A Novel Spatial Framework for Understanding Population Structure

Gideon S. Bradburd[1,a], Peter L. Ralph[3,b], Graham M. Coop[1,c]

[1]Center for Population Biology, Department of Evolution and Ecology, University of California, Davis, CA 95616

[3]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089

[a]gbradburd@ucdavis.edu; [b]pralph@usc.edu; [c]gmcoop@ucdavis.edu

Abstract

Patterns of genetic variation in modern populations reflect the evolutionary outcome of a complex history of demographic and migrational events. Many methods have been developed to visualize and infer aspects of population history from genetic data to illuminate how populations have diverged, been connected by gene flow, and how population sizes have fluctuated over time. One popular set of methods focus on estimating a population phylogeny, using allele frequency covariances, in which shared branch length on a tree represents shared evolutionary history between a pair of populations sampled in the modern day. There has been renewed interest in such methods as a null model to detect gene flow (admixture) between disprate parts of the population phylogeny by looking for anomolously patterns of allele frequency covariance. However, patterns of population differentiation are rarely tree-like, as migration and colonization continuously re-shape patterns of relatedness between populations. Isolation by distance (IBD), in which the covariance of allele frequencies between pairs of populations decreases with the distance between populations, may offer a more natural null hypothesis. Here, we present a novel analytical framework, SpaceMix, for the study of spatial genetic variation and genetic admixture under a model of isolation by distance. The method uses a simple spatial-statistical model of the decay of allele frequency covariance with geographical distance to estimate a visual geogenetic map on which populations are located in such a way as to maximize the posterior probability of the allele frequencies given their geogenetic distances to each other. In this framework we model admixture as unusually high covariance in allele frequencies between locations given their geographic (geogenetic) distance to each other. We estimate the admixture proportion and geogenetic source of admixture for each sample, and visualize these admixture events through arrows added to the geogenetic map. To illustrate the strengths and pitfalls of Spacemix we applied the method to data from Green Warblers and Human populations.

(I think too much of the abstract is about time and history, when we don't explicitly model time; shouldn't it focus on visualization (of, say, historical relationships)?)

Introduction

There are many different population genetic approaches to learn how population structure and historical contingencies of demographic have left their mark on patterns of genetic variation within and between populations. At one end there are model-based approaches that focus on developing a detailed view of the migrational history of a small number of populations. There has been considerable recent progress in this area, using a variety of summaries such as the allele frequency spectrum ([Gutenkunst et al., 2009], Song, Excoffier), or approximations to the coalescent applied to sequence data (Song, Li and Durbin,). At the other end of this axis sits a broad set of approaches that focus on visualizing patterns of genetic relatedness and population structure. Such methods are usually less dependent on a particular population genetic model, and can deal with many populations or individuals as the unit of analysis. Examples of this second set of methods include cluster-based methods [**???**] and reduced dimensionality representations of the data, such as Principal Components (PC) Analysis (e.g. Menozzi, Piazza, and Cavalli-Sforza 1978, Novembre et al 2008, Novembre and Stephens 2008, McVean 2009,Yang et al 2012, Yang et al 2014).

Among the wide variety of approaches that sit between these two end points, one large set of methods for visualizing patterns of relatedness work by estimating from from allele frequencies a population phylogeny of the evolutionary history shared between sampled individuals or populations. This approach of modeling relatedness between populations within species as a tree-like structure was pioneered by Cavalli-Sforza & Edwards, (&Thompson?), as were methods to test whether a tree was a good model of population history [**?**] (see [**?**] for a review).

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, where admixture is defined as covariance between samples in excess of that predicted by their tree-like history (Reich and friends). Others, such as Treemix [**?**] and mixmapper (cite mixmapper), focus on data analysis and visualization, and fit a population model represented by a directed acyclic graph between populations. TreeMix accommodates reticulate population structure by allowing branches on the population tree to be connected by arrows of admixture that explain excess residual covariance between populations or population 'clades'.

These tree-based methods are valuable as both genetic inference and visual-

ization tools. However, gene flow is frequently pervasive, and the history of populations will often not be well represented by a tree. A more natural framework for modeling genetic differentiation between sampled populations or individuals may be that of isolation by distance (IBD) (Wright). The empirical pattern of IBD, in which genetic differentiation increases with the geographic distance between populations, is ubiquitous in nature (Meirmans 2012). As a result, the first principal axes of genetic variance (in, e.g., PC analysis) within continents or other geographically restricted areas often correspond strongly to geographic coordinates [?], which allows estimation of the location of individuals whose origin of ancestry is unknown [???]. Modeling such empirical patterns relies only on the weak assumption that an individual's mating opportunities are geographically limited by dispersal, and 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 (REF).

how about we call the framework "geographically continuous models" or some such; and the observed pattern IBD? since "IBD" is not a framework or a model.

brief pointer to theory on continuous pops, e.g. Felsenstein, Pain in the Torus; Barton, Depaulis & Etheridge?

Under a spatially continuous model, "admixture" can be thought of as the outcome of unusually long-distance gene flow, and its signature as relatedness between a set of populations that is anomalously high given the distance that separates them. Here, we present an analytical framework for studying the spatial distribution of genetic variation and genetic admixture based on a model of isolation by distance. 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. These 'geogenetic' maps are simple, intuitive, low-dimensional summaries of population structure, and they provide a natural framework for the inference and visualization of spatial patterns of genetic variation and the signature of genetic admixture. *(contrast to PCA?)*

Data

Our data consist of L unlinked variable loci sampled across K populations, as well as the geographic coordinates, denoted G_k, (latitude and longitude)at which those populations were sampled (although G_k may be missing for some or all of the populations). For convenience, we use bi-allelic single nucleotide polymorphisms (SNPs) as the genetic data in our descriptions below, but we note that this method could be applied equally well to other types of data, including microsatellites. We refer to the K accessions as populations, but each "population" could be a single individual. We summarize these genetic

needs to know is lat/long? also, k not yet defined

data by the allele counts, arbitrarily choosing an allele to count at each locus, and denoting the number of the counted allele at locus ℓ in population k as $C_{\ell,k}$ out of a total sample size of $S_{\ell,k}$ alleles. The sample frequency of the counted allele at locus ℓ in population k is therefore $\hat{f}_{\ell,k} = C_{\ell,k}/S_{\ell,k}$. Our method models the covariance in these allele frequencies between populations across loci.

The Normal Approximation to Drift

In order to model this covariance, we draw on previous work (Cavalli-Sforza & Edwards, Nicholson et al 2002, Coop et al 2010, Pickrell & Pritchard 2012), and model the process of genetic drift - the compouding of binomial sampling over the generations - as approximately Gaussian. *(Let's frame this not in tree-land? And separate calculations of covariance (general) from Gaussian assumption? Instead, could set up like: " The covariance between allele frequencies is given by a function of their unknown shared history (which we could write down here if it's the place for it?); in simple treelike models it looks like this: now pretend, as others have, that allele frequencies are Gaussian.")* Briefly, if the population frequency of allele ϵ in an ancestral population A is ϵ_A, then we model the population frequency of that allele in a daughter population, B, as follows:

$$p \sim N(\epsilon_A, \delta(\epsilon)(1-\epsilon)) \qquad (1)$$

where the parameter δ_B is the amount of genetic drift separating A and B, a parameter shared by all neutral loci in the entire genome. The binomial variance term in the normal variance in (1) ignores the fact that the drift variance will change from generation to generation due to changes in the frequency of ϵ within a population over time. This approximation will work best for alleles at intermediate frequency, and in situations where the extent of drift is limited (short time-scales or large population sizes).

define "amount of drift"?

Among populations, the differentiating process of drift is counteracted by the homogenizing force of migration, so that populations with higher levels of historical or ongoing migration can be thought of as having more shared drift, i.e. stronger covariance in their allele frequency deviations around the ancestral frequency ϵ. We model the population frequencies at a locus *(with allele frequency ϵ)* as multivariate normal (MVN) with mean ϵ and covariance matrix $\epsilon(1-\epsilon)\Omega$. The multivariate normal distribution offers a natural statistical framework for describing this covariance, *(in the setup so far, we are modeling allele frequencies, not the covariance (haven't said that empirical covariance is a sufficient*

statistic), so the framework is for describing the allele frequencies, and we then put a parametric form on the covariance? Hm, to transition into next paragraph need some description of why we are modeling the covariance.) which may be straightforwardly modeled as a parametric function of any pairwise distance variable (see, e.g., *(others, first)*; Bradburd, Ralph, and Coop 2013).

Spatial Covariance Model We wish to model the covariance between populations as the result of a spatial process, in which migration rates between nearby populations are higher than those between distant ones, so that a population has higher covariance with a close neighbor than with a more distant population. This modeling framework is inspired by that of Wasser et al (2004), in which the authors assign individuals onto a spatially smoothed landscape of geographically placed allele frequencies. We choose as the form of our parametric covariance matrix a simple and flexible model with exponential decay of allele frequency covariance with geographic distance (Diggle et al 1998, Wasser et al 2004, Bradburd, Ralph, and Coop 2013). The covariance between *(normalized)* population allele frequencies in populations i and j is

explain

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

where $D_{i,j}$ is the geographic distance between population i and j (and therefore a function of their locations, G_i and G_j), α_0 controls the within-population variance, or the covariance when distance between points is 0 (the sill of the covariance matrix), α_1 controls the rate of the decay of covariance per unit pairwise distance, and α_2 determines the shape of that decay. While previously we and others have used a logit link function [**??**], or truncated allele frequencies at 0 and 1 [**??**], combined with binomial sampling to model the sample frequencies [**??**], here we chose to treat these sample frequencies themselves as multivariate normal across populations [**?**]. To accommodate the effect of finite sample size, we introduce "nuggets" of population-specific variance terms on the diagonal of allelic covariance matrix, which accommodate the effects of both sampling and unshared genetic drift.

explain nuggets with less jargon?

wait, $\Omega_{i,j}$ defined twice. I infer that $\delta = 0$. drop previous one.

$$\Omega_{i,j} = \frac{1}{\alpha_0}\exp\left((-\alpha_1 D_{i,j})^{\alpha_2}\right) + \delta_{i,j}\left(\frac{1}{\bar{S}_i} + \eta_i\right), \tag{3}$$

where $\delta_{i,j}$ is the indicator function 1 if $i = j$, and 0 otherwise. Our η is a vector of population-specific variances, where η_k, the nugget estimated in population k,

subscript on S and η; write $1/S$ instead so it doesn't parse as a matrix

represents any genetic drift (or more generally excess allele frequency variance) not captured by our spatial model. Finally, $\bar{S}$ is a vector of the mean sample sizes in each population, where $\bar{S}_k$ is the mean sample size across all loci in population k. We then assume that the vector of sample allele frequencies at locus ℓ, $\hat{f}_\ell$, has distribution

$$\hat{f}_\ell \sim MVN(\epsilon_\ell, \epsilon_\ell(1-\epsilon_\ell)\Omega) \tag{4}$$

with Ω given by equation (3).

Given the assumption of multivariate normality for our sample frequencies, it follows that the sample covariance of our standardized sample frequencies calculated across loci ~~($\widehat{\Omega} = \hat{f}\hat{f}^T$)~~ is Wishart distributed with degrees of freedom equal to the number of loci (L) across which the covariance is calculated. *(P&G confirm?). (Besides the factor of L, need to subtract mean and divide by $\epsilon(1-\epsilon)$ for this to be true; but (6) says the right thing.)* That is,

$$L\widehat{\Omega} \sim \mathcal{W}(\Omega, L) \tag{5}$$

We can calculate an estimate of the covariance matrix (Ω) across loci as

$$\widehat{\Omega} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{(\hat{f}_\ell - \epsilon_\ell)(\hat{f}_\ell - \epsilon_\ell)^T}{\epsilon_\ell(1-\epsilon_\ell)}. \tag{6}$$

Then, if we define our standardized sample allele frequencies, $\hat{X}_\ell$, as

$$\hat{X}_\ell = (\hat{f}_\ell - \epsilon_\ell)/\sqrt{\epsilon_\ell(1-\epsilon_\ell)}, \tag{7}$$

the expression given in Eqn. (6) gives the sample covariance matrix of our standardized sample allele frequencies. Given the assumption of multivariate normality for our sample frequencies, it follows that this sample covariance of our standardized sample frequencies calculated across loci ($L\widehat{\Omega} = \hat{X}\hat{X}^T$) is Wishart distributed with degrees of freedom equal to the number of loci (L) across which the covariance is calculated. That is,

$$P(\widehat{\Omega} \mid \Omega) = \mathcal{W}\left(L\widehat{\Omega} \mid \Omega, L\right) \tag{8}$$

In practice, we do not get to observe ϵ_ℓ and so we replace ϵ_ℓ with $\bar{f}_\ell$, the mean sample allele frequency at locus ℓ. We discuss the implications of this move in the Methods.

so put what we actually do in the definition.

