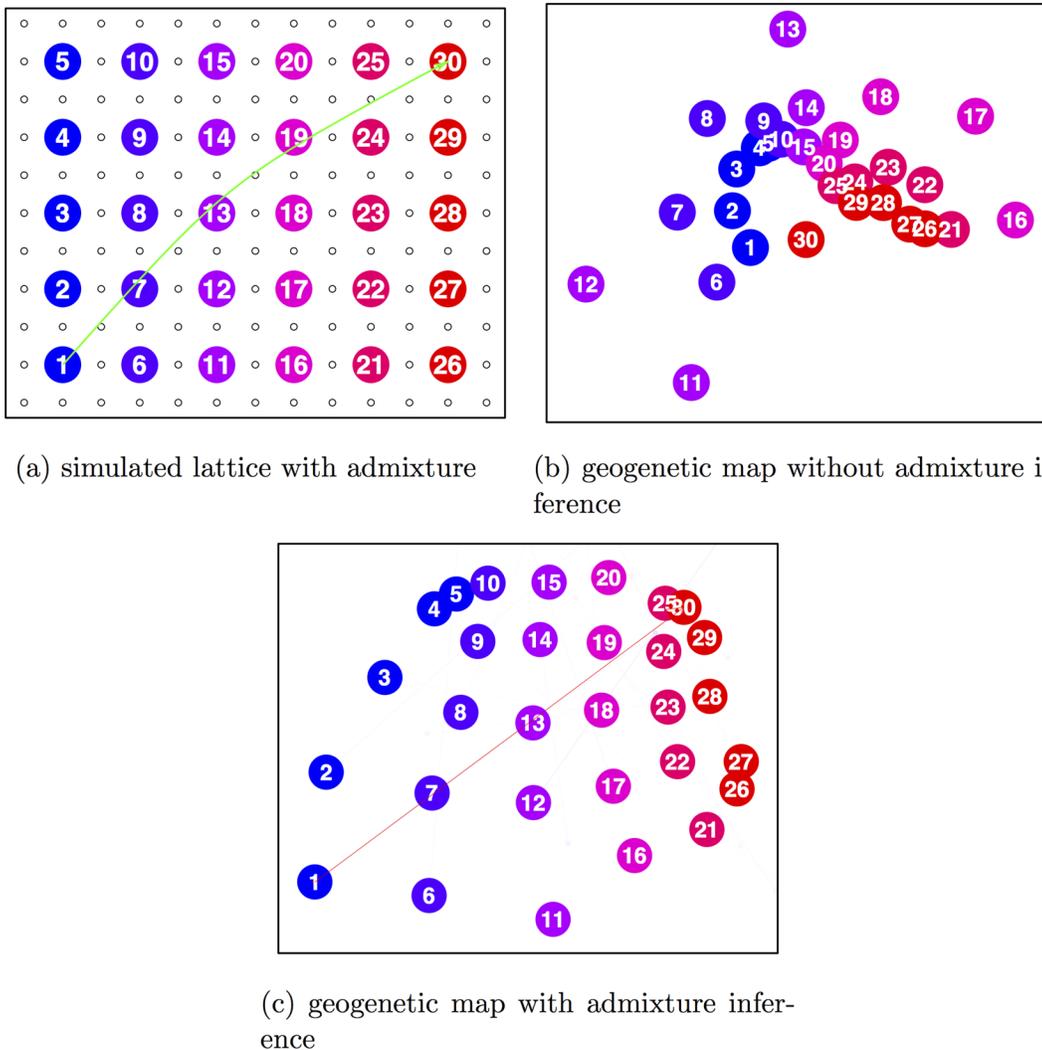


Fig 2. Simulation scenarios and SpaceMix inference. **a)** a lattice with a recent admixture event between population 1 in the southwest corner and population 30 in the northeast corner, so that population 30 is drawing half of its ancestry from population 1 (a plot showing the first two Principal Component axes for this simulation is given in [S2 Fig](#) panel b); **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.

doi:10.1371/journal.pgen.1005703.g002

Principal Component axis 1 (PC1) against PC2 diverge more and more from the true geographic configuration of the samples ([S10 Fig](#)), quickly resembling the triangular shape commonly seen in PCA plots of real datasets. SpaceMix also becomes less certain about placement of some samples, but to a much smaller extent, and in all scenarios SpaceMix produces geogenetic maps that are more faithful to the true geographic configuration of the samples than those generated using PCA.

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.

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 Eq (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 $\hat{X}_{i,\ell}$ and $\hat{X}_{j,\ell}$, for each ℓ , as

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

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 Eq (2), and we reintroduce the nugget, η_k , and the sample size effect, $1/\bar{S}_k$, for each population as above in Eq (3).

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 Eq (4), except with Ω replaced by Ω^* . The posterior probability of these parameters can be expressed as a function of this parametric admixed covariance, Ω^* ,

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

as specified by the parameters w , G^* , $\vec{\alpha}$, and η , 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 the Methods.

The models described above may be used in various combinations. In the simplest model, locations are not estimated for populations, nor do they draw admixture; the only parameters to be estimated are those of the spatial covariance function given in Eq (2), and the population-specific variance terms (η_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. We discuss the utility of these different models in the Methods.

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).

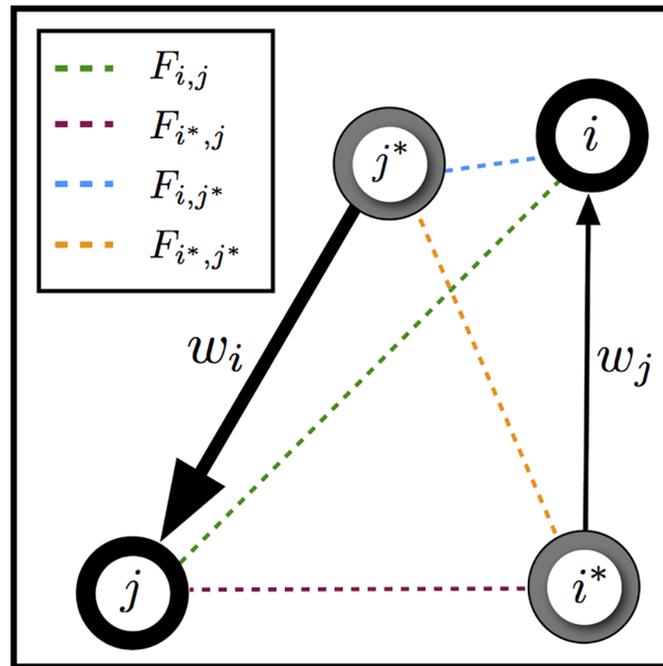


Fig 3. An illustration of the form of the admixed covariance. Following Eq (6), populations i and j are drawing admixture in proportions w_i and w_j from their respective sources of admixture, i^* and j^* , and all pairwise spatial covariances (the F 's) are shown. In this cartoon example, population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

doi:10.1371/journal.pgen.1005703.g003

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) draws 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 (Fig 4d); however, its intermediate genetic relationships are explained through an estimated location close to its true admixture source (population 13) and source of admixture (95% CI: 0.04–0.14) 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, the model is able to explain population 18's higher covariance with population 13 via its estimated 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.

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, S1 and S2 Tables. For all analyses presented below, we centered the priors on location

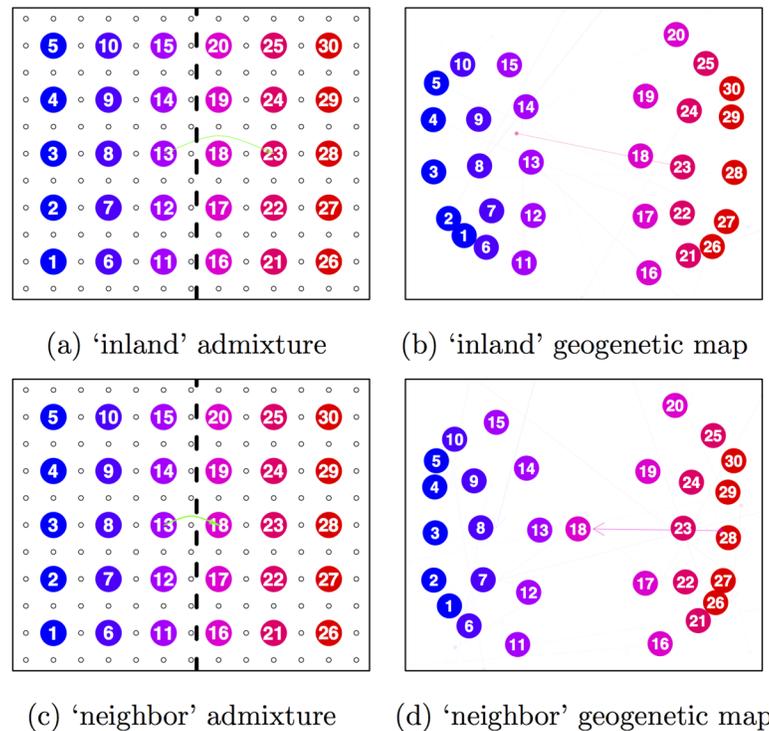


Fig 4. Simulation scenarios and inferred population maps for two different admixture scenarios. Green arrows denote admixture from a source to a target population in the simulation. **a)** lattice with a barrier and an admixture event (10%) across the barrier to an 'inland' population; **b)** the inferred 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).

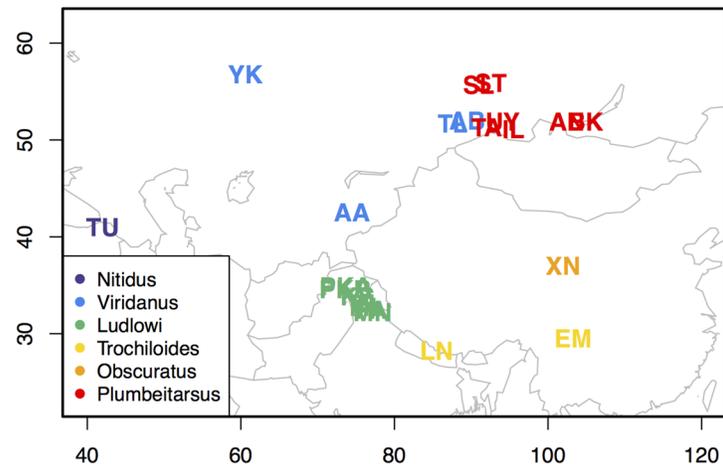
doi:10.1371/journal.pgen.1005703.g004

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). As with the simulations described above, the axes of the geogenetic maps are presented in Eastings and Northings.

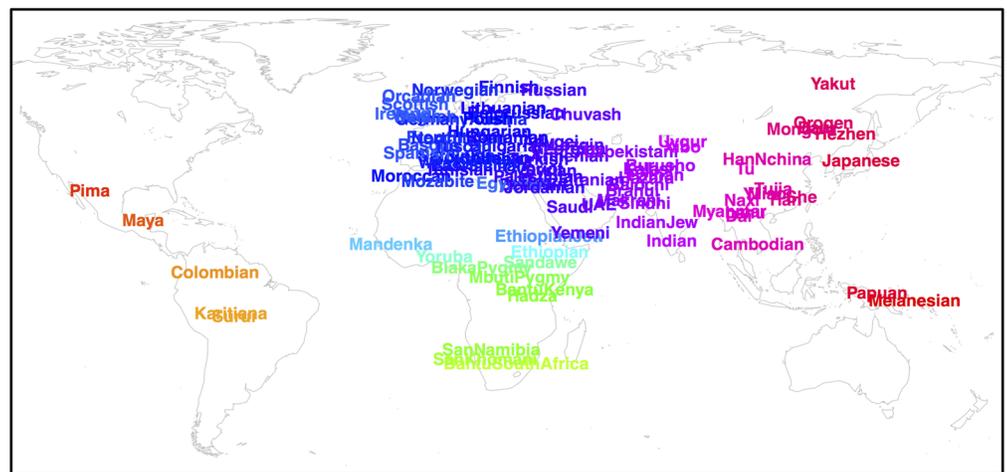
Greenish Warblers

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 [38–40]. 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 [39, 41]. 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



(a) Warbler subspecies distribution map



(b) Human sample distribution map

Fig 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 [46]; (b) sampling map for human dataset, consisting of 1,490 individuals from 95 population samples [50].

doi:10.1371/journal.pgen.1005703.g005

facilitated periods of differentiation in genotype or phenotype without gene flow [42–44] (see Wake and Schneider [45] for discussion). Alcaide *et. al* [46] 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* [46], 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. These loci were treated as

independent (i.e., un-linked). We discuss ways to accommodate linkage disequilibrium further in the Discussion.

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 has a relatively high nugget parameter (see S11 Fig panel a), 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 S12 Fig).

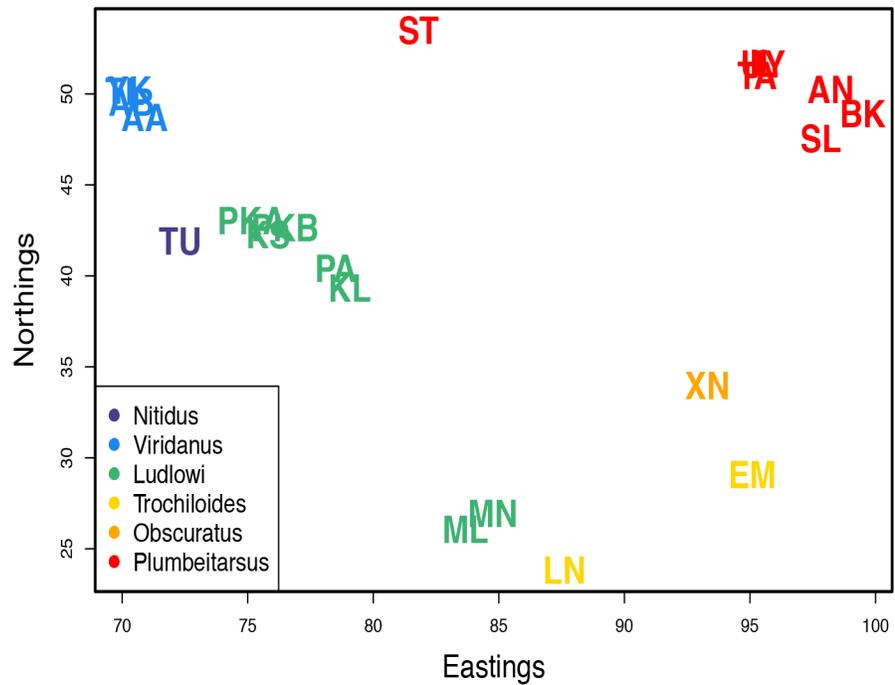
We then ran the method allowing admixture (Fig 6b). The only population sample with appreciable admixture is the Stolby sample (ST; $w = 0.19$, 95% credible interval: 0.146–0.238; S13 Fig). This sample is known to be composed of an equal mixture of eastern *plumbeitarsus* and western *viridanus* individuals [46]. Multiple runs agreed well on the level of admixture of the Stolby sample (see S14 Fig). What does vary across runs is whether the Stolby sample has an estimated location by the *viridanus* cluster while drawing admixture from near the *plumbeitarsus* cluster, or vice versa; however, this is to be expected given the 50/50 nature of the sample's makeup (S14 Fig). The somewhat intermediate position of the Stolby sample, and its non-50/50 admixture proportion, likely partially reflect the influence of the priors (S15 Fig).

We repeated these analyses (with and without admixture) with a dataset in which we treated each individual as the unit of analysis (Fig 7). No individual drew appreciable admixture (see S16 Fig for admixture proportions), and so we discuss the results with admixture (those without admixture are nearly identical, see S17–S19 Figs). As with the analysis on multi-sample populations, the results approximately mirror the geography of the individuals.

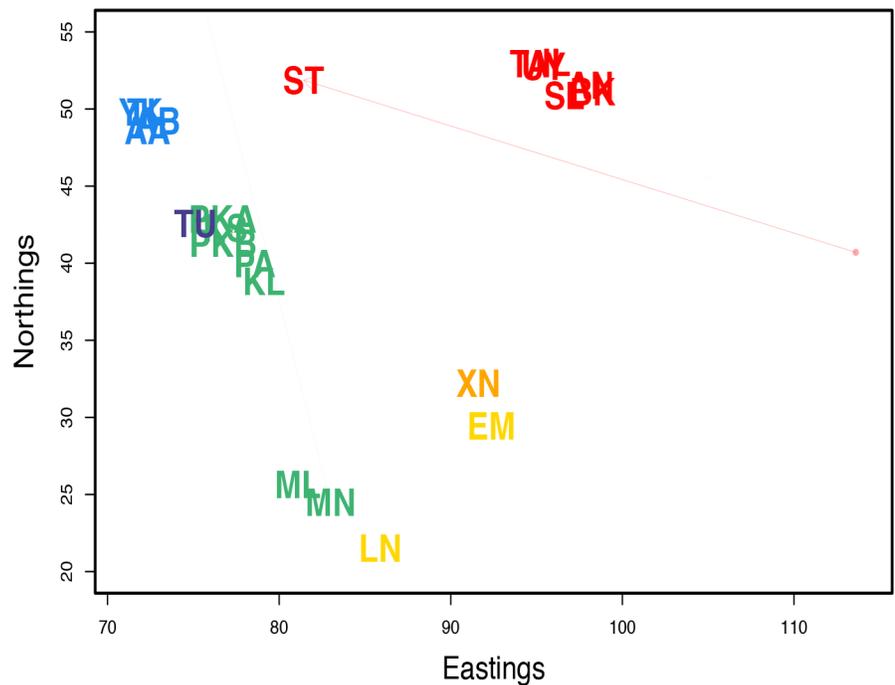
There are, however, a number of obvious departures in the individual geogenetic map from the population map. The most obvious is that the location of a pair of *nitidus* samples (in purple) is very far from the rest of the samples. These individuals appear to be closely related, and in the population-level analysis, this increase in shared ancestry was accounted for by a large nugget for the *nitidus* population (S11 Fig panel a). However, in the individual-level analysis, a nugget is estimated separately for each sample, so, the model must accommodate the much higher relatedness between this pair of individuals through estimated locations that are 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 S20 Fig).

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 the estimated locations of individuals from the Stolby population are near their respective clusters. Although the geogenetic separation between the *viridanus* and *plumbeitarsus* individuals is greater than their geographic separation, 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 configuration, with *viridanus* and *plumbeitarsus* at its tips, is steady within and among runs of the MCMC and choice of position priors (see S18 Fig).