We can estimate the parameters of our simple isolation by distance model, $\alpha_0, \alpha_1, \alpha_2$, and η, by using (8) to give us the likelihood of the standardized sample frequencies across loci, as this contains all of the information about our parameters of interest. 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. We use a Bayesian approach to estimating parameter values, and we place priors on the $\vec{\alpha}$ parameters as well as independent priors on the nugget in each population. These priors are detailed in Table 2, and for more details on our inference procedure, please see the Methods. *(Have some link to an appendix describing what to do with linked loci?)* *(still want to try adding them together and appealing to central limit theorem)*

also, fact that sample covariance is a sufficient statistic for mean-centered multivariate Gaussian.

Accommodating heterogeneous processes The model assumes that the variance in allele frequencies is the same in all locations and that the covariance between pairs of populations decays in the same way with geographic distance in all portions of the sampled range (i.e. that the process is homogeneous and isotropic). However, in many cases these assumptions will not be met. 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.

but, nugget?

One way that we can accommodate these heterogeneous processes is to allow populations to choose their own locations with respect to each other in order to maximize the resemblance to isolation by distance. *(or, "...to infer the locations of populations on a map that reflects 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 choose estimated 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 choose locations that are farther apart. The result is a "geogenetic" map in which the distances between populations are indicative of the way that populations perceive the distances between themselves.

To generate this map, we can treat populations' geographic locations as random variables and estimate them as part of our Bayesian inference procedure. The set of coordinates where the populations were sampled is G, and we designate the set of location parameters (i.e. coordinates in the geogenetic map) for our

populations as G'. Our parametric covariance matrix Ω, as given by equation (3), is a function of the pairwise matrix of distances between these inferred locations, $D(G')$, and our $\vec{\alpha}$ and η parameters. We acknowledge this dependence by writing $\Omega(\vec{\alpha}, D(G'), \eta)$. We write the posterior probability as

$$P(G', \vec{\alpha}, \eta \mid \widehat{\Omega}, L, G) \propto P(\widehat{\Omega} \mid \Omega(\vec{\alpha}, D(G'), \eta) P(\vec{\alpha}) P(G') P(\eta) \tag{9}$$

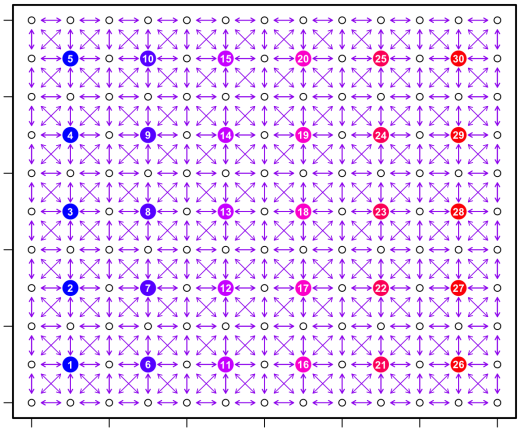
where the constant of proportionality is the normalization constant of the posterior. Here $P(G')$ are the priors for the spatial locations,which we assume are independent across populations. *(We use a very weak)* prior on population k's location parameter, G'_k: a bivariate normal spatial prior centered on the observed location G_k and with a variance of half the mean pairwise distance between the observed population locations. We then use Markov chain Monte Carlo to estimate the posterior distribution on the parameters; for more details see the Methods. This method is implemented in a program called SpaceMix. *(can we insert space-age sound effects whenever this name appears?)*

distinguish from actual locations

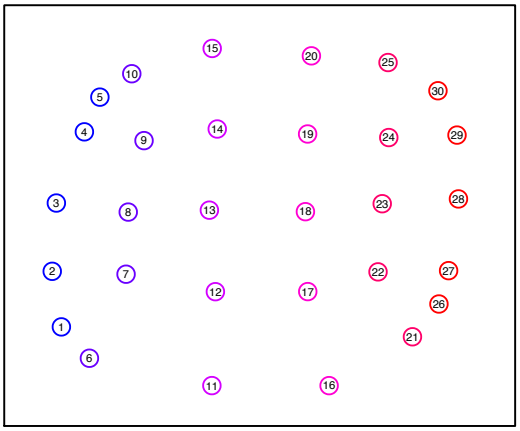
The sampled locations are a natural prior on where populations are positioned under the null of isolation by distance. This prior also encourages the resulting inferred geogenetic map to be anchored in the observed locations and to represent (informally) the minimum distortion to space necessary to satisfy the constraints placed by genetic similarities of populations. However, one concern is that if the method returns a spatial configuration that strongly resembles the observed population map, this could reflect the possibility that there is insufficient information in the data to overcome the influence of the prior, or that the MCMC has not been run long enough. To address these concerns, we can also use random locations as the "observed" locations (G_k) to remove any influence of the observed map on the output via the prior, or we can change the variance on the spatial priors to ascertain the effect of the prior on inference.

perhaps this discussion of the prior belongs lower down; only mention "a non-informative prior is used" or something?

To illustrate this inference procedure, we present several scenarios simulated under the coalescent (using ms, (Hudson)), as well as the results of SpaceMix analyses run on them. In all scenarios, we simulate variations of a stepping stone model in which populations are arranged on a grid with symmetric nearest neighbor migration (1a), with 10 haploid individuals sampled from every other population at 10,000 unlinked loci (for details on all simulations, see Methods). On all simulated datasets, we perform SpaceMix analyses in which we treat population locations as random variables to be estimated as part of the model, and randomly place the 'observed' location of each population so as to remove the influence of the prior. For clarity, we present the full Procrustes superimposition of the inferred locations around the coordinates used to simulate the data.



(a) simulated lattice



(b) lattice inference

Figure 1: Simulation scenarios and SpaceMix inference. a) configuration of simulated populations on a simple lattice at equilibrium; b) inference of population locations under this scenario.

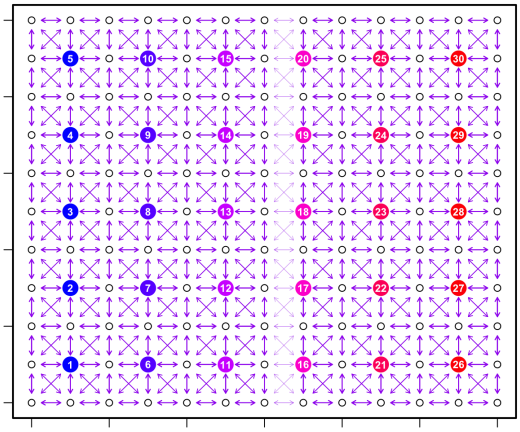
In the first scenario (output illustrated in Fig. 1b), we simulate a stepping stone model at migration-drift equilibrium with homogeneous migration rates across the grid. In Figure 1b, the reader can see that the configuration estimated for the populations by SpaceMix matches the lattice structure used to simulate the data, and that populations are correctly choosing their nearest neighbors. Populations at the edge of the lattice have the least amount of data informing their exact placement, and are therefore pulled in somewhat.

In our second scenario (Fig. 2b), we simulate under the same stepping stone model, but introduce a longitudinal barrier to dispersal (between populations 11:15 and 16:20), across which migration rates are attenuated by a factor of 5. In Figure 2b, the population configuration matches that of the lattice used to simulate the data, but due to the influence of the barrier, the two halves of the map have pushed farther away from one another. This gap between populations on either side of the barrier reflects the way those populations perceive the increased effective distance between them.

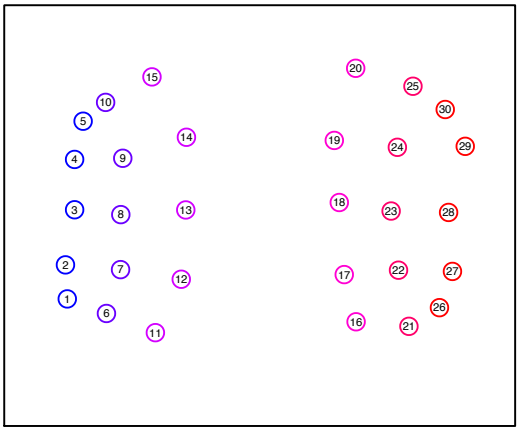
In the third scenario (Fig. 3b), we simulate an expansion event in which all populations in the last five columns of the grid have expanded in the recent past from the nearest population in their row (e.g., populations 25 and 30, as well as the three unsampled populations that bracket them, have all expanded at the same point in the recent past from population 20). In the scenario of recent expansion (Fig. 3b), the daughter populations of the expansion event cluster with their parent populations, reflecting the higher relatedness (per unit geographic separation) between them. In SuppMat Figs 13, we show the relationship between covariance in allele frequencies and geographic distance and inferred geogenetic distance for these simulations.

If our data are well fit by a model of isolation by distance, then (1) a population's genetic makeup should be well predicted by that of its neighbors, and (2) populations should not show excess covariance with distant populations. Violation of either of these two points will result in a poor fit of a simple isolation by distance model, and particularly a violation of point (2) may indicate that long distance admixture has taken place.

To examine the behavior of SpaceMix when there is long distance covariance between populations, we simulated an admixture event on the stepping-stone model we had used previously. Specifically, (using Fig. 4a as a reference) we allowed population 30, in the northeast corner of the grid, to draw half of its ancestry from population 1, in the southwest corner. The result of a SpaceMix

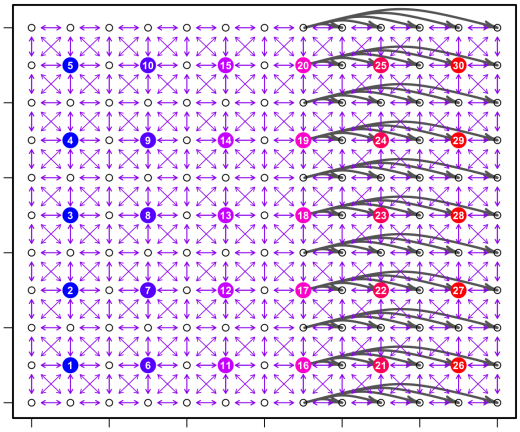


(a) simulated lattice with barrier

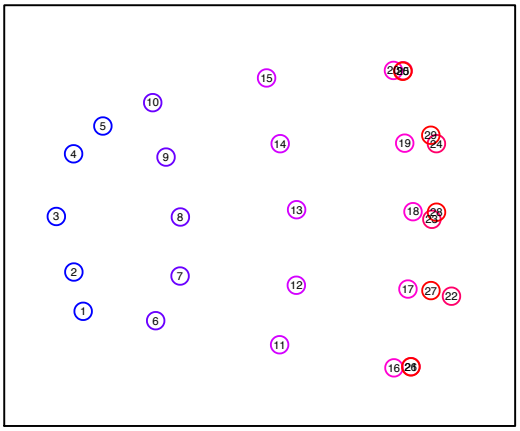


(b) barrier inference

Figure 2: Simulation scenarios and SpaceMix inference. a) a lattice with a barrier across the center line of longitude; b) inference of population locations under this scenario.



(a) simulated lattice with expansion



(b) Expansion

Figure 3: Simulation scenarios and SpaceMix inference. a) a lattice with recent expansion on the eastern margin; b) inference of population locations under this scenario.

analysis in which the locations of these populations were estimated is shown in Figure 4b.

This signal of excess covariance over anomalously long distances is clearly difficult to accommodate within the "choose-your-own-location" framework described above. In Figure 4b, the reader can see the torturous lengths to which the method goes to come up with a configuration of populations that accommodates their genetic relationships. The admixed population 30 is estimated to have a location intermediate 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 especially so in the visualization of genetic relationships in empirical data for which a researcher does not know the true demographic history. It would therefore be of great utility to directly model the action of admixture on spatial patterns of genetic variation.