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 S21 Fig and [46]). However, as discussed by Novembre and Stephens [23], such patterns in PC analysis



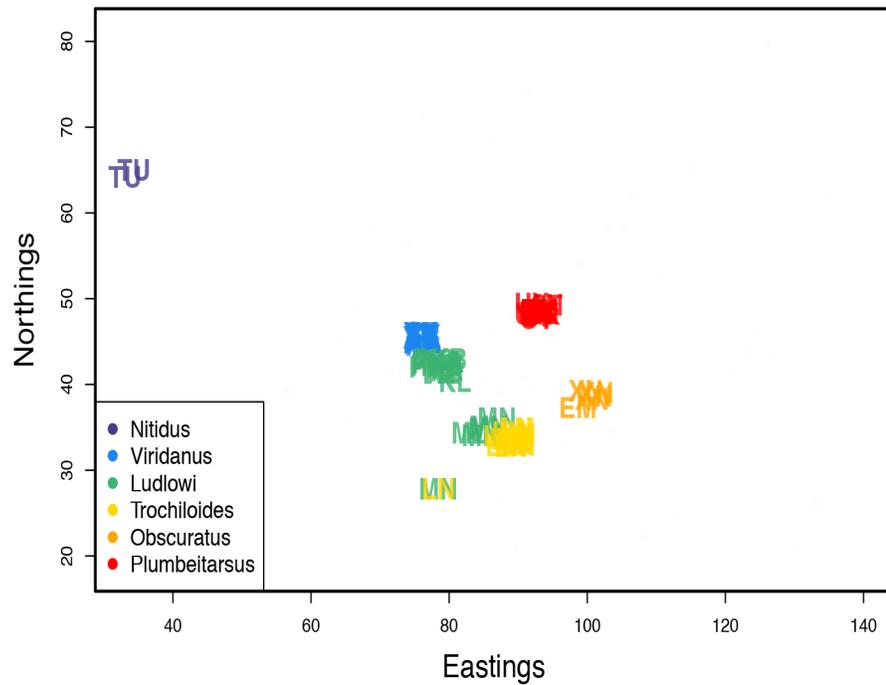
(a) Warbler geogenetic map, no admixture



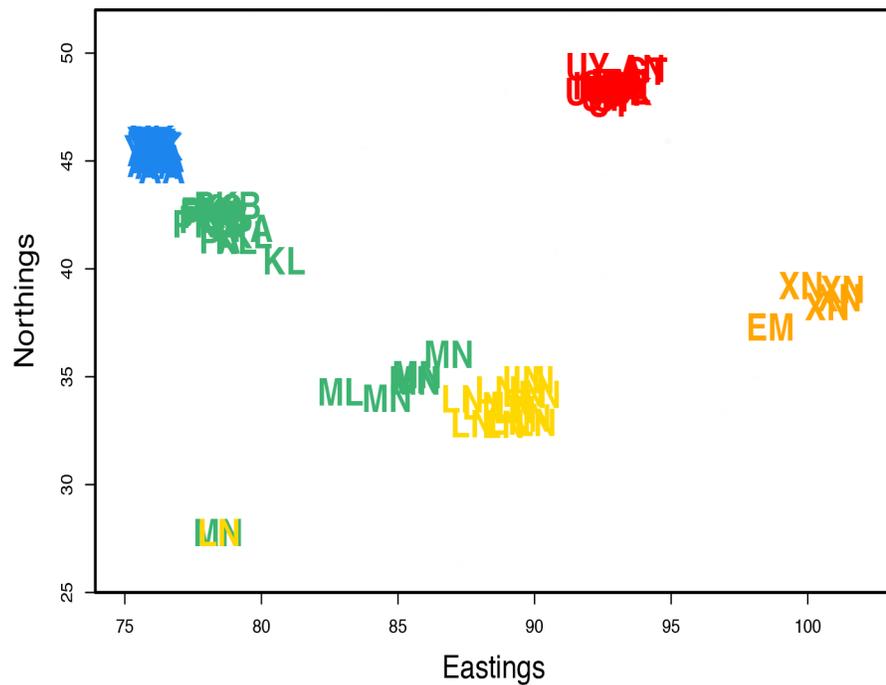
(b) Warbler geogenetic map, with admixture

Fig 6. Inferred maps for warbler populations. Population labels are colored as in Fig 5a. **a)** the map inferred with no admixture inference; **b)** the map inferred with admixture inference.

doi:10.1371/journal.pgen.1005703.g006



(a) Warbler individuals map, admixture



(b) Close-up of non-*nitidus* samples

Fig 7. Inferred maps for warbler individuals with admixture inference. Individual labels are colored by subspecies as in Fig 5a. **a)** map inferred with admixture; **b)** close-up of all non-*nitidus* samples in the admixture map.

doi:10.1371/journal.pgen.1005703.g007

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 [S22 Fig](#) panel b) not because of gene flow connecting the tips, but rather because of the orthogonality requirement of PCs (see [\[23\]](#) 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 [S22 Fig](#) panels c and d). 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.* [\[46\]](#), 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.* [\[46\]](#), makes the *ludlowi* cluster unusual compared to the estimated spatial distributions of the other subspecies (see [S23 Fig](#)), and suggests a break in historic or current gene flow.

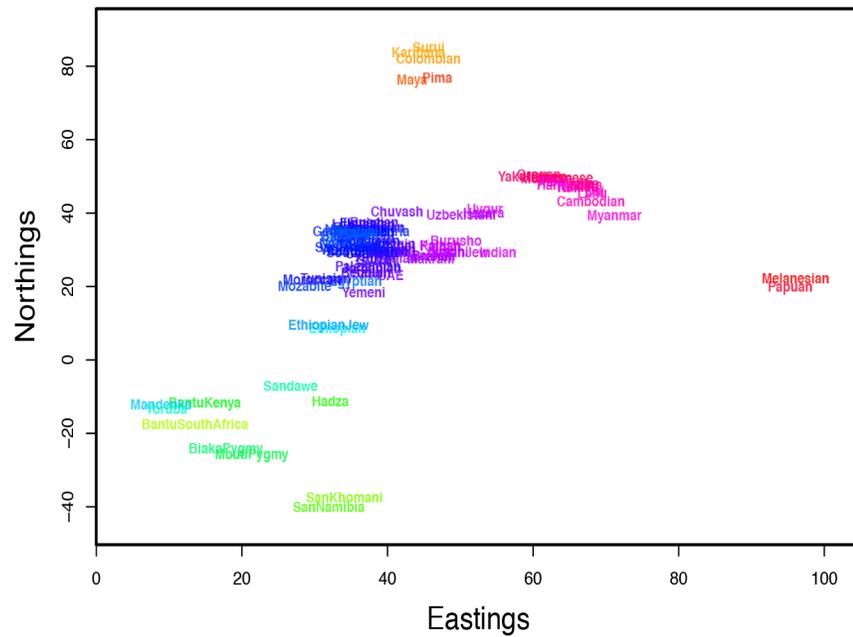
Human Populations

Human population structure is a complex product of the forces of migration and drift acting on both local and global scales, patterned by geography [\[21, 47\]](#), time [\[48, 49\]](#), admixture [\[50\]](#), landscape and environment [\[36, 51, 52\]](#), and shaped by culture [\[16, 53, 54\]](#). 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 of our results, and view them as a simplistic but rich visualization of patterns of population structure.

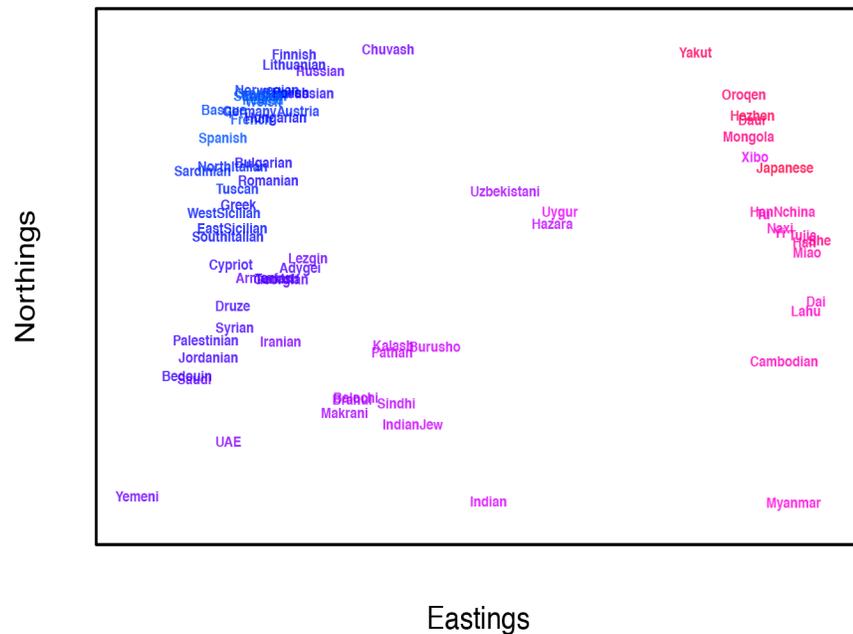
We used the dataset of Hellenthal *et al.* [\[50\]](#), comprised of 1,490 individuals from 95 population samples (see [Fig 5b](#) for map of sampling), as well as the latitude and longitude attributed to each sample. In the analyses presented on human genotype data below, we have thinned the total dataset for LD in windows of 50 base-pairs, with a step-size of 5 base-pairs, and an upper limit of 0.2 on pairwise r^2 [\[55, 56\]](#). We then used a random subset of 10,000 SNPs to estimate the sample covariance.

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 these samples that are in the HGDP dataset [\[50, 57–60\]](#). 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



(a) Inferred map of human samples



(b) Close-up of Eurasian samples

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

doi:10.1371/journal.pgen.1005703.g008

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 S24 Fig), 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 [61–63] following an out-of-Africa expansion, and subsequent expansions into Western Eurasia, East Asia, the Americas, and Oceania (although see Pickrell and Reich [64] for a discussion of other models).

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 shifts the estimated locations of admixed samples intermediate between Sub-Saharan Africa and North Africa/the Middle East toward one cluster or the other, which, in turn, pushes each of those major clusters to move relatively farther apart. The Ethiopian and Ethiopian Jewish samples have estimated locations closer to the Sub-Saharan samples than those of the North African samples, but draw substantial amounts of admixture (~40%) from close to where the Egyptian sample has positioned itself in the the Middle East cluster, as do the Sandawe [65, 66]. The SanKhomani draw admixture from near Syria, which may reflect multiple distinct geographic sources of admixture [50, 67]. Interestingly the Bantu South African sample, though it has an estimated location near 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 [50, 66–68]. The inferred sample-specific drift parameters (the nuggets) are similar between runs with and without admixture (S25 Fig).