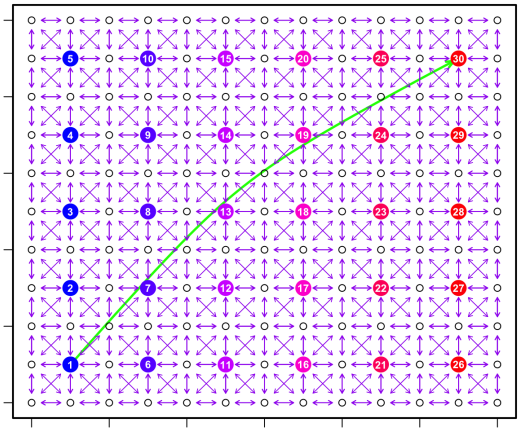
Inference of Spatial Admixture

We can incorporate recent admixture directly into out inference framework. We imagine that population k draws the majority of its ancestry from G'_k, but that a proportion $0 \geq w_k \leq 0.5$ of its ancestry comes from another location G^*_k, which we refer to as its source of admixture. The mean standardized population allele frequency at locus ℓ in population k is a weighted average of the allele frequencies at the geographic location of the sampled population chooses for itself ($p_{\ell,k'}$) and those at the coordinates of the source (p_{ℓ,k^*}) from which the observed population draws admixture:

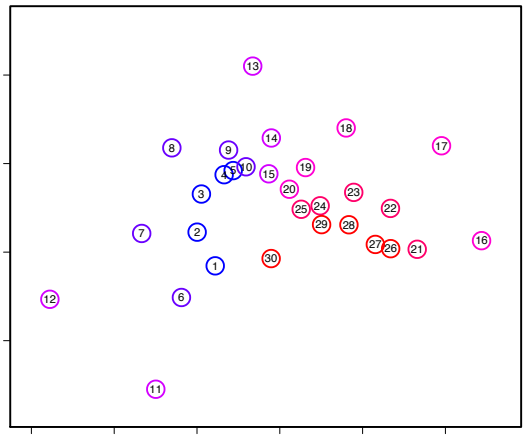
$$w_k p_{\ell,k'} + (1 - w_k) p_{\ell,k^*} \tag{10}$$

We can allow each of our populations to have an independent spatial source of admixture. Our $p_{\ell,k'}$ and p_{ℓ,k^*} are well defined, as, given our population allele frequency covariance matrix, equation 2, we can get the distribution of allele frequencies at any point in space simply by plugging its spatial coordinates into our $\Omega^{(P)}(D(G))$.

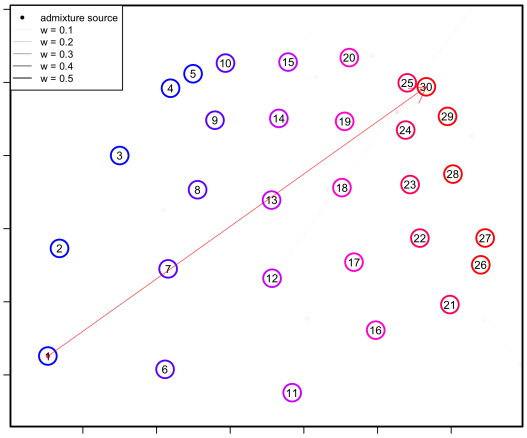
We can then consider the resulting parametric covariance matrix that follows from the form of the population frequencies in equation (10). The covariance between the standardized allele frequencies of population i and j can be modeled



(a) simulated lattice with admixture



(b) location Inference



(c) location and admixture inference

Figure 4: Simulation scenarios and SpaceMix inference. a) a lattice with recent admixture event between population 1 in the southwest corner and population 30 in the northeast corner; b) inference of population locations under this scenario; c) inference of population locations and their sources of admixture under this scenario.

as

$$\begin{aligned}\Omega^*_{i,j} = (1-w_i)(1-w_j)\Omega^{(P)}_{i\ ,\ j} \ \times \\ (w_i)(1-w_j)\Omega^{(P)}_{i^*,\ j} \ \times \\ (w_j)(1-w_i)\Omega^{(P)}_{i\ ,\ j^*} \times \\ (w_i)(w_j)\Omega^{(P)}_{i^*,\ j^*} + \\ \delta_{i,j}(\eta_i + \frac{1}{\bar{S}})\end{aligned} \tag{11}$$

where i^* and j^* are the sources from which populations i and j are drawing their admixture with proportions p_i and p_j, and the spatial covariance function of population frequencies, Ω, is given by (2) (this lacks the sample-specific variance terms on the diagonal). Note that we then reintroduce the nugget, η_k, and the sample size effect, ${\bar{S}_k}^{-1}$, for each population, to model drift or excess variance in population k on top of that predicted by the mixture of frequencies given by our spatial model. As in equation (3), $\delta_{i,j}$ is the indicator function 1 if $i = j$, and 0 otherwise.

The admixed covariance between samples i and j, $\Omega^*_{i,j}$ is then a function of all the pairwise spatial covariances between populations i and j and the points from which they draw admixture, i^* and j^*. Those spatial covariances in turn are a function of all combinations of pairwise distances between their locations: G_i, G_j, G^*_i, and G^*_j. This parametric covariance form is illustrated in Figure 5.

As we only get to observe the sample frequencies and we standardize our allele frequencies using the sample mean, our predicted admixture covariance matrix needs to be transformed to accommodate these sampling considerations. We can do this as before (see Methods), and we again treat the likelihood of our sample covariance matrix as Wishart:

$$P(\widehat{\Omega} \mid \Omega^*) = \mathcal{W}\left(L\widehat{\Omega} \mid \Omega^*\left(G', G^*, w, \vec{\alpha}, \eta\right), L\right). \tag{12}$$

As before G', $\vec{\alpha}$, η are treated as random variables to be estimated. Now the set of K admixture sources G^* are also parameters to be estimated, along with the vector of K admixture proportions, w. The posterior probability of these parameters can be expressed as a function of this parametric admixed covariance, Ω^*,

$$P(G', G^*, w, \vec{\alpha}, \eta \mid \widehat{\Omega}, L, G) \propto P(\widehat{\Omega} \mid \Omega^*)P(\vec{\alpha})P(G')P(G^*)P(w)P(\eta) \tag{13}$$

as specified by the parameters w, G^*, $\vec{\alpha}$, and η, and the observed locations, G. We place the same priors as stated above on G', $\vec{\alpha}$, η, and we now specify

priors on w and G^*. The admixture proportions, w, are capped at 0.5, to prevent populations from swapping identities with their source of admixture, and are heavily weighted towards small values to be conservative with respect to admixture inference. Our admixture proportions are independently beta-distributed: $2w_k \sim \beta(\alpha = 1, \beta = 100)$. The priors on the sources of admixture, G^* are taken independently as bivariate normal spatial distributions, all with the same mean at the centroid of the observed population locations, G, and variance equal to twice the mean pairwise observed distance: $G^*_k \sim \mathcal{N}(\mu = \bar{G}, \sigma = 2\bar{D}(G))$.

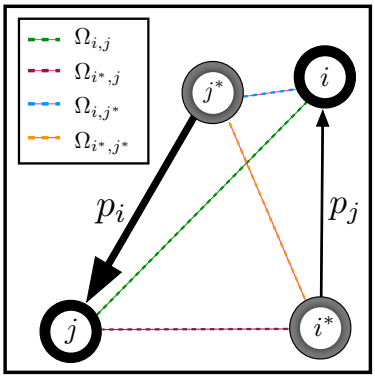


Figure 5: An illustration of the form of the admixed covariance given in equations (11). Populations i and j are drawing admixture in proportions p_i and p_j from their respective sources of admixture, i^* and j^*, and all pairwise spatial covariances (the Ω's) are shown. In this cartoon example, population j is drawing more admixture from its source j^* than i is from its source i^* (i.e., $p_j > p_i$).

The models described above may be used in various combinations. In the simplest model, populations do not choose their own locations, nor are they allowed to draw admixture, and the only parameters to be estimated are those of the spatial covariance function given in equation (2), and the population-specific variance terms (η). In the most complex model, population locations, the locations of their sources of admixture, and the proportions of that admixture are all estimated jointly in addition to the parameters of the spatial covariance function and the population specific variances.

An natural concern is whether all of these parameters are separately identifiable, most notably whether population locations, admixture locations, and proportions can be estimated. That is, if a population has received some level of admixture from another population, what is to stop it from simply moving

toward that population in geogenetic space to satisfy its increased resemblance to that population, rather than choosing admixture from that location?

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 Figure 4b). To illustrate this point, consider the simple spatial admixture scenario shown in Figure 6. Our populations A-D are spatially arrayed along a line, and their allele frequencies are specified by a simple isolation by distance model, but there is recent admixture from D into B (such that 40% of the alleles in B are drawn from D). The lines show the expected covariance under IBD 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 resulting admixed covariance between B and the three other populations, as well as B's variance with itself (B-B) as specified by equation (11), with no nugget or sampling effect.

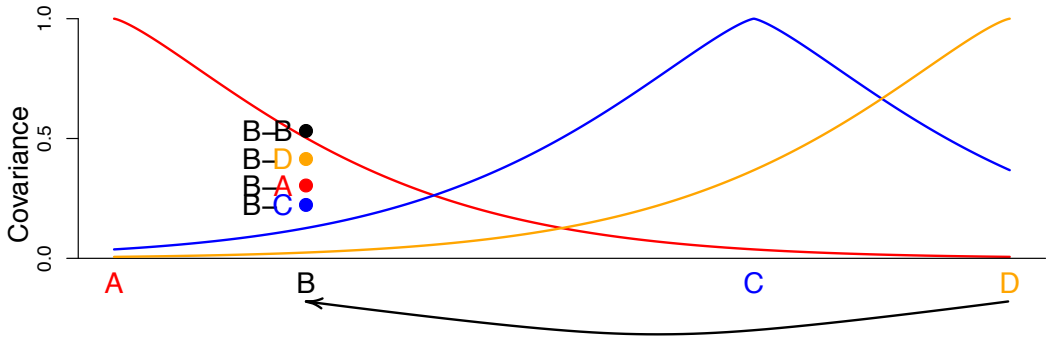


Figure 6

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 non identifiable. For the sake of simplicity, assume that we hold the locations of A, C, and D constant, as well as the decay of covariance with distance (as could be the case if A-D are part of a larger analysis). The covariance relationships of B to the other populations cannot be simply satisfied by moving B towards D. Doing so would

better match B's covariance to D, but B would then have a covariance with C that is higher, and a covariance with A that is lower, than that we actually observe.

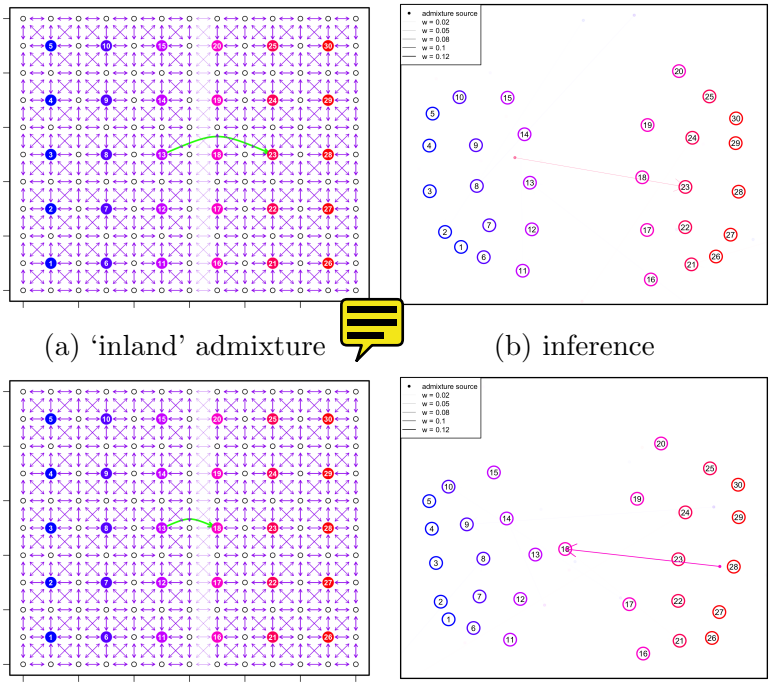
Moving B also does not resolve the mystery that its variance is lower than that of the other populations. Introducing admixture into the model allows it to satisfy all of these conditions: it can draw ancestry from D but keep part of its resemblance to A, and it also avoids B having to move closer to C. Even in the absence of a sample from population C, B's covariance with A and D and its low variance could not be satisfied simply by moving B. 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 non-identifiability between estimated population location and that of its source of admixture. We describe an example of this phenomenon below.

We now return to the example admixture scenario described above in Figure 4a, we demonstrate the inference of populations' sources and strengths of admixture and illustrate the results in Figure 4c. The reader can see that only the admixed population (population 30) is drawing admixture from the location of the source of admixture that was used to simulate the data, and that all other populations, which are not admixed, are choosing to draw admixture in only negligible amounts.