The majority of North African samples (Egyptian, Tunisian, Moroccan, 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 [50, 69].

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 [50], 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 [54, 70].

The other significant signal of admixture is between East and West Eurasia, a signal documented by many authors [50, 57, 58, 71]. The majority of samples maintain their relative positions within each of these groups; however, there are several samples that show admixture between eastern and western Eurasia. The Uzbekistani and Hazara samples have estimated locations close to the East Asian samples and draw a substantial admixture proportion from

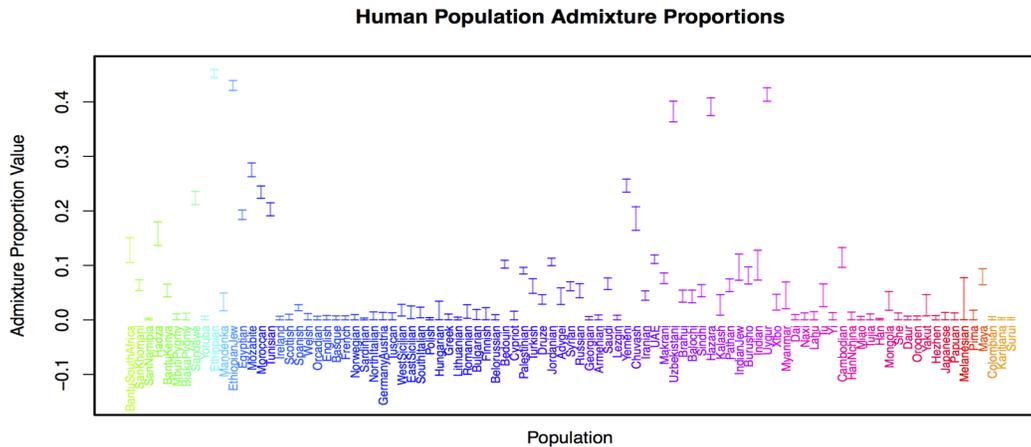


Fig 10. Admixture proportions (95% CIs) for each human population sample. Labels are colored as in as in Fig 5b.

doi:10.1371/journal.pgen.1005703.g010

close to the Georgian and Armenian estimated locations. Conversely, the Uyghur sample has an estimated location close to the Burusho, Kalash, and Pathan samples, and draws admixture from near the Mongola and Hezhen samples. The Tu sample (with a geogenetic location in East Asia) draws a small amount of ancestry from close to the estimated location of the Uyghur. The estimated location of the Chuvash sample is near the Russian and Lithuanian samples, and the Chuvash draw admixture from close to the Yakut (as do the Turkish, to a smaller extent). 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 Egyptian sample [17, 50].

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 have estimated admixture locations that average over those sources. This may be the case for the Hadza and Bantu Keynan samples [50]. A second possibility is that the relatively steep prior on admixture proportion forces samples to draw 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 the limited geographic sampling of current day populations, or because of old admixture events from populations for which there are not other more direct modern descendant 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.* [16] and Moorjani *et al.* [72] 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 S26 and S27 Figs, 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 the east-west axis within the Eurasian cluster. For some samples that draw a significant amount of admixture, such as the central Asian populations (Uyгур, Hazara and Uzbekistani), the estimated location switches 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 have estimated 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 uneven sampling, 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 [17], and MixMapper [18]), and producing powerful visualizations of genetic structure on a map (like PCA [11] and SPA [73]).

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 model-based clustering-based methods [7] or TreeMix/MixMapper-like methods [17, 18]. 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.

Comparison to PCA

SpaceMix, like PCA, is well suited to describing population structure in a continuous fashion—but it also has a number of advantages over PCA. PCA is a general-purpose tool for exploratory visualization of high-dimensional data; in application to genetic data, PCA can quickly identify problematic samples and major axes of variation. Since geography is a major cause of differentiation, the first one or two PC axes often correspond to geography [23]. However, because PCs are linear functions of the genotypes, sometimes many PCs must be used to depict patterns produced by simple isolation by distance [23].

These higher order PCs can be hard to interpret in empirical data (see discussion in the warbler section). The recently introduced SPA approach [74], which also assumes allele frequencies are monotonically increasing in a given direction, may suffer from the same problem, which SpaceMix avoids (although PCA and SPA are both significantly faster than SpaceMix). Similarly, unevenness of sampling can greatly distort PC maps, as illustrated in the comparison of the uneven subsampling simulation scenarios shown in S8 and S9 Figs.

The generality of PCA is also its weakness: it displays any structure, not necessarily geographical structure. SpaceMix, since it works explicitly with local correlations on maps, is designed to visualize the relationships between samples induced by geographically limited dispersal, and so is less easily misled by other types of structure. Our explicit modeling of admixture is also helpful; in PCA, admixed individuals appear in intermediate locations in PC biplots, but are not distinguished from individuals in intermediate populations.

The application of SpaceMix to humans illustrates the utility of our approach: the first two PCs of this dataset resemble a triangle (S28 Fig), 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). SpaceMix's explicitly geographic model avoids the tendency of PC biplots towards triangular plots, as was also seen when applied to unevenly sampled datasets (S3–S9 Figs).

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, interpretation and visualization of these higher dimensions is more difficult, and so we have stuck to 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.

Our method shows the utility of representing both isolation by distance and long-distance admixture on a 2-D geogenetic map. While we generate this map using likelihood-based inference relating a parametric covariance matrix to the observed empirical matrix, it would be interesting to explore other methods of creating this geogenetic map (e.g., [30, 74–76]). Such methods may offer computational speedups and also potentially help place SpaceMix within a broader statistical framework.

Admixture Arrows

One of the greatest strengths of SpaceMix is the introduction of admixture arrows. Although PCA can be interpreted in light of simple admixture events [20], and new methods can locate the recent, spatially admixed ancestry of out-of-sample individuals [73, 74], neither approach explicitly models admixture between multiple geographically distant locations, as SpaceMix does. Assignment methods are designed to deal with many admixed samples [7], 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 in a SpaceMix analysis 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 have clustered estimated locations, 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 [50] using a different approach), despite evidence of substantial ancient admixture [77]. 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.

Linkage Between Loci

In this paper, we have treated the loci in the dataset as independent, and, where necessary, we have thinned empirical datasets to decrease LD between loci. One possible approach that avoids the necessity of thinning the data would be to calculate the sample covariance in large (e.g., megabase), non-overlapping windows along the genome, then average those sample covariances across all windows. Another approach is to use empirical LD between loci to estimate the effective number of independent loci in the dataset, and use this quantity as the number of degrees of freedom in the Wishart likelihood calculation. Additionally, although we have focused on the covariance among alleles at the same locus, linkage disequilibrium (covariance of alleles among loci) holds rich information about the timing and source of admixture events [50, 72, 78, 79] as well as information about isolation by distance [47]. Just as population graph approaches have been extended to incorporate information from LD [59], a spatial covariance approach could be informed by LD. A null model inspired by models of LD under isolation by distance models [80, 81] 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.

Future Work

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 [64] 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 [49]. 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 Eq (7) of the population locations, G , their sources of admixture, G^* , their admixture proportions, w , their independent drift parameters, η , 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 ϵ_ℓ , 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 ϵ_ℓ), 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}_ℓ in place of ϵ_ℓ , which is what we do to define $\hat{X}_{k,\ell}$ in Eq (1). Recall that the sample allele frequency at locus ℓ 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 ℓ 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

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

where $\delta_{ij} = 1$ if $i = j$ and is 0 otherwise. If the covariance matrix of Y is Ω^* , then the covariance matrix of \hat{X}_ℓ 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}_ℓ 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 reduce the covariance among populations in \hat{X}_ℓ compared to \hat{f}_ℓ , 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 Ψ 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 $\hat{\Omega} = \hat{X}^T \hat{X}$ as

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

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) population locations are estimated, but not admixture; (3) populations may draw admixture, but their own locations are fixed; or (4) population locations and admixture are both estimated. The free parameters for each of options are given in [Table 1](#).

Although we anticipate most empirical researchers will be interested in the joint inference of a geogenetic map with admixture (Model 4), we have presented these models separately, as we believe each have their own utility. Model 1 and Model 3 can each be used to infer landscapes of allele frequencies, upon which genotyped individuals can be probabilistically placed (following [35]). This application may be useful to determine the geographic origin of potentially contraband biological samples (e.g., ivory), or the most likely source of museum specimens missing sampling metadata. Model 3 has the potential to improve the performance of these spatial assignment methods over Model 1, as the inclusion of admixture in the model may allow for more accurate inference of allele frequency surfaces. Model 2 directly parallels

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

| Model | # of Free Parameters | Parameters |
|---|----------------------|---|
| (1) stationary population locations, no admixture | $K + 3$ | $\alpha_0, \alpha_1, \alpha_2, \eta$ |
| (2) inferred population locations, no admixture | $2K + 3$ | $\alpha_0, \alpha_1, \alpha_2, \eta, G$ |
| (3) stationary population locations, inferred admixture | $2K + 3$ | $\alpha_0, \alpha_1, \alpha_2, \eta, G^*, w$ |
| (4) inferred population locations, inferred admixture | $3K + 3$ | $\alpha_0, \alpha_1, \alpha_2, \eta, G, G^*, w$ |

doi:10.1371/journal.pgen.1005703.t001

Principal Component Analysis. Informally the visual comparison of Models 2 and 4 can allow investigators to understand how ignoring long distance admixture distorts relationships among populations. Formally, the fit of these various models could be compared by cross validation, but we do not implement that here.

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 ℓ in population k ; 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 ℓ 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^{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 $\hat{\Omega}$ 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 $(\hat{\Omega}, G^{(obs)}, 1/\bar{S}_k)$, 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

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 $G^{(obs)}$.

| Parameter | Description | Prior |
|------------|--|---|
| α_0 | controls the sill of the covariance matrix | $\alpha_0 \sim \text{Exp}(\lambda = 1/100)$ |
| α_1 | controls the rate of the decay of covariance with distance | $\alpha_1 \sim \text{Exp}(\lambda = 1)$ |
| α_2 | controls the shape of the decay of covariance with distance | $\alpha_2 \sim U(0.1, 2)$ |
| η_k | the nugget in population k (population specific drift parameter) | $\eta_k \sim \text{Exp}(\lambda = 1)$ |
| G_k | the geogenetic location of population k | $G_k \sim \mathcal{N}(\mu = G_k^{(obs)}, \sigma = \frac{1}{2}\bar{D}(G^{(obs)}))$ |
| w_k | the proportion of admixture in population k | $2w_k \sim \beta(\alpha = 1, \beta = 100)$ |
| G_k^* | the geogenetic location of the source of admixture in population k | $G_k^* \sim \mathcal{N}(\mu = \bar{G}^{(obs)}, \sigma = 2\bar{D}(G^{(obs)}))$ |

doi:10.1371/journal.pgen.1005703.t002

parameter value (e.g. η_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 Θ' , which differs from Θ 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 η parameters, gives the admixed covariance matrix described in Eq (6), Ω^* . The likelihood of the two sets of parameters Θ and Θ' , calculated with Eq (12) and the priors of Table 2, combine to give the Metropolis-Hastings ratio, R , the probability of accepting the proposed parameter values Θ' :

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

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

Updates for $\vec{\alpha}$, η , and w . We propose updates to the values of the $\vec{\alpha}$, η , 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 $\delta \sim \mathcal{N}(0, \sigma_{\alpha_0}^2)$ and σ_{α_0} is the tuning parameter for α_0 . For η 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 proposed move is rejected, in which case the current parameter value is sampled in that iteration and reused in the next 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 Δ_{G_i} , where, e.g., $\Delta_{G_i} \sim |\mathcal{N}(0, \sigma_{G_i})|$, and a bearing, γ , where $\gamma \sim U(0, 2\pi)$. Then we use the following equations to calculate the latitude and longitude of the proposed location:

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

and

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

where latitude and longitude are given in radians and are taken mod 2π . As with η 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 [82, 83]. 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 [84, 85]. We implement this mechanism by creating, for each variable i , an associated variable ζ_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 ζ .

When we start our MCMC, ζ_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 ζ_i by an “adaption amount”, which is added to ζ_i if the acceptance proportion in the n th set of 50 iterations ($r_i^{(n)}$) has been above 0.44, and subtracted from ζ_i if not. The magnitude of the adaption amount is a decreasing function of the index n , so that updates to ζ_i proceed as follows:

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

We choose to cap the maximum adaption amount at 20 (which is the equivalent of capping the variance of the proposal distribution at 4.85×10^8) 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 S29 Fig, 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` [37]. The full command line arguments for all simulations are included in the S1 Text. Briefly, we simulated populations on a lattice, with nearest neighbor 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 (S22 Fig), 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.

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 having an estimated location near that population in geogenetic space to satisfy its increased resemblance to that population, rather than drawing 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 Eq (6), with no nugget or sampling effect.

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 if B had an estimated location near 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 an estimated location

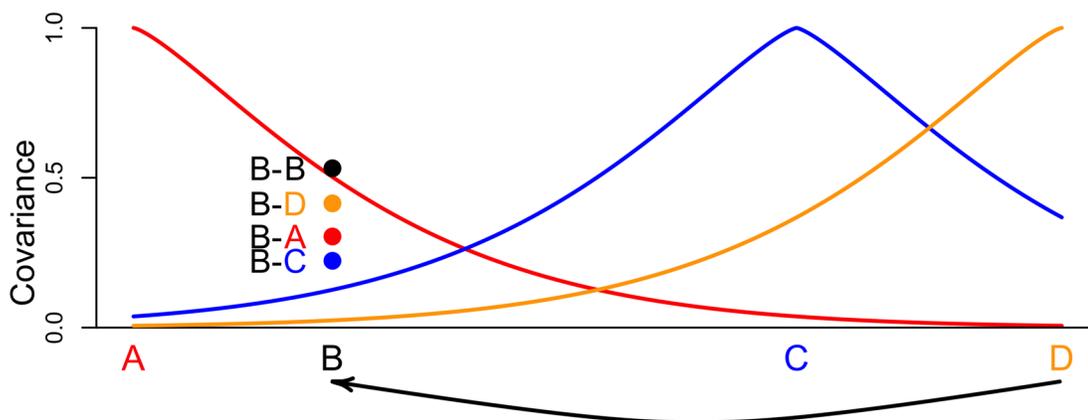


Fig 11. Illustrated example of spatial covariance and the effects of admixture. Lines show the covariances populations A, C, and D would have with population B as a function of B’s location with no admixture, under the parametric form of Eq (2). The colored dots above ‘B’ show the covariances observed with B at that location given that B has 40% admixture from D. There is no spatial configuration that induces unadmixed covariances remotely similar to those observed.

doi:10.1371/journal.pgen.1005703.g011

too close 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 [Fig 2c](#), [Fig 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 model is able to explain the high covariance the admixed populations have with their sources of admixture via their inferred location, rather than just via their inferred source of admixture and admixture proportion. The reason the model explains these admixed populations' 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 5×10^6 MCMC iterations each in which population locations were estimated (but no admixture). Population locations were initiated at the origin (i.e. at iteration 1 of the MCMC, $G_i = (0,0)$), or at uniformly distributed coordinates between the minimum and maximum observed range of latitude and longitude, 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 [[11](#)]. 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., [21]). For consistency of presentation, we show the full Procrustes superimposition of the PC coordinate space around the geographic sampling locations.

Treemix Comparisons

To contrast our approach to tree-based approaches to admixture we applied TreeMix [17] to our spatial simulations. We took the `ms` output on which we had also run Spacemix (Figs 1, 2 and 4), and converted it into TreeMix format. We also included an outgroup sequence for each dataset, which consisted of a single haploid individual who carried the 0 (ancestral allele) at every locus. We ran TreeMix without migration edges on the processed `ms.treemix.file.gz` file to construct the initial tree, which was rooted using the `myoutgroup` sequence using the following command:

```
treemix -i ms.treemix.file.gz -root myoutgroup -o treemix_output
```

We then sequentially added admixture migration edges, using the following command to add another edge to the existing tree (“`prev.treemix`”):

```
treemix -i ms.treemix.file.gz -root myoutgroup -g prev.treemix.vertices.gz prev.treemix.edges.gz -m 1 -o treemix_output
```

The TreeMix graphs and residual covariance matrices were visualized using the scripts provided with TreeMix.

In S30 Fig we show the tree and admixture graphs produced when TreeMix is run on a lattice stepping stone model (a smaller scale version of this exercise was previously done by Pickrell and Pritchard [17]). The tree produced by running TreeMix is rake-like, showing the lack of deep shared sub-division. However, while the tree captures some features of isolation by distance (e.g., neighboring samples are often sister to each other), the tree structure forces many unnatural splits of geographically neighboring populations (as was previously found [17]; see their S14 Fig). The admixture migration arrows act to mitigate the strongest departures from the tree, such as geographically neighboring samples that were forced into separate places on the tree, but are unable to fully accommodate the spatial relationships between samples. Different runs of TreeMix on the same dataset result in quite different trees and orders of migration events, reflecting both the high degree of symmetry in our simulated samples on a grid, and also the poor fit of the tree model to spatial data. In S31 and S32 Figs, we also present the results of TreeMix run on our expansion and barrier simulations.

In S33, S34 and S35 Figs, we show the application of TreeMix to the scenarios simulated with admixture events. For none of our scenarios was the true admixture the first migration edge added; in fact, only for the corner admixture scenario was the true admixture event in the first three edges added. This reflects the fact that TreeMix has to add migration edges to cope with the residual covariance induced by the poor fit of a tree to spatially simulated data, and so misses more subtle (but real) admixture events. The poor performance of TreeMix here is the result of the spatially simulated data not conforming to the underlying assumption of the TreeMix tree-like model.