To demonstrate the use of the model in which the location of each population as well as the location of its source of admixture are estimated jointly, we used the spatial stepping-stone coalescent simulation procedure described above to generate a dataset of populations on a lattice in which there is both a barrier to dispersal and a more subtle admixture event (admixture proportion = 10%, see Fig. 7a). In the SpaceMix analysis (Fig. 7b), the separation of the east and west sides of the grid accommodates the effect of the barrier to migration, and the admixed population (population 23) chooses admixture from very close to its true source (population 13), and in close to the correct amount ($\bar{p}_{(23)} = 0.05; 95\%$ credible interval $= 0.02 - 0.08$).

We also show a scenario in which there has been admixture (40%) between two populations on either side of a barrier. Here, the admixed population 23 is the only population that chooses a non-negligible proportion of admixture, but rather than drawing this admixture from its true source (population 13), it draws admixture from a location on the far margin of the half of the grid on its own side of the barrier. The sampling scheme here illustrates more fully

the example laid out in 6: because there is no sampled intervening population between admixed population 23 and its source of admixture 13, there is nothing to stop 23 from explaining its higher covariance with 13 via its chosen location $G'_{(23)}$ rather than via that of its source of admixture ${G^*}_{(23)}$. However, 23 still has a lower variance than the other populations, and therefore must choose a proportion of admixture to accommodate that fact. This analysis is a somewhat artificial example, as the biological interpretation of "admixture" (defined in our spatial framework as anomalously long distance covariance) between near neighbors is unclear.



(a) 'inland' admixture (b) inference

(c) 'neighbor' admixture (d) inference

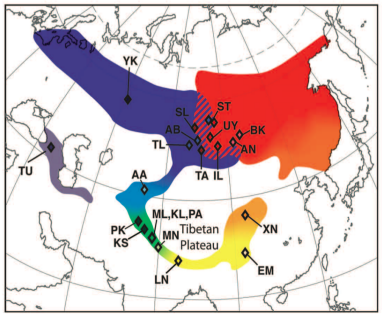
Figure 7: Simulation scenarios and inferred population maps for two different admixture scenarios: a) lattice with a barrier and an admixture event across the barrier to an 'inland' population; b) the inferred population map for the scenario in (a); c) lattice with a barrier and an admixture event across the barrier to a 'neighbor' population on the border of the barrier; (d) the inferred population map for the scenario in (c).

Empirical Applications

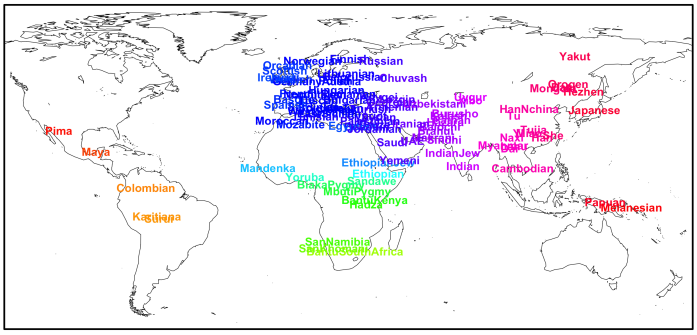
To demonstrate the applications of this novel method, we analyzed population genomic data from two systems: the greenish warbler ring species complex, and a global sampling of contemporary human populations. Maps showing our sampling in these two systems are shown in Figure 8, and information on the specific populations included is given in the Supplementary Materials, Tables 3 and 4. To minimize the potential influence of the spatial prior on population locations, for each empirical analysis, we performed a SpaceMix run in which random, uniformly distributed locations between, for longitude, the minimum and maximum observed longitude, and, for latitude, the minimum and maximum observed latitude, were used as the prior on population locations (in addition to two runs for which the observed locations are used as priors). For clarity and ease of interpretation, we then 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.

Greenish Warblers

The greenish warbler (Phylloscopus trochiloides) species complex is broadly distributed around the Tibetan plateau, and exhibits gradients around the ring in a range of phenotypes including song, as well in allele frequencies (Ticehurst (1938), Irwin et al 2001, Irwin et al 2005, Irwin et al 2008). At the northern end of the ring in central Siberia, where the eastern and western arms of population expansion meet, there are discontinuities in call and morphology, as well as reproductive isolation and a genetic discontinuity (Irwin et al 2001, Irwin et al 2008). It is proposed that the species complex represents a ring species, in which selection and/or drift, acting in the populations as they spread northward on either side of the Tibetan plateau, have led to the evolution of reproductive isolation between the terminal forms(REFs). The question of whether it constitutes a ring species, in purest form, focuses on whether gene flow along the margins of the plateau has truly been continuous throughout the history of the expansion or if, alternatively, discontinuities in migration around the species complex's range have facilitated periods of differentiation in genotype

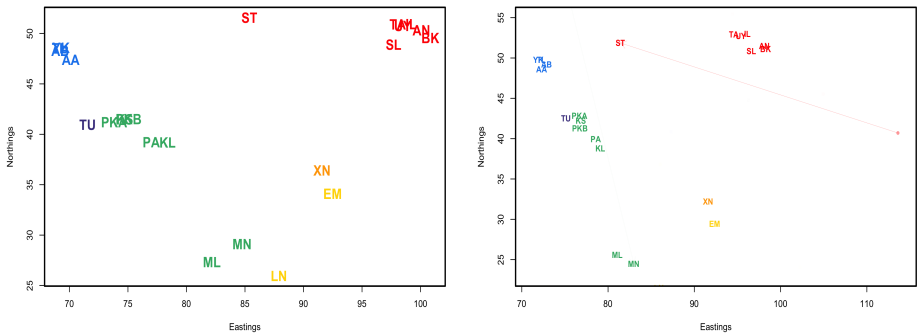


(a) Warbler subspecies distribution map



(b) Human sample distribution map

Figure 8: Sampling maps of both empirical systems analyzed. (a) greenish warbler subspecies distributions of all 22 sampled populations, consisting of 95 individuals; (b) sampling map for human dataset, consisting of 1,490 individuals from 95 sampling locations.



(a) Warbler population map, no admixture

(b) Warbler population map, with admixture

Figure 9: Inferred population maps with population labels colored as in 8a: a) the map inferred with no admixture inference; b) the map inferred with admixture inference.

or phenotype without gene flow (Mayr 1942, Mayr 1970, Coyne and Orr 2004). However, we note that many would still classify this as a ring species even if that condition were not met, just not as a case of speciation-by-distance (see Wake and Schneider 1998 for discussion).

Alcaide et al. [2014] have suggested that the greenish warbler species complex constitutes a 'broken' ring species, in which historical discontinuities in gene flow have facilitated the evolution of reproductive isolation between adjacent forms. Because the questions in this system are fundamentally both geographic and genetic in nature, it is eminently SpaceMix-able, and, within this spatial framework, we performed a number of analyses to investigate the geographic context of population differentiation in the greenish warbler species complex. For these analyses, we used the dataset from Alcaide et al. [2014], which consisted 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.

We first ran SpaceMix on the population dataset, with no admixture, setting the prior locations of the populations at random (as described above). The inferred map (Figure 9a) largely recapitulates the geography of the sampled populations. Populations choose locations around a large ring, with ordering similar to that of their true geographic locations. The Turkish population (*Phylloscopus trochiloides* ssp. *nitidus*) clustered with the populations in the subspecies *ludlowi*, but also chose a relatively high nugget parameter (see Supp-

Mat Fig **??**), reflecting the independent drift it does not share with its *ludlowi* neighbors. The Yekat population of *viridanus* individuals clusters closely with the other, less far-flung *viridanus* individuals, indicating that differentiation within that subspecies is incommensurate with the amount of IBD expected for samples separated by that much distance.

In the North, where the twin waves of expansion around the Tibetan Plateau are hypothesized to meet, the inferred geogenetic distance between populations identified as *Phylloscopus trochiloides* ssp. *plumbeitarsus* and ssp. *viridanus* was much greater than their observed geographic separation, reflecting the reproductive isolation between these adjacent forms (see Figure 14). Interestingly, the ST population, which consists of six individuals sampled in Stolby, Russia, chooses a location intermediate between the *plumbeitarsus* and *viridanus* groups. The Stolby sample is composed of three individuals that belong to the eastern *plumbeitarsus* and three individuals that belong to the western *viridanus* [Alcaide et al., 2014]. In the case where no admixture is allowed, this population is forced to adopt an intermediate position to incorporate its admixed nature.

We then ran the method allowing admixture, and again discuss the results from the analysis using random locations as priors on G' (Figure 9b). The Stolby population chooses the highest admixture proportion, with a mean of 0.19 (95% credible interval: 0.146-0.238). Multiple runs agreed well on the level of admixture of the Stolby sample (see caption of Supplementary Figure 16). What does vary across runs is whether the Stolby population chooses to locate itself by the *viridanus* cluster and draw admixture from near the *plumbeitarsus* cluster or vise versa; however, this is to be expected given the 50/50 nature of the sample's makeup (Supplementary Figure 16).

Because *a priori* assigned population membership may be artificial (individuals from more than one population may be sampled at a single site), we repeated these analyses (with and without admixture) on an individual level (Figure 10). In these analyses, the sample size in each 'population' was 2 (for the two haplotypes in a diploid), and each individual chose its own location as well as the location of its source of admixture, the proportion of that admixture, and its nugget. Because the analyses with and without admixture were nearly identical (individuals in the admixture analysis choose very low levels of admixture (see SuppMat Figure 24); the population with the highest admixture proportion draws has a mean of 0.012 (1.46×10^{-4} to 4.73×10^{-2})), we discuss the analysis with admixture inference. As with the analysis on multi-sample populations,

the results approximately mirror the geography of the individuals.

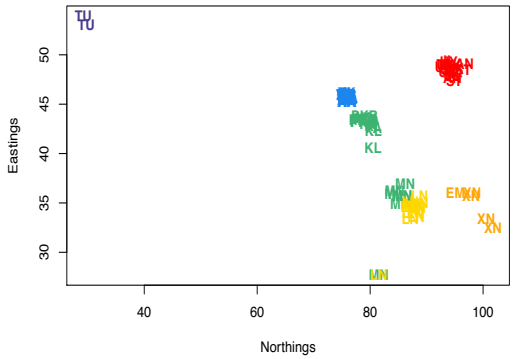
There are, however, a number of obvious departures in the individual inferred geogenetic map from the observed map. The most obvious again is the clear split between *viridanus* and *plumbeitarsus* individuals in the north at the contact zone of the two waves of expansion. This is clearer now than in the population-based analysis as individuals from the Stolby population have moved to near their respective *viridanus* and *plumbeitarsus* clusters.

Despite the fact that *viridanus* and *plumbeitarsus* individuals have moved away from each other in our geogenetic map, they are still closer to each other than we might expect if their drift were truly independent (e.g. our populations could form along a line, with *viridanus* at one end and *plumbeitarsus* at the other). This horseshoe, with *viridanus* and *plumbeitarsus* at its tips, is steady within and among runs of the MCMC and choice of position priors (see Supplementary Figures 21).

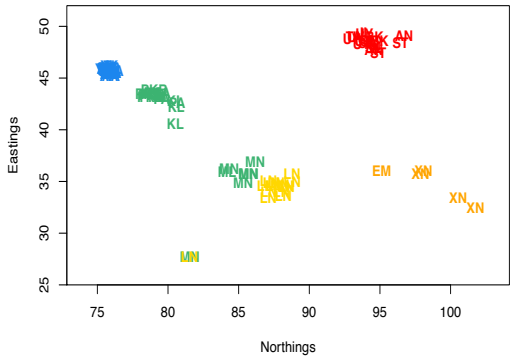
Is this biologically meaningful? A similar horseshoe shape appears when a principal components (PC) analysis is conducted and individuals are plotted on the first two PCs [Alcaide et al., 2014, see SuppMat Figure 15]. However, as discussed by Novembre & Stephens (2008) such patterns in PC analysis can arise for somewhat unintuitive reasons. If populations are simulated under a one dimensional stepping stone model, then plotting individuals on the first two PCs results in a horseshoe (e.g. see SuppMat Figure 32b) not because of gene flow connecting between the tips, but rather because of the orthogonality requirement of PCs (see Novembre & Stephens for more discussion). In contrast, when SpaceMix is applied to one dimensional stepping stone data, the placement of samples is consistent with a line (see SuppMat Figures 32c, 32d).

In addition, when we run SpaceMix on the greenish warbler individuals, specifying their location priors to fall along a straight line with samples located at their approximate positions around the horseshoe, the posterior positions of the populations still curl up to form a rough horseshoe (SuppMat Figure 27). 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. [2014], who observed evidence of hybridization between *viridanus* and *plumbeitarsus* using assignment methods.

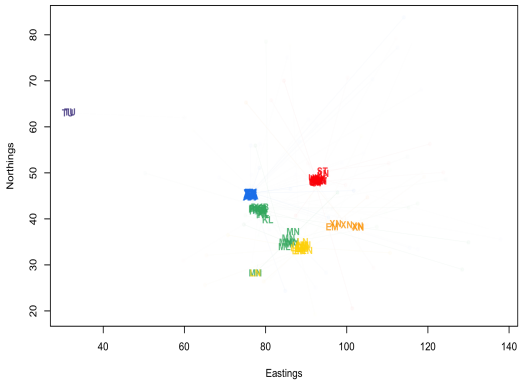
A second difference between the observed and inferred maps is a pair of indi-



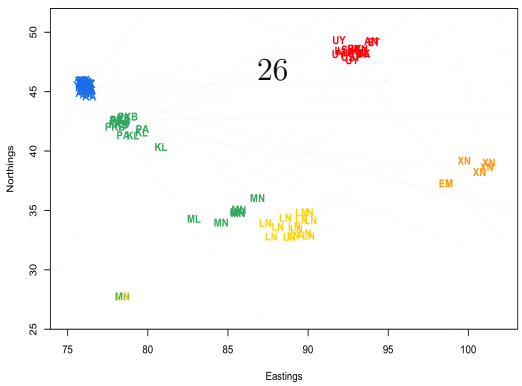
(a) Warbler individuals map, no admixture



(b) Closeup of non-*nitidus* samples



(c) Warbler individuals map, admixture



viduals, one identified as *P. t. ludlowi* (Lud-MN3), one as *P. t. trochiloides* (Tro-LN11), that choose locations very close to one another and also away from the other individuals sampled at their locations. Examining pairwise sequence divergence reveals that these two individuals show unusually recent common ancestry (see SuppMat Figure 26), and therefore are likely expressing their shared ancestry (drift unshared with other *ludlowi* and *trochiloides* individuals) by choosing locations that are close to each other and far from their respective clusters of individuals that were sampled at the same sites.

The SpaceMix map also diverges from the observed map in the distribution of individuals from the subspecies *ludlowi*. 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. [2014], makes the *ludlowi* cluster unusual compared to the estimated spatial distributions of the other subspecies (see SuppMat Figure 25).

Human Populations

The geography of human population structure is a complex product of the forces of migration, drift, and selection acting on both local and global scales. Recent work in this area has demonstrated that the signature of these forces can be read in the genomes of modern humans (e.g. Novembre et al 2008, Ralph & Coop 2013, Moorjani, Reich, ARG folks, Hellenthal et al 2014). This research has revealed that patterns of spatial genetic differentiation across human populations are byzantine in their complexity, varying across both space (Ralph & Coop 2013) and time (Skoglund et al 2012, 2014), and shaped by culture (Reich et al 2009, Atzmon et al 2010, Moorjani et al 2011), landscape (Bradburd, Ralph, Coop 2013), and environment (Beall et al 2010, Bigham et al 2010).

To visualize the patterns these processes have induced, we create a geogenetic map for a worldwide sample of modern human populations. In doing so we fully acknowledge that human history at these geographic scales has many aspects that are not well captured by isolation distance or simple admixture models. To simplify the discussion of our results, we talk about samples' locations and those of their sources of admixture, but of course both reflect the compounding of drift and gene flow over many historical processes. We therefore urge caution in the interpretation our results, and view them as an overly simplistic but rich

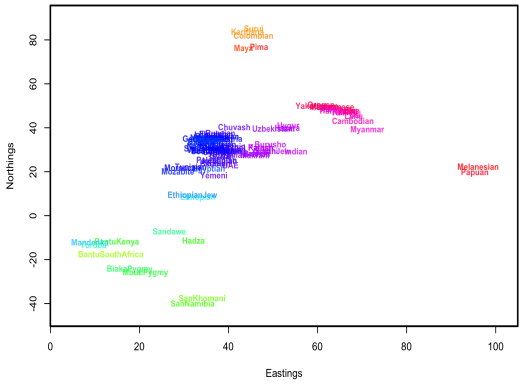
visualization of patterns of population structure in humans.

We used a random subset of 10,000 SNPs from the SNP dataset of Hellenthal et al (2014), which is comprised of 1,490 sampled in 95 populations (see Fig. 8b for map of sampling), as well as the latitude and longitude attributed to each population. We ran two sets of SpaceMix analyses: in the first, we estimated population locations, and in the second, we estimated population locations, as well as their admixture sources and amounts.

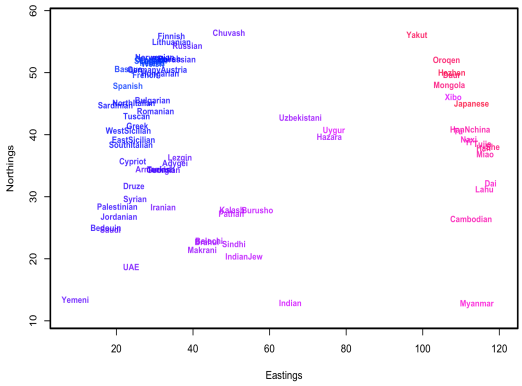
In the analysis in which populations choose their own locations, the map roughly recapitulates the geography of the samples (Fig. 11). Populations generally cluster with other populations sampled on their continent, and the relative placement of populations is similar to that of their observed geography. Sub-Saharan African populations are distributed in a manner consistent with their sampled latitude, with the San populations in the South and the Ethiopian populations closest to the Eurasian populations in the North. North African populations, such as the Moroccan, Tunisian, and Egyptian populations, cluster with Middle Eastern populations such as the Yemeni and the UAE, and are quite close to both the Ethiopian populations and the western European populations, such as the Sicilians, the Cypriots, the Tuscans, and the Sardinians.

Within Eurasia, we see populations grouping roughly in the order of their longitude on the continent, with British, continental European, and Russian populations at the western margin, and Han, Japanese, and Southeast Asian populations such as the Cambodians, the Lahu, and Myanmar on the eastern margin. Interspersed between are populations of the Sub-continent, such as the Indian, Indian Jewish, and Sindhi populations, as well as populations from farther north in Eurasia that choose intermediate locations between European and East Asian populations, such as the Uygur, the Hazara, the Uzbekistani, and the Chuvash. Populations from the Americas cluster together and choose a location close to the East Asian populations, as do the populations from Oceania.

We recover evidence for several major population expansions and colonization routes in human pre-history. We can see evidence of these expansions by examining the relationship between observed and estimated pairwise distance between the daughter populations of a putative expansion event. As can be seen in Figure 3b, populations that have recently expanded on the landscape are less differentiated both from each other and from their parent population than would be expected from their geographic separation, and this fact is reflected in



(a) Inferred map of human populations



(b) Closeup of Eurasian populations

Figure 11: Map of human populations, inferred without admixture. (a) complete map; (b) close up of Eurasian populations.

the decreased ratio between estimated and observed pairwise distances between members of an expansion event.

In the SpaceMix analysis of the human genetic data, the scale of inferred inter-population distance within Africa is much greater than that between any other group (see SuppMat Fig. 31), and the slope of the relationship between observed and estimated 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, including an out-of-Africa expansion, an expansion into East Asia, and twin expansions into the Americas and Oceania (REFS). In addition, we see that both the populations in the Americas and those in Oceania cluster close to the East Asian populations, but that the two clusters are 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 Asian populations(REF?), but that the expansion events induced independent drift trajectories in both waves of colonization.

We also see intriguing evidence for potential admixture in the placement of the Chuvash, Uzbekistani, Hazara, and Uygur populations, which choose locations intermediate between Europe and East Asia. Notably, the Xibo who are sampled from a geographic location close to the Uygur, choose a geogenetic location within the East Asian samples. The placement of the Moroccan, Mozabite, and Ethiopian populations, which choose locations between the western Eurasian cluster and the African cluster, is also suggestive of potential admixture. To investigate possible patterns of admixture further, we ran a SpaceMix analysis in which we estimated the parameters of the spatial covariance matrix along with, for each population, a location, a source of admixture, and the magnitude of that admixture.

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 drift not shared with the Sub-Saharan African samples. The inclusion of admixture allows samples that fall intermediate between Sub-Saharan Africa and North Africa and the Middle East to move closer to one or the other, which, in turn, allows each of those major clusters to move relatively farther apart. The Ethiopian and Ethiopian Jewish samples move to be closer to the Sub-Saharan samples than

the rest of the North African samples, but draw substantial amounts of admixture ($\sim 40\%$) from close to where the Egyptian sample has positioned itself in the the Middle East cluster, as do the Sandawe. The SanKhomani draw admixture from near Syria, which may reflect multiple distinct geographic sources of admixture as discussed by [**?**] and [**?**]. *(CHECK Joe's San paper)* Interestingly the Bantu South African sample, though it moves to join the other Bantu populations, draws admixture from close to the San populations. This is consistent with previous signals of the expansion of Bantu-speaking peoples into southern Africa (Pickrell et al 2012;Jakoboson, Hellenthal 2014).

The majority North African populations (Egyptian, Tunisian, Morocan, Mozabite) move to join the Middle Eastern populations (positioning 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 location chosen by the Ethiopian samples and where most of North African samples draw admixture from. It seems likely that this spread of populations represents North African influence.

A number of other populations draw admixture from Africa. The Sindhi, Makrani, and Brahui draw admixture from close to the location of the Bantu samples, and the Balochi and Kalash draw admixture from some distance away from African populations. Of the European samples, the Spanish and the East and West Sicilian samples all draw small amounts of admixture from close the Ethiopian samples presumably reflecting a North African ancestry component (Bustamante, and Reich).

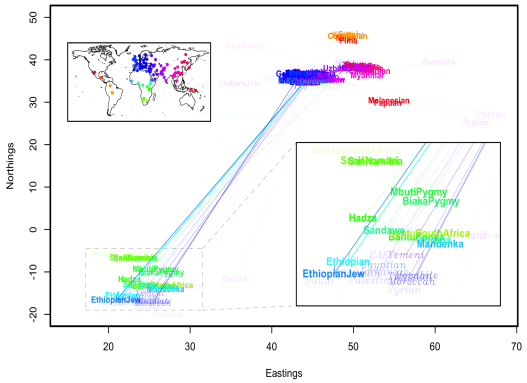
The other dominant signal of admixture is between East and West Eurasia. The majority of samples maintain their relative positions within each of these groups; however, several of the populations that chose locations intermediate between eastern and western Eurasia now move towards one side, and draw admixture from the other. The Uzbekistani and Hazara samples move to be closer to the East Asian samples, while drawing a substantial admixture proportion from close to where the Georgian and Armenian samples have located themselves, while conversely the Uygur sample moves to be close to the Burusho, Kalash, and Pathan samples. The Tu sample (who locate to East Asia) draw a small amount of ancestry from close to where the Uygur have positioned themselves. The Chuvash move close to Russian and Lithuanian samples, drawing admixture from close to the Yakut; the Turkish sample also draws a smaller amount of admixture from here. There are a number of other East-West connections: the Russian, Adygei samples have admixture from a location "north"

of our East Asian samples, and Cambodia draws admixture from close to the Eygptian sample (as was also noted by Treemix, Globetrotter).

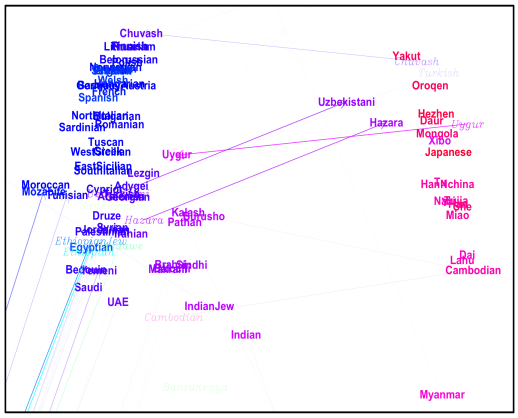
There are a number of samples that draw admixture from locations that are not immediately interpretable. For example, the Hadza and Bantu Kenyan samples draw admixture from somewhat close to India, and the Xibo and Yakut from close to "northwest" of Europe. The Pathan draw admixture from a location far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

There are a number of possible explanations for these results. As we only allow a single admixture arrow for each sample, populations with multiple, geographically distinct, sources of admixture may be choosing admixture locations that average over the sources. This may be the case for the Hadza and Bantu Keynan samples. A second possibility is that the relatively harsh prior on admixture proportion forces samples to choose lower proportions of admixture from locations that overshoot their true sources; this may explain the XIbo and Yakut admixture locations. A final explanation is that good proxies for the sources of admixture may not be included in our sampling, either because of of the limited geographic sampling of current day populations, or because of old admixture events from populations that are no long extant. The admixture into the Indians and Pathan may be an example of this, as has been hypothesized by Moorjani and Reich.