Supporting Information

S1 Fig. 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. (TIF)

S2 Fig. Plots of the first two Principal Component axes (with variance explained labeled on the relevant axes) of the mean-centered covariance matrix from simulated spatial scenarios. a) the basic lattice scenario shown in [Fig 1a](#). b) the lattice scenario with admixture from Population 1 into Population 30, shown in [Fig 2a](#). (TIF)

S3 Fig. The effect of uneven sampling on inference of geogenetic maps. This figure shows the complete grid with corresponding PC map and SpaceMix geogenetic map (ellipses denote 95% credible intervals after a 50% burn-in). This grid is successively subsampled with greater unevenness of sampling across six subsampled scenarios. (TIF)

S4 Fig. The effect of uneven sampling on inference of geogenetic maps. This figure shows the subsampling Scenario 1, in which the lower left, top left, and top right squares are ‘protected,’ and 2 samples are randomly removed from all other squares. Also shown are the corresponding PC map and SpaceMix geogenetic map (ellipses denote 95% credible intervals after a 50% burn-in). (TIF)

S5 Fig. The effect of uneven sampling on inference of geogenetic maps. This figure shows the subsampling Scenario 2, in which the lower left, top left, and top right squares are ‘protected,’ and 4 samples are randomly removed from all other squares. Also shown are the corresponding PC map and SpaceMix geogenetic map (ellipses denote 95% credible intervals after a 50% burn-in). (TIF)

S6 Fig. The effect of uneven sampling on inference of geogenetic maps. The effect of uneven sampling on inference of geogenetic maps. This figure shows the subsampling Scenario 3, in which the lower left, top left, and top right squares are ‘protected,’ and 6 samples are randomly removed from all other squares. Also shown are the corresponding PC map and SpaceMix geogenetic map (ellipses denote 95% credible intervals after a 50% burn-in). (TIF)

S7 Fig. The effect of uneven sampling on inference of geogenetic maps. The effect of uneven sampling on inference of geogenetic maps. This figure shows the subsampling Scenario 4, in which the lower left, top left, and top right squares are ‘protected,’ and 8 samples are randomly removed from all other squares. Also shown are the corresponding PC map and SpaceMix geogenetic map (ellipses denote 95% credible intervals after a 50% burn-in). (TIF)

S8 Fig. The effect of uneven sampling on inference of geogenetic maps. The effect of uneven sampling on inference of geogenetic maps. This figure shows the subsampling Scenario 5, in which the lower left, top left, and top right squares are ‘protected,’ and all 9 samples are removed from all other squares. Also shown are the corresponding PC map and SpaceMix geogenetic map (ellipses denote 95% credible intervals after a 50% burn-in). (TIF)

S9 Fig. The effect of uneven sampling on inference of geogenetic maps. The effect of uneven sampling on inference of geogenetic maps. This figure shows the subsampling Scenario 6, in which the lower left, middle, and bottom right squares are ‘protected,’ and all 9 samples are removed from all other squares, except for the bottom left and top right squares, which are each randomly subsampled down to one sample. Also shown are the corresponding PC map

and SpaceMix geogenetic map (ellipses denote 95% credible intervals after a 50% burn-in). (TIF)

S10 Fig. The effect of uneven sampling on inference of geogenetic maps. The effect of uneven sampling on inference of geogenetic maps. This figure shows, for each scenario, the median distance between the true sampling coordinates and the sample coordinates in PC-space and SpaceMix's geogenetic space. For the SpaceMix analyses, the geogenetic coordinates used were chosen from the MCMC iteration with the highest posterior probability. (TIF)

S11 Fig. Credible intervals on estimated warbler population nugget parameters. a) analysis without admixture b) analysis with admixture. (TIF)

S12 Fig. 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. (TIF)

S13 Fig. Warbler population admixture proportions. Credible intervals on estimated warbler population admixture proportion parameters. (TIF)

S14 Fig. 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. (TIF)

S15 Fig. 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. (TIF)

S16 Fig. Warbler individual admixture proportions. Credible intervals on estimated warbler individual admixture proportion parameters. (TIF)

S17 Fig. 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. (TIF)

S18 Fig. Maps of the posterior distributions on population locations in three separate SpaceMix analyses on the warbler individual dataset. a) analysis with randomly generated priors on geogenetic location parameters; b) one analysis with true geographic locations as priors on geogenetic location parameters; c) a second analysis with true geographic locations as priors on geogenetic location parameters. (TIF)

S19 Fig. Warbler individual nugget parameters. Credible intervals on estimated warbler individual nugget parameters. **a)** analysis without admixture; **b)** analysis with admixture.

(TIF)

S20 Fig. Warbler individual pairwise- π . 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.

(TIF)

S21 Fig. PCA map of warbler individuals. The map of warbler individuals derived from a Principal Components analysis, plotting PC1 against PC2. The PC coordinates have undergone a full Procrustes transformation around the actual sampling coordinates.

(TIF)

S22 Fig. 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.

(TIF)

S23 Fig. Comparison of geographic and geogenetic pairwise distances for warbler individuals. **a)** comparisons between populations in different subspecies. **b)** comparisons between populations in the same subspecies.

(TIF)

S24 Fig. Comparison of geographic and geogenetic pairwise distances for human samples individuals. Comparisons are 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.

(TIF)

S25 Fig. Credible intervals on estimated human sample nugget parameters. **a)** analysis without admixture; **a)** analysis with admixture.

(TIF)

S26 Fig. Geogenetic map of human samples. Map of human populations from a different SpaceMix analysis than that reported in the main text ("Real_Prior1"—inferred with admixture), using real geographic coordinates as population location priors. **a)** complete map; **b)** close-up of Eurasian populations.

(TIF)

S27 Fig. Geogenetic map 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.

(TIF)

S28 Fig. 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).
(TIF)

S29 Fig. Example parameter acceptance proportions. Example parameter acceptance proportions for the α_2 parameter and the nugget parameter, η , using the adaptive Metropolis-within-Gibbs proposal mechanism.

(TIF)

S30 Fig. TreeMix analysis of lattice spatial coalescent simulation. The tree, residual covariance matrix, and first three migration admixture arrows are shown.

(TIF)

S31 Fig. TreeMix analysis of lattice spatial coalescent simulation with a barrier. The tree, residual covariance matrix, and first three migration admixture arrows are shown.

(TIF)

S32 Fig. TreeMix analysis of lattice spatial coalescent simulation with an expansion event. The tree, residual covariance matrix, and first three migration admixture arrows are shown.

(TIF)

S33 Fig. TreeMix analysis of lattice spatial coalescent simulation with a long-range admixture event. The tree, residual covariance matrix, and first three migration admixture arrows are shown.

(TIF)

S34 Fig. TreeMix analysis of lattice spatial coalescent simulation with a barrier and short-range admixture event. The tree, residual covariance matrix, and first three migration admixture arrows are shown.

(TIF)

S35 Fig. TreeMix analysis of lattice spatial coalescent simulation with a barrier and mid-range admixture event. The tree, residual covariance matrix, and first three migration admixture arrows are shown.

(TIF)

S1 Table. Warbler sample metadata. Subspecies and geographic meta-data for greenish warbler individuals included in analysis.

(TIF)

S2 Table. Human sample metadata. Sample size and geographic meta-data for human samples included in analysis.

(TIF)

S1 Text. Command line arguments to simulate lattice datasets. Command line arguments used in `ms` to simulate lattice datasets for SpaceMix analysis.

(TXT)

Acknowledgments

We would like to gratefully acknowledge the Coop Lab, Yaniv Brandvain, Razib Khan, Joe Pickrell, Jonathan Pritchard, 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.

Author Contributions

Conceived and designed the experiments: GSB PLR GMC. Performed the experiments: GSB. Analyzed the data: GSB. Contributed reagents/materials/analysis tools: GSB PLR GMC. Wrote the paper: GSB PLR GMC.

References

1. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet.* 2009 10; 5(10): e1000695. doi: [10.1371/journal.pgen.1000695](https://doi.org/10.1371/journal.pgen.1000695) PMID: [19851460](https://pubmed.ncbi.nlm.nih.gov/19851460/)
2. Bhaskar A, Wang YXR, Song YS. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *bioRxiv.* 2014;.
3. Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet.* 2013 10; 9(10):e1003905. doi: [10.1371/journal.pgen.1003905](https://doi.org/10.1371/journal.pgen.1003905) PMID: [24204310](https://pubmed.ncbi.nlm.nih.gov/24204310/)
4. Paul JS, Steinrücken M, Song YS. An Accurate Sequentially Markov Conditional Sampling Distribution for the Coalescent With Recombination. *Genetics.* 2011;. doi: [10.1534/genetics.110.125534](https://doi.org/10.1534/genetics.110.125534)
5. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature.* 2011 July; 475(7357):493–496. doi: [10.1038/nature10231](https://doi.org/10.1038/nature10231) PMID: [21753753](https://pubmed.ncbi.nlm.nih.gov/21753753/)
6. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet.* 2014 Aug; 46(8):919–925. doi: [10.1038/ng.3015](https://doi.org/10.1038/ng.3015) PMID: [24952747](https://pubmed.ncbi.nlm.nih.gov/24952747/)
7. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics.* 2000 Jun; 155(2):945–959. PMID: [10835412](https://pubmed.ncbi.nlm.nih.gov/10835412/)
8. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research.* 2009; 19(9):1655–1664. doi: [10.1101/gr.094052.109](https://doi.org/10.1101/gr.094052.109) PMID: [19648217](https://pubmed.ncbi.nlm.nih.gov/19648217/)
9. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet.* 2012 01; 8(1):e1002453. doi: [10.1371/journal.pgen.1002453](https://doi.org/10.1371/journal.pgen.1002453) PMID: [22291602](https://pubmed.ncbi.nlm.nih.gov/22291602/)
10. Cavalli-Sforza LL, Menozzi P, Piazza A. *The History and Geography of Human Genes.* Princeton, NJ: Princeton University Press; 1994.
11. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. *PLoS Genet.* 2006 12; 2(12): e190. doi: [10.1371/journal.pgen.0020190](https://doi.org/10.1371/journal.pgen.0020190) PMID: [17194218](https://pubmed.ncbi.nlm.nih.gov/17194218/)
12. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006 Aug; 38(8):904–909. doi: [10.1038/ng1847](https://doi.org/10.1038/ng1847) PMID: [16862161](https://pubmed.ncbi.nlm.nih.gov/16862161/)
13. Cavalli-Sforza LL, Edwards AWF. Phylogenetic Analysis: Models and Estimation Procedures. *Evolution.* 1967; 21(3):pp. 550–570. doi: [10.2307/2406616](https://doi.org/10.2307/2406616)
14. Cavalli-Sforza LL, Piazza A. Analysis of evolution: Evolutionary rates, independence and treeness. *Theoretical Population Biology.* 1975; 8(2):127–165. doi: [10.1016/0040-5809\(75\)90029-5](https://doi.org/10.1016/0040-5809(75)90029-5) PMID: [1198349](https://pubmed.ncbi.nlm.nih.gov/1198349/)
15. Felsenstein J. How can we infer geography and history from gene frequencies? *Journal of Theoretical Biology.* 1982; 96(1):9–20. doi: [10.1016/0022-5193\(82\)90152-7](https://doi.org/10.1016/0022-5193(82)90152-7) PMID: [7109659](https://pubmed.ncbi.nlm.nih.gov/7109659/)
16. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. *Nature.* 2009 Sep; 461(7263):489–494. doi: [10.1038/nature08365](https://doi.org/10.1038/nature08365) PMID: [19779445](https://pubmed.ncbi.nlm.nih.gov/19779445/)
17. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 2012; 8(11):e1002967. doi: [10.1371/journal.pgen.1002967](https://doi.org/10.1371/journal.pgen.1002967) PMID: [23166502](https://pubmed.ncbi.nlm.nih.gov/23166502/)
18. Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. *Molecular Biology and Evolution.* 2013 Aug; 30(8):1788–1802. doi: [10.1093/molbev/mst099](https://doi.org/10.1093/molbev/mst099) PMID: [23709261](https://pubmed.ncbi.nlm.nih.gov/23709261/)
19. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science.* 1978 Sep; 201(4358):786–792. doi: [10.1126/science.356262](https://doi.org/10.1126/science.356262) PMID: [356262](https://pubmed.ncbi.nlm.nih.gov/356262/)
20. McVean G. A Genealogical Interpretation of Principal Components Analysis. *PLoS Genet.* 2009 Oct; 5(10):e1000686. doi: [10.1371/journal.pgen.1000686](https://doi.org/10.1371/journal.pgen.1000686) PMID: [19834557](https://pubmed.ncbi.nlm.nih.gov/19834557/)
21. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. *Nature.* 2008 Nov; 456(7218):98–101. doi: [10.1038/nature07331](https://doi.org/10.1038/nature07331) PMID: [18758442](https://pubmed.ncbi.nlm.nih.gov/18758442/)