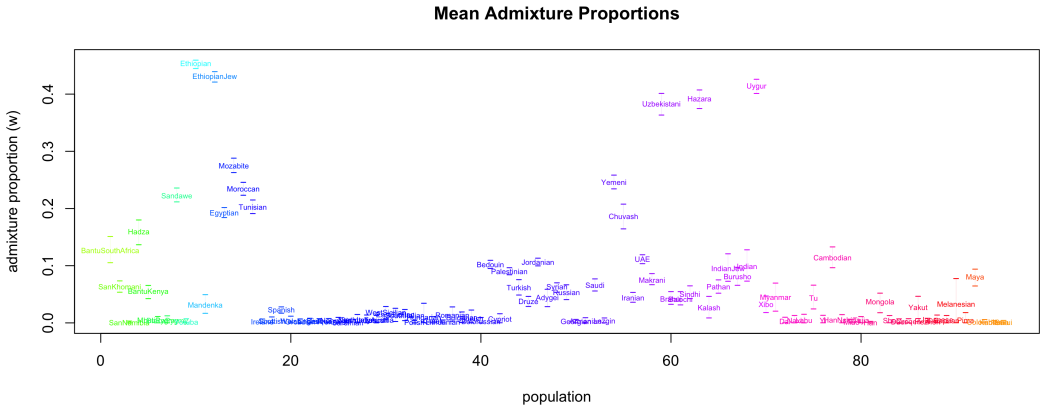
In the Supplementary Figures XXXX-ZZZZZ we show the results of other independent MCMC analyses on these data. The broad-scale patterns and results discussed above are consistent across these runs. However, as is to be expected, there is significant heterogeneity in the exact layout of sample and admixture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



(a) Inferred map of human populations



(b) Closeup of Eurasian populations



(c) Inferred admixture proportions for human populations

Figure 12: Map of human populations, inferred with admixture. (a) complete map; (b) close up of Eurasian populations; c) mean admixture proportions for each population

Conclusion

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 of a variety of simulated scenarios, which included the effects of spatially heterogeneous migration, population expansion, and population admixture. In our 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 powerful summaries of patterns of population structure.

Within the constellation of existing methods for inferring and illustrating patterns of population structure, SpaceMix falls between methods that use model-based inference to infer population trees, such as TreeMix, mixmapper, or Reich et al's f-statistics, and methods that use dimensionality reduction to visualize patterns of structure on a map. SpaceMix, unlike methods such as STRUCTURE or Geneland, does not infer hierarchical population structure, but it can be used as a useful guide in those analyses.

Caveats

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 process of spatially determined migration, and, although we have included statistical elements to accommodate deviations from spatial expectations (by, e.g., estimating population locations, allowing population-specific drift effects in the nugget parameters, and allowing the inference of admixture), it goes without saying that the true history of a sample of empirical populations is vastly more complex than any model we could create, and, as such, SpaceMix results should be interpreted with caution.

Three factors in particular stand out as potential sources of difficulty for the model. The first is that there may be empirical datasets for which the assumption of IBD is inappropriate. In taxa that migrate from their sampled locations and breed more or less panmictically elsewhere (or in the case of some broadcast

spawners), a geogenetic map inferred by SpaceMix may bear no relation to the observed sample locations, and thus may be difficult to parse. However, cases in which there is little or no concordance between the observed and inferred geography may also prove the most interesting and informative.

The second caveat is that, in the current release of SpaceMix, each population is allowed only a single source of admixture. In many cases, modern day populations will be amalgams of many different historical groups, and the inability to model that history of multiple admixtures may produce results that are difficult to interpret (e.g., the inferred source of admixture falls in the middle of the locations of a population's true parental populations). Although it is statistically and computationally possible 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, but there may be empirical datasets in which such a modeling scheme is required to effectively map patterns of population structure.

Third, the landscape of allele frequencies on which the sources of admixture for the sampled populations are estimated is entirely informed by the placement of those modern samples. However, it will almost certainly be the case that the true history of population admixture occurred at different points in space (and deeper in time), and between populations that are ancestral to those in the modern sample. The increased sequencing of ancient DNA (skoglund, jakobsson, etc.) 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).

Future Directions

As noted in the third caveat above, the ancestral populations that have donated to the genetic makeup of modern day samples will not be among the samples in the current analysis, and therefore the inferred geography of admixture may only coarsely approximate the truth (depending on the complexity of the demographic history and the extent to which daughter populations of that ancestral admixture source both resemble their parents and have stayed put, geographically). The inclusion of ancient DNA samples in the analyzed sample offers an intriguing way forward in getting better representation of the ancestral popu-

lations from which the ancestors of modern samples received their admixture. However, it is also possible to model shared drift as a spatiotemporal Gaussian process, in which covariance in allele frequencies decays with distance both in space and in time. In this model, which we are currently developing, ancient DNA samples can 'calibrate' allele frequency landscapes at points in the past, and modern day samples may draw admixture from estimated coordinates in space-time.

Methods

As described above, we have developed an algorithm, called SpaceMix, that uses the standardized sample allele frequency covariance and the the likelihood equation in equation (13) to simultaneously estimate the posterior distribution 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, $\vec{\alpha}$. The inference procedure has two main components: (1) the derivation of the standardized sample allele frequency covariance matrix, and (2) a Markov chain Monte Carlo algorithm that samples parameter values from the posterior distribution. Here, we describe these procedures in detail.

Deriving the standardized sample covariance

We then assume that our sample allele frequencies at a locus across populations are given by

$$\hat{f}_\ell \sim MVN(\epsilon_\ell, \epsilon_\ell(1-\epsilon_\ell)\Omega) \tag{14}$$

with Ω given by equation (3), where the $\epsilon_\ell(1-\epsilon_\ell)\bar{S}^{-1}$ on the diagonal of the covariance matrix captures, to first approximation, the effects of sampling.

If we knew ϵ we can calculate an estimate of the covariance matrix (Ω) across loci as

$$\widehat{\Omega} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{(\hat{f}_{\ell_\ell} - \epsilon_\ell)(\hat{f}_\ell - \epsilon_\ell)^T}{\epsilon_\ell(1-\epsilon_\ell)}. \tag{15}$$

Then, if we define our standardized sample allele frequencies, X_ℓ, as

$$X_\ell = (\hat{f}_\ell - \epsilon_\ell)/\sqrt{\epsilon_\ell(1-\epsilon_\ell)}, \tag{16}$$

the expression given in equation (6) gives the sample covariance matrix of our standardized sample allele frequencies. Then $L\widehat{\Omega} = XX^T$ is Wishart distributed with degrees of freedom equal to the number of loci (L) across which the covariance is calculated.

However, we do not have the 'ancestral' frequency ϵ. Instead we mean-center and normalize our observations at a locus using the weighted mean sample frequency in place of ϵ. Recall that the sample allele frequency at locus ℓ in population k is given by $\hat{f}_{\ell,k} = C_{\ell,k}/S_{\ell,k}$. We wish to calculate a sample mean frequency at each locus weighted by the sample size in each population. As sample size may vary across loci, we first calculate $\bar{S}_k$, the mean sample size in population k, as $\bar{S}_k = \frac{1}{L}\sum_{\ell=1}^{L} S_{\ell,k}$. We then calculate the weighted sample mean frequency at locus ℓ

$$\bar{f}_\ell = \frac{1}{\sum_K S_{\ell,k}} \sum_K \hat{f}_{\ell,k} S_{\ell,k} \tag{17}$$

We call the mean standardized allele frequency in population k $X'_{\ell,k}$, and calculate them as follows:

$$\hat{X_{\ell,k}} = \frac{\hat{f}_{\ell,k} - \bar{f}_\ell}{\sqrt{\bar{f}_\ell(1-\bar{f}_\ell)}} \tag{18}$$

This is equivalent to taking our normalized vector of sample frequencies

$$Y_\ell = \frac{\hat{f}_{\ell,k}}{\sqrt{\bar{f}_\ell(1-\bar{f}_\ell)}} \tag{19}$$

and writing $X'_\ell = TY_\ell$ where T is the mean centering matrix, whose elements are given by

$$T_{ij} = \delta_{ij} - \frac{\bar{S}_j}{\sum_{k=1}^{K} \bar{S}_j} \tag{20}$$

Therefore, we assume that

$$\hat{X}_\ell \sim MVN(0, T^T \Omega T) \tag{21}$$

the mean centering acts to to reduce the covariance among populations in $\hat{X}_\ell$ compared to X_ℓ, and can induce negative covariance between distant populations (as they are often on either side of the mean). In addition the covariance matrix of our mean centered frequencies has rank $K-1$ rather than K, as any one of our mean centered frequencies could be written as a linear function of the others.

Model	No. Free Parameters	Parameters
stationary population locations, no admixture	$K+3$	$\alpha_0, \alpha_1, \alpha_2, \eta$
inferred population locations, no admixture	$2K+3$	$\alpha_0, \alpha_1, \alpha_2, \eta, G'$
stationary population locations, inferred admixture	$2K+3$	$\alpha_0, \alpha_1, \alpha_2, \eta, G^*, w$
inferred population locations, inferred admixture	$3K+3$	$\alpha_0, \alpha_1, \alpha_2, \eta, G', G^*, w$

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

This means that $T^T \Omega T$ is not invertible, and so can not be used directly in our Wishart likelihood. To circumvent this problem we take the QR decomposition of the matrix T, and form a projection matrix $\Psi = Q_{,-K}$ dropping the K^{th} column of Q. We then write our Wishart likelihood as

$$P(\widehat{\Omega} \mid \Omega) = \mathcal{W}\left(\Psi^T Y Y^T \Psi \mid \Psi^T \Omega \Psi, L\right). \tag{22}$$

this procedure avoids the reduced rank of our covariance matrices.

Markov chain Monte Carlo Inference Procedure

The inference algorithm described here may be used to estimate the parameters of any of four possible models: (1) population locations are stationary, and they do not draw any admixture; (2) populations may choose their own locations, but not admixture; (3) populations may draw admixture, but are themselves stationary; (4) populations may both choose their own locations and draw admixture. The free parameters in each of these models are given in Table 1.

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

We use a Bayesian MCMC approach to parameter inference, and specify priors on all parameters. A table of all parameters, their descriptions, and their priors is given in Table 2.

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

Table 2: List of parameters used in the SpaceMix models, along with their descriptions and priors.

Our inference algorithm proceeds is described in excruciating detail below.

We assume that the user has specified the following data:

- the allelic count data, C, from K population over L variant loci, where $C_{\ell,k}$ 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 $S_{\ell,k}$ gives the number of haplotypes sampled at locus ℓ in population k.
- the geographic sampling locations, G, from each of the K populations, where G_k gives the longitude and latitude of the sampled individual(s).

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

Initiating the MCMC We then calculate the standardized sample covariance matrix $\widehat{\Omega}'$ as outlined in the section "Deriving the standardized sample covariance" above. We calculate $\bar{S_k}$, the mean sample size across loci for population k, for each population as part of the standardization of the sample allele frequencies, and for use as $\bar{S_k}^{-1}$ as part of the spatial covariance function detailed in (3).

Armed with the standardized sample covariance, the geographic sampling locations, and the inverse mean sample sizes across samples ($\widehat{\Omega}'$, G, $\bar{S_k}^{-1}$), we may embark upon our analysis.

To initiate our MCMC, we specify starting values for each parameter. We draw initial values for α_0, α_1, α_2, η, 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 parameter $\Theta = \{\alpha_0, \alpha_1, \alpha_2, \eta, w, G', G^*\}$ 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 parameters, a single population, k is randomly selected and only its 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^{NEW}, 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 (11), Ω^*. We then calculate the likelihood of the standardized sample covariance $\widehat{\Omega}'$ following equation (22) for the current set of parameters $P(\widehat{\Omega}' \mid \Omega^*(\Theta))$ and our proposed update to our set of parameters $P(\widehat{\Omega}' \mid \Omega^*(\Theta^{NEW}))$. We then calculate the prior probabilities of both sets of parameter values, $P(\Theta^{NEW})$, $P(\Theta)$, following the priors given in Table 2.

We combine these together to give the Metropolis-Hastings ratio, $\underline{\mathrm{R}}$, the probability of accepting the proposed parameter values Θ^{NEW}:

$$R = \min\left(1, \frac{(\widehat{\Omega}' \mid \Omega^*(\Theta^{NEW}))}{P(\widehat{\Omega}' \mid \Omega^*(\Theta))} \frac{P(\Theta^{NEW})}{P(\Theta)}\right), \tag{23}$$

Note that all of our moves, described below, are symmetric so a Hastings ratio cancels through. If we accept our proposed move Θ is replaced by Θ^{NEW} and this is recorded, otherwise Θ^{NEW} is disgarded and we remain at Θ.

Udates 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 its own variance given by the scale of the tuning parameter for that parameter. For example $\alpha_0' \sim \alpha_0 + \delta$, where $delta \sim \mathcal{N}(0, \sigma_{\alpha_0})$ and σ_{α_0} is scale of the tuning parameter. For η and w, each of which consists of K parameters, each parameter receives its own idependent scale of the tuning parameter. If the proposed move takes the parameter outside the range of its prior, the move is rejected and we do not move in that iteration of the MCMC.

Updates for geographic co-ordinates G' and G^* . Updates to the location parameters, G' and G^*, are somewhat more complicated due to the curvature of the Earth (Eratosthenes, *personal observation*). Implementing updates via a symmetric normal density on estimated latitude and longitude directly, which 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 for a bearing and a distance traveled, and, given these two quantities and a starting position, calculate the latitude and longitude of the proposed location. For example, in an update to the location of population i, G_i', we propose a distance traveled $\Delta_{G'_i}$, where, e.g., $\Delta_{G'_i} \sim |\mathcal{N}(0, \sigma_{G'_i})|$, and a bearing, γ, where $\gamma \sim U(0, 2\pi)$. Then we use the following equations to calculate the latitude and longitude of the proposed location:

$$\begin{aligned} \text{proposed latitude} = \arcsin(&\sin(\text{current latitude}) \times \\ &\cos(\Delta) \times \cos(\text{current latitude}) \times \\ &\sin(\Delta) \times \cos(\gamma)) \end{aligned} \tag{24}$$

and

$$\begin{aligned}\text{proposed longitude} = &-\pi + (\text{current longitude} - \\ &\arctan(\sin(\gamma) \times \sin(\Delta) \times \cos(\text{current latitude}), \\ &\cos(\Delta) - \sin(\text{current latitude}) \times \\ &\sin(\text{proposed latitude})) + \pi) \mod (2\pi)\end{aligned} \tag{25}$$

where latitude and longitude are given in radians. As with η and w, the scales of the tuning parameters for the different populations and different location parameters (G' and G^*) are independent.

Adaptive Metropolis-within-Gibbs proposal mechanism . We use an adaptive Metropolis-within-Gibbs proposal mechanism on each parameter (Roberts and Rosenthal 2008, Rosenthal 2010). This mechanism attempts to maintain an acceptance proportion for each parameter as close as possible to 0.44 (optimal for one-dimensional proposal mechanisms; Roberts et al 1997, Roberts and Rosenthal 2001). We implement this mechanism by creating, for each variable i, an associated variable 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 nth set of 50 iterations, we adjust ζ_i by an "adaption amount," which is added to ζ_i if the acceptance proportion in the nth set of 50 iterations (${r^{(n)}}_i$ 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:

$${\zeta^{n+1}}_i = \begin{cases} {\zeta^n}_i + \min(\min(0.01, n^{-\frac{1}{2}}), 20), & \text{if } {r^{(n)}}_i > 0.44 \\ {\zeta^n}_i - \min(0.01, n^{-\frac{1}{2}}), & \text{if } {r^{(n)}}_i < 0.44 \end{cases} \tag{26}$$

We choose to cap the maximum adaption amount at 20 (which is, recall, 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 Figure 33, and in more efficient mixing and decreased autocorrelation time of parameter estimates.

Simulations

We ran our simulations using a coalescent framework in the program *ms* (Hudson). Briefly, we simulated populations on a lattice, with nearest neighbor (separated by a distance of 1) migration rate $m_{i,j}$, as well as migration on the diagonal of the unit square at rate $m_{i,j}/\sqrt{2}$. For each locus in the dataset, we used the **-s** option to specify a single segregating site, and then we simulated 10,000 loci independently, which were subsequently conglomerated into a single dataset for each scenario. For all simulations, except the "Populations on a line" scenario (Fig. 32), we sampled only every other population, and, from each population, we sampled 10 haplotypes (corresponding to 5 diploid individuals). In the "Populations on a line" scenario, we simulated no intervening populations, such that every population was sampled.

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

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 populations were allowed to choose their own locations (but no admixture). Population locations were initiated at the origin (i.e. - at iteration 1 of the MCMC, $G'_i = (0, 0)$), and all other parameters were drawn randomly from their priors at the start of each chain.

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

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

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

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

References

Miguel Alcaide, Elizabeth S. C. Scordato, Trevor D. Price, and Darren E. Irwin. Genomic divergence in a ring species complex. Nature, advance online publication, May 2014. ISSN 14764687. URL http://dx.doi.org/10.1038/nature13285.

Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS Genet, 5(10):e1000695, 10 2009. doi: 10.1371/journal.pgen.1000695. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1000695.

Supplementary Materials

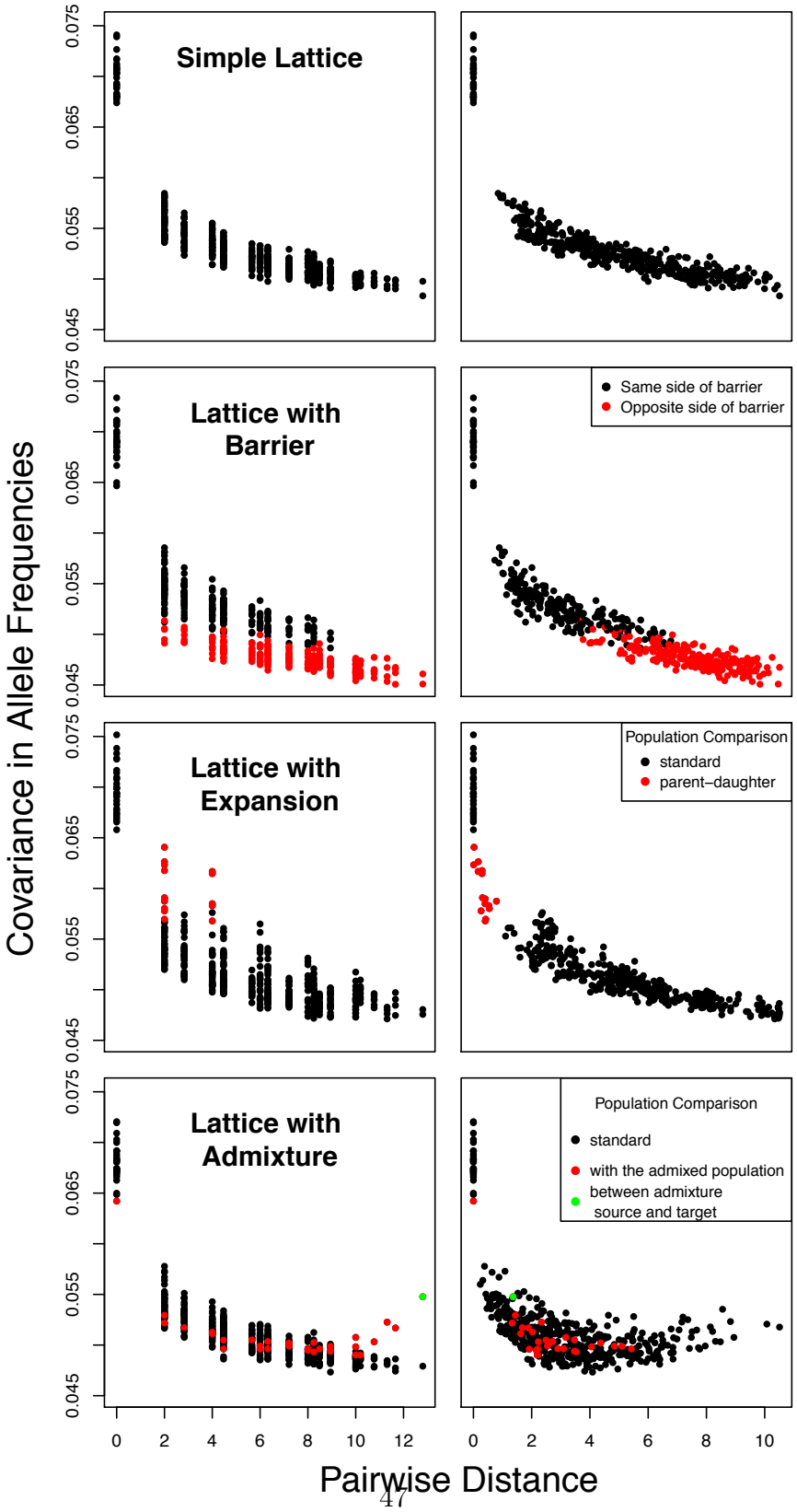


Figure 13: Decays in covariance for four different simulation scenarios (from top to bottom: simple lattice; lattice with barrier; lattice with expansion; lattice with admixture). Left column: sample covariance plotted against observed pairwise distance. Right column: sample covariance plotted against inferred geogenetic distance.

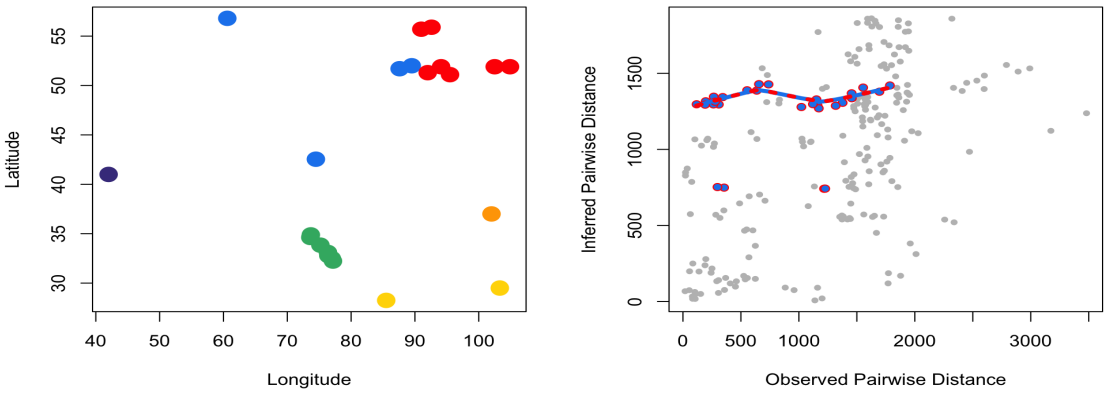


Figure 14: Comparing observed to estimated pairwise distance between warbler populations: a) observed population coordinates; b) observed pairwise distance between populations compared to that between their inferred locations. The highlighted points show distances between populations from the *plumbeitarsus* and *viridanus* subspecies. Notice that, regardless of their observed distance, their inferred separations are roughly constant, and much larger than their observed distance.

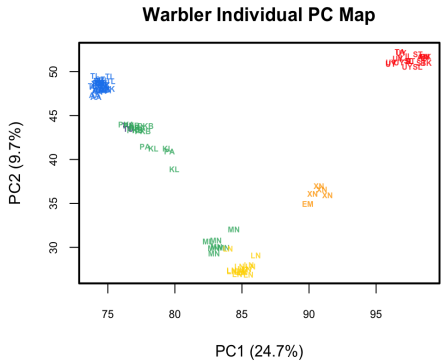
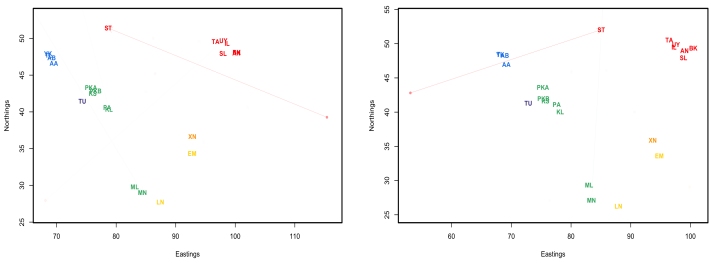
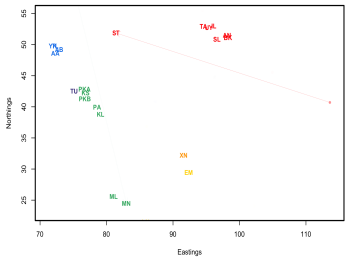


Figure 15: The map of warbler individuals derived from a Principal Components analysis.



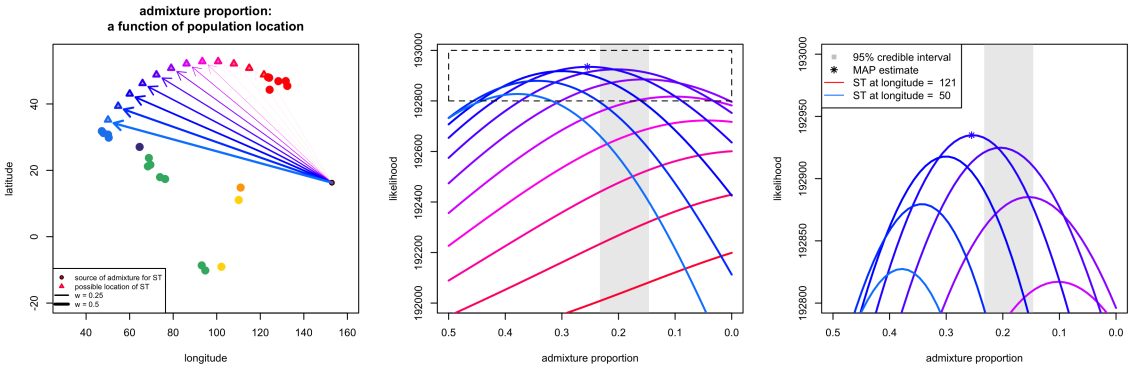
(a)

(b)

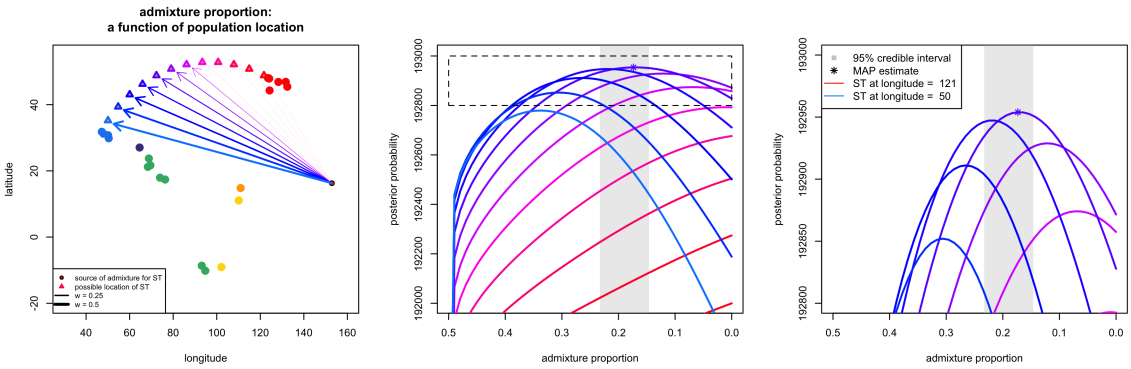


(c)

Figure 16: 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.



(a) Likelihood



(b) Posterior probability

Figure 17: Likelihood surfaces for different placements of population ST between *plumbeitarsus* and *viridanus* clusters: (a) log likelihood surface; (b) posterior probability surface, incorporating the priors.

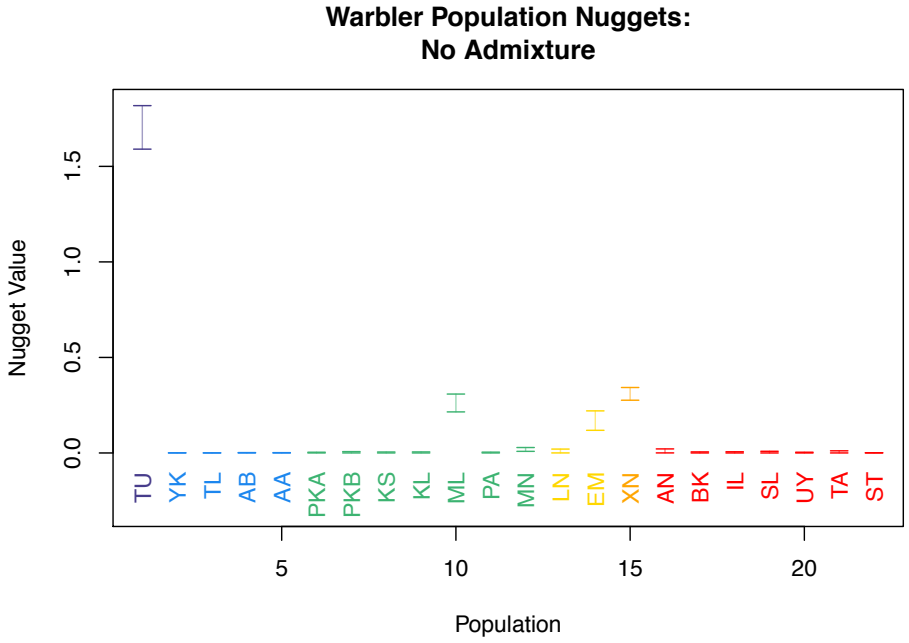


Figure 18: Credible intervals on estimated warbler population nugget parameters in an analysis without admixture.

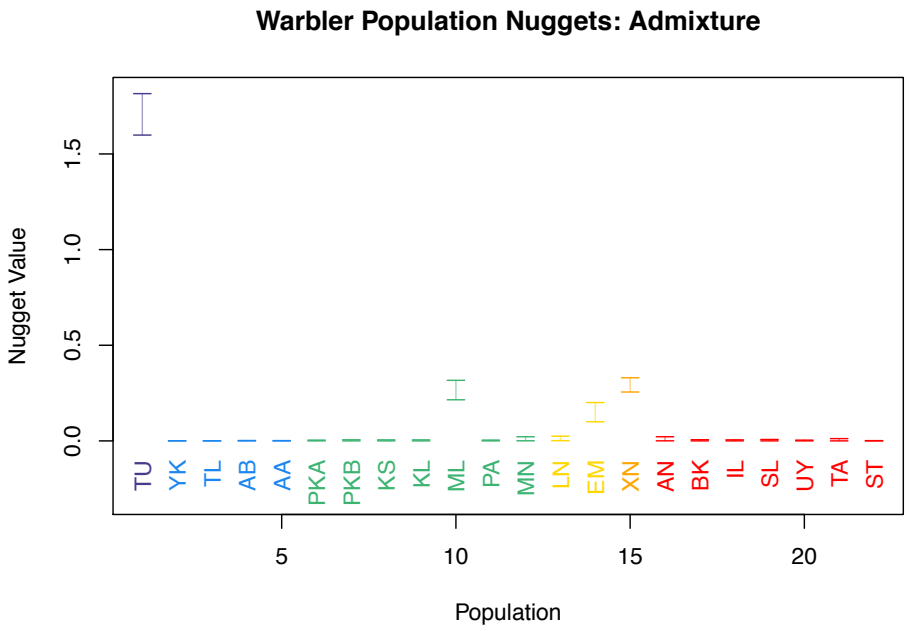


Figure 19: Credible intervals on estimated warbler population nugget parameters in an analysis with admixture.

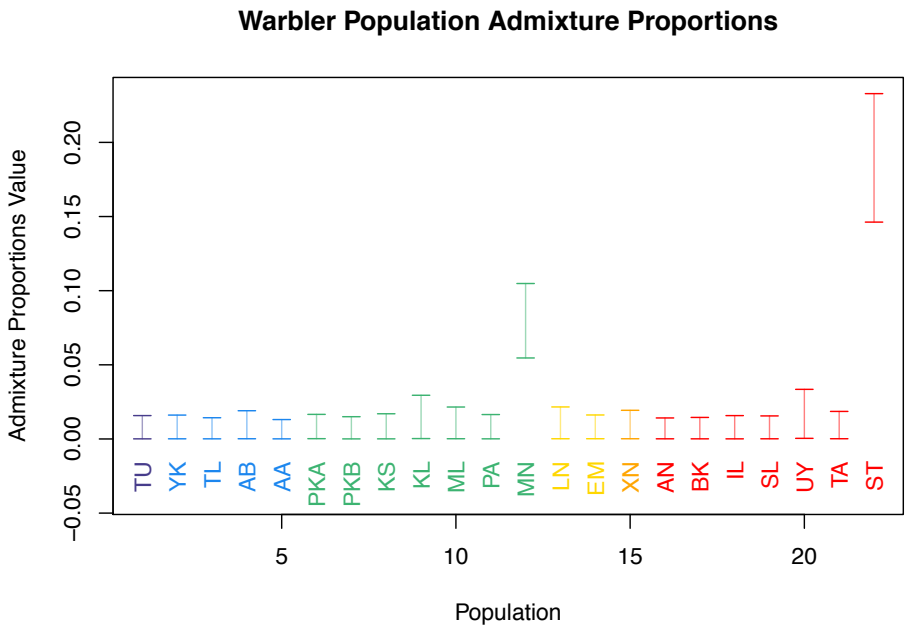
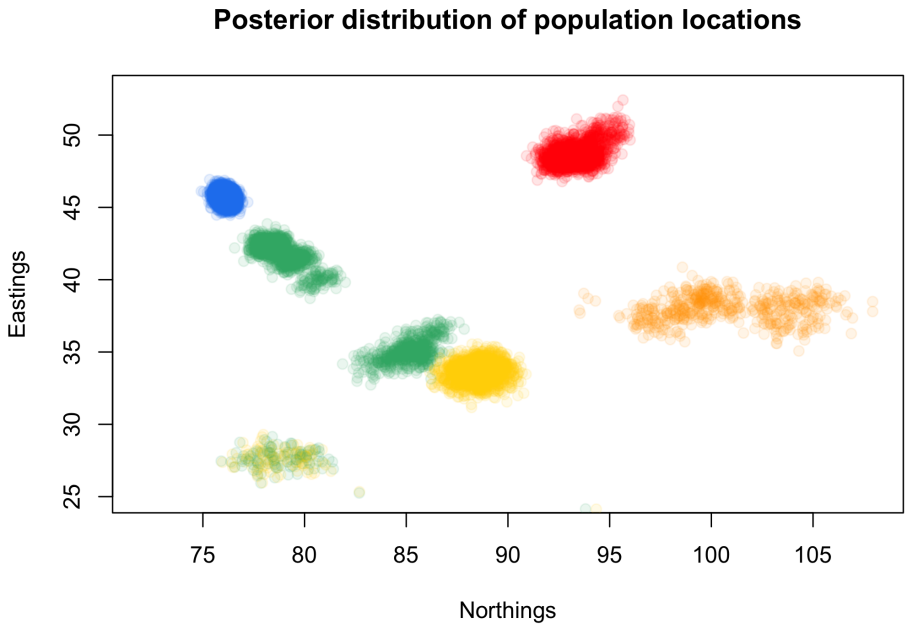
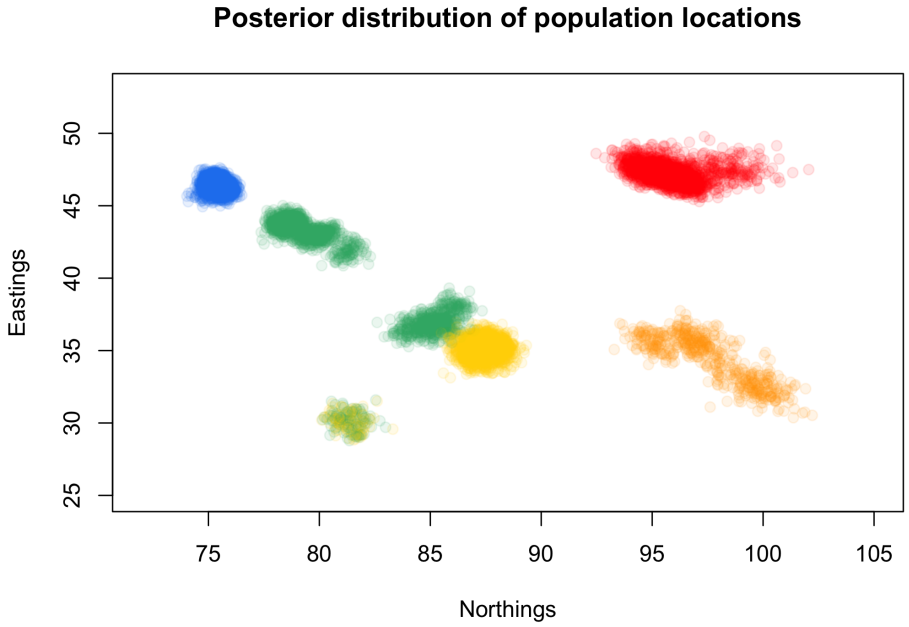