22. Wang C, Zöllner S, Rosenberg NA. A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLoS Genet.* 2012 08; 8(8):e1002886. doi: [10.1371/journal.pgen.1002886](https://doi.org/10.1371/journal.pgen.1002886) PMID: [22927824](https://pubmed.ncbi.nlm.nih.gov/22927824/)
23. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics.* 2008 May; 40(5):646–649. doi: [10.1038/ng.139](https://doi.org/10.1038/ng.139) PMID: [18425127](https://pubmed.ncbi.nlm.nih.gov/18425127/)
24. Francois O, Currat M, Ray N, Han E, Excoffier L, Novembre J. Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture. *Molecular Biology and Evolution.* 2010; 27(6):1257–1268. doi: [10.1093/molbev/msq010](https://doi.org/10.1093/molbev/msq010) PMID: [20097660](https://pubmed.ncbi.nlm.nih.gov/20097660/)
25. Frichot E, Schoville SD, Bouchard G, Francois O. Correcting principal component maps for effects of spatial autocorrelation in population genetic data. *Frontiers in Genetics.* 2012; 3(254). doi: [10.3389/fgene.2012.00254](https://doi.org/10.3389/fgene.2012.00254) PMID: [23181073](https://pubmed.ncbi.nlm.nih.gov/23181073/)
26. Malécot G. Heterozygosity and relationship in regularly subdivided populations. *Theoretical Population Biology.* 1975; 8(2):212–241. doi: [10.1016/0040-5809\(75\)90033-7](https://doi.org/10.1016/0040-5809(75)90033-7) PMID: [1198353](https://pubmed.ncbi.nlm.nih.gov/1198353/)
27. Nagylaki T. A diffusion model for geographically structured populations. *Journal of Mathematical Biology.* 1978; 6(4):375–382. doi: [10.1007/BF02463002](https://doi.org/10.1007/BF02463002)
28. Felsenstein J. A Pain in the Torus: Some Difficulties with Models of Isolation by Distance. *The American Naturalist.* 1975; 109(967):359–368. doi: [10.1086/283003](https://doi.org/10.1086/283003)
29. Barton NH, Depaulis F, Etheridge AM. Neutral Evolution in Spatially Continuous Populations. *Theoretical Population Biology.* 2002 February; 61(1):31–48. doi: [10.1006/tpbi.2001.1557](https://doi.org/10.1006/tpbi.2001.1557) PMID: [11895381](https://pubmed.ncbi.nlm.nih.gov/11895381/)
30. Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces. *bioRxiv.* 2014;
31. Wright S. Isolation by distance. *Genetics.* 1943; 28(2):114–138. PMID: [17247074](https://pubmed.ncbi.nlm.nih.gov/17247074/)
32. Meirmans PG. The trouble with isolation by distance. *Molecular Ecology.* 2012; 21(12):2839–2846. doi: [10.1111/j.1365-294X.2012.05578.x](https://doi.org/10.1111/j.1365-294X.2012.05578.x) PMID: [22574758](https://pubmed.ncbi.nlm.nih.gov/22574758/)
33. Nicholson G, Smith AV, Jónsson F, Gútafsson Ó, Stefánsson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal Of The Royal Statistical Society Series B.* 2002; 64(4):695–715. doi: [10.1111/1467-9868.00357](https://doi.org/10.1111/1467-9868.00357)
34. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. *Journal of the Royal Statistical Society Series C (Applied Statistics).* 1998; 47(3):299–350. doi: [10.1111/1467-9876.00113](https://doi.org/10.1111/1467-9876.00113)
35. Wasser SK, Shedlock AM, Comstock K, Ostrander E, Mutayoba B, Stephens M. Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *PNAS.* 2004 Oct; 101(41):14847–52. doi: [10.1073/pnas.0403170101](https://doi.org/10.1073/pnas.0403170101) PMID: [15459317](https://pubmed.ncbi.nlm.nih.gov/15459317/)
36. Bradburd GS, Ralph PL, Coop GM. Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution.* 2013; 67(11):3258–3273. doi: [10.1111/evo.12193](https://doi.org/10.1111/evo.12193) PMID: [24102455](https://pubmed.ncbi.nlm.nih.gov/24102455/)
37. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics.* 2002; 18(2):337–338. doi: [10.1093/bioinformatics/18.2.337](https://doi.org/10.1093/bioinformatics/18.2.337) PMID: [11847089](https://pubmed.ncbi.nlm.nih.gov/11847089/)
38. Ticehurst CB. *A Systematic Review of the Genus Phylloscopus.* Trustees of the British Museum, London; 1938.
39. Irwin DE, Bensch S, Price TD. Speciation in a ring. *Nature.* 2001 Jan; 409(6818):333–337. doi: [10.1038/35053059](https://doi.org/10.1038/35053059) PMID: [11201740](https://pubmed.ncbi.nlm.nih.gov/11201740/)
40. Irwin DE, Staffan B, Irwin Jessica H, Price D Trevor. Speciation by distance in a ring species. *Science.* 2005 Jan; 307(5708):414–416. doi: [10.1126/science.1105201](https://doi.org/10.1126/science.1105201) PMID: [15662011](https://pubmed.ncbi.nlm.nih.gov/15662011/)
41. Irwin DE, Thimgan MP, Irwin JH. 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.* 2008; 21(2):435–448. doi: [10.1111/j.1420-9101.2007.01499.x](https://doi.org/10.1111/j.1420-9101.2007.01499.x) PMID: [18205774](https://pubmed.ncbi.nlm.nih.gov/18205774/)
42. Mayr E. *Systematics and the origin of species, from the viewpoint of a zoologist.* Harvard University Press; 1942.
43. Mayr E. *Populations, species, and evolution; an abridgment of Animal species and evolution.* Belknap Press of Harvard University Press Cambridge, Mass; 1970.
44. Coyne HAOJA. *Speciation.* Sunderland, Mass: Sinauer Associates; 2004.
45. Wake DB, Schneider CJ. Taxonomy of the Plethodontid Salamander Genus *Ensatina*. *Herpetologica.* 1998; 54(2):pp. 279–298.
46. Alcaide M, Scordato ESC, Price TD, Irwin DE. Genomic divergence in a ring species complex. *Nature.* 2014 July; 511 (7507). doi: [10.1038/nature13285](https://doi.org/10.1038/nature13285) PMID: [24870239](https://pubmed.ncbi.nlm.nih.gov/24870239/)
47. Ralph P, Coop G. The Geography of Recent Genetic Ancestry across Europe. *PLoS Biol.* 2013 05; 11(5):e1001555. doi: [10.1371/journal.pbio.1001555](https://doi.org/10.1371/journal.pbio.1001555) PMID: [23667324](https://pubmed.ncbi.nlm.nih.gov/23667324/)

48. Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, et al. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science*. 2012; 336(6080):466–469. doi: [10.1126/science.1216304](https://doi.org/10.1126/science.1216304) PMID: [22539720](https://pubmed.ncbi.nlm.nih.gov/22539720/)
49. Skoglund P, Sjödin P, Skoglund T, Lascoux M, Jakobsson M. Investigating Population History Using Temporal Genetic Differentiation. *Molecular Biology and Evolution*. 2014 Sep; 31(9):2516–2527. doi: [10.1093/molbev/msu192](https://doi.org/10.1093/molbev/msu192) PMID: [24939468](https://pubmed.ncbi.nlm.nih.gov/24939468/)
50. Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. *Science*. 2014 Feb; 343(6172):747–751. doi: [10.1126/science.1243518](https://doi.org/10.1126/science.1243518) PMID: [24531965](https://pubmed.ncbi.nlm.nih.gov/24531965/)
51. Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, Knight J, et al. Natural selection on EPAS1 (HIF2?) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences*. 2010; 107(25):11459–11464. doi: [10.1073/pnas.1002443107](https://doi.org/10.1073/pnas.1002443107)
52. Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, et al. Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. *PLoS Genet*. 2010 09; 6(9):e1001116. doi: [10.1371/journal.pgen.1001116](https://doi.org/10.1371/journal.pgen.1001116) PMID: [20838600](https://pubmed.ncbi.nlm.nih.gov/20838600/)
53. Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, Palamara PF, et al. 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*. 2010; 86(6):850–859. doi: [10.1016/j.ajhg.2010.04.015](https://doi.org/10.1016/j.ajhg.2010.04.015) PMID: [20560205](https://pubmed.ncbi.nlm.nih.gov/20560205/)
54. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLoS Genet*. 2011 Apr; 7(4):e1001373. doi: [10.1371/journal.pgen.1001373](https://doi.org/10.1371/journal.pgen.1001373) PMID: [21533020](https://pubmed.ncbi.nlm.nih.gov/21533020/)
55. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. 2007; 81(3):559–575. doi: [10.1086/519795](https://doi.org/10.1086/519795) PMID: [17701901](https://pubmed.ncbi.nlm.nih.gov/17701901/)
56. Purcell S. PLINK v1.07; 2009. <http://pngu.mgh.harvard.edu/purcell/plink/>
57. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. *Science (New York, NY)*. 2002 Dec; 298(5602):2381–2385. doi: [10.1126/science.1078311](https://doi.org/10.1126/science.1078311)
58. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, NY)*. 2008 Feb; 319(5866):1100–1104. doi: [10.1126/science.1153717](https://doi.org/10.1126/science.1153717)
59. Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*. 2013 Apr; 193(4):1233–1254. doi: [10.1534/genetics.112.147330](https://doi.org/10.1534/genetics.112.147330) PMID: [23410830](https://pubmed.ncbi.nlm.nih.gov/23410830/)
60. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient Admixture in Human History. *Genetics*. 2012 Nov; 192(3):1065–1093. doi: [10.1534/genetics.112.145037](https://doi.org/10.1534/genetics.112.145037) PMID: [22960212](https://pubmed.ncbi.nlm.nih.gov/22960212/)
61. Harpending H, Rogers A. Genetic perspectives on human origins and differentiation. *Annu Rev Genomics Hum Genet*. 2000; 1:361–385. doi: [10.1146/annurev.genom.1.1.361](https://doi.org/10.1146/annurev.genom.1.1.361) PMID: [11701634](https://pubmed.ncbi.nlm.nih.gov/11701634/)
62. Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. *Current biology:CB*. 2005 Mar; 15(5):R159–R160. doi: [10.1016/j.cub.2005.02.038](https://doi.org/10.1016/j.cub.2005.02.038) PMID: [15753023](https://pubmed.ncbi.nlm.nih.gov/15753023/)
63. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA*. 2005 Nov; 102(44):15942–15947. doi: [10.1073/pnas.0507611102](https://doi.org/10.1073/pnas.0507611102) PMID: [16243969](https://pubmed.ncbi.nlm.nih.gov/16243969/)
64. Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet*. 2014 Sep; 30(9):377–389. doi: [10.1016/j.tig.2014.07.007](https://doi.org/10.1016/j.tig.2014.07.007) PMID: [25168683](https://pubmed.ncbi.nlm.nih.gov/25168683/)
65. Hodgson JA, Mulligan CJ, Al-Meerri A, Raaum RL. Early Back-to-Africa Migration into the Horn of Africa. *PLoS Genet*. 2014 Jun; 10(6):e1004393. doi: [10.1371/journal.pgen.1004393](https://doi.org/10.1371/journal.pgen.1004393) PMID: [24921250](https://pubmed.ncbi.nlm.nih.gov/24921250/)
66. Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Guldemann T, et al. The genetic prehistory of southern Africa. *Nat Commun*. 2012; 3:1143. doi: [10.1038/ncomms2140](https://doi.org/10.1038/ncomms2140) PMID: [23072811](https://pubmed.ncbi.nlm.nih.gov/23072811/)
67. Pickrell JK, Patterson N, Loh PR, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci USA*. 2014 Feb; 111(7):2632–2637. doi: [10.1073/pnas.1313787111](https://doi.org/10.1073/pnas.1313787111) PMID: [24550290](https://pubmed.ncbi.nlm.nih.gov/24550290/)
68. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science (New York, NY)*. 2012 Oct; 338(6105):374–379. doi: [10.1126/science.1227721](https://doi.org/10.1126/science.1227721)
69. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *PLoS Genet*. 2012 Jan; 8(1):e1002397. doi: [10.1371/journal.pgen.1002397](https://doi.org/10.1371/journal.pgen.1002397) PMID: [22253600](https://pubmed.ncbi.nlm.nih.gov/22253600/)

70. Botigué LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, et al. 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*. 2013 Jul; 110(29):11791–11796. doi: [10.1073/pnas.1306223110](https://doi.org/10.1073/pnas.1306223110) PMID: [23733930](https://pubmed.ncbi.nlm.nih.gov/23733930/)
71. Xu S, Jin L. A Genome-wide Analysis of Admixture in Uyghurs and a High-Density Admixture Map for Disease-Gene Discovery. *The American Journal of Human Genetics*. 2008 Sep; 83(3):322–336. doi: [10.1016/j.ajhg.2008.08.001](https://doi.org/10.1016/j.ajhg.2008.08.001) PMID: [18760393](https://pubmed.ncbi.nlm.nih.gov/18760393/)
72. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, et al. Genetic Evidence for Recent Population Mixture in India. *American Journal of Human Genetics*. 2013 Sep; 93(3):422–438. doi: [10.1016/j.ajhg.2013.07.006](https://doi.org/10.1016/j.ajhg.2013.07.006) PMID: [23932107](https://pubmed.ncbi.nlm.nih.gov/23932107/)
73. Yang WY, Platt A, Chiang CWK, Eskin E, Novembre J, Pasaniuc B. Spatial Localization of Recent Ancestors for Admixed Individuals. *G3: Genes|Genomes|Genetics*. 2014 Dec; 4(12):2505–2518. doi: [10.1534/g3.114.014274](https://doi.org/10.1534/g3.114.014274) PMID: [25371484](https://pubmed.ncbi.nlm.nih.gov/25371484/)
74. Yang WY, Novembre J, Eskin E, Halperin E. A model-based approach for analysis of spatial structure in genetic data. *Nature genetics*. 2012 May; 44(6):725–731. doi: [10.1038/ng.2285](https://doi.org/10.1038/ng.2285) PMID: [22610118](https://pubmed.ncbi.nlm.nih.gov/22610118/)
75. Bookstein FL. Principal warps: thin-plate splines and the decomposition of deformations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 1989 Jun; 11(6):567–585. doi: [10.1109/34.24792](https://doi.org/10.1109/34.24792)
76. Sampson PD, Guttorp P. Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal of the American Statistical Association*. 1992; 87(417):108–119. doi: [10.1080/01621459.1992.10475181](https://doi.org/10.1080/01621459.1992.10475181)
77. Lazaridis I, Patterson N, Mitnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. 2014 Sep; 513(7518):409–413. doi: [10.1038/nature13673](https://doi.org/10.1038/nature13673) PMID: [25230663](https://pubmed.ncbi.nlm.nih.gov/25230663/)
78. Chakraborty R, Weiss KM. 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*. 1988 Dec; 85(23):9119–9123. doi: [10.1073/pnas.85.23.9119](https://doi.org/10.1073/pnas.85.23.9119) PMID: [3194414](https://pubmed.ncbi.nlm.nih.gov/3194414/)
79. Gravel S. Population Genetics Models of Local Ancestry. *Genetics*. 2012 Jun; 191(2):607–619. doi: [10.1534/genetics.112.139808](https://doi.org/10.1534/genetics.112.139808) PMID: [22491189](https://pubmed.ncbi.nlm.nih.gov/22491189/)
80. De A, Durrett R. Stepping-Stone Spatial Structure Causes Slow Decay of Linkage Disequilibrium and Shifts the Site Frequency Spectrum. *Genetics*. 2007; 176(2):969–981. doi: [10.1534/genetics.107.071464](https://doi.org/10.1534/genetics.107.071464) PMID: [17409067](https://pubmed.ncbi.nlm.nih.gov/17409067/)
81. Barton NH, Etheridge AM, Kelleher J, Véber A. Inference in two dimensions: Allele frequencies versus lengths of shared sequence blocks. *Theoretical Population Biology*. 2013; 87(0):105–119. Coalescent Theory doi: [10.1016/j.tpb.2013.03.001](https://doi.org/10.1016/j.tpb.2013.03.001) PMID: [23506734](https://pubmed.ncbi.nlm.nih.gov/23506734/)
82. Roberts GO, Rosenthal JS. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*. 2009; 18(2):349–367. doi: [10.1198/jcgs.2009.06134](https://doi.org/10.1198/jcgs.2009.06134)
83. Rosenthal JS. Optimal Proposal Distributions and Adaptive MCMC. In: Brooks S, Gelman A, Jones GL, Meng XL, editors. *Handbook of Markov Chain Monte Carlo*. 1st ed. Handbooks of Modern Statistical Methods. Florida, USA: Chapman & Hall, CRC; 2011.
84. Roberts GO, Gelman A, Gilks WR. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*. 1997; 7:110–120. doi: [10.1214/aoap/1034625254](https://doi.org/10.1214/aoap/1034625254)
85. Roberts GO, Rosenthal JS. Optimal scaling for various Metropolis-Hastings algorithms. *Statist Sci*. 2001; 16(4):351–367. doi: [10.1214/ss/1015346320](https://doi.org/10.1214/ss/1015346320)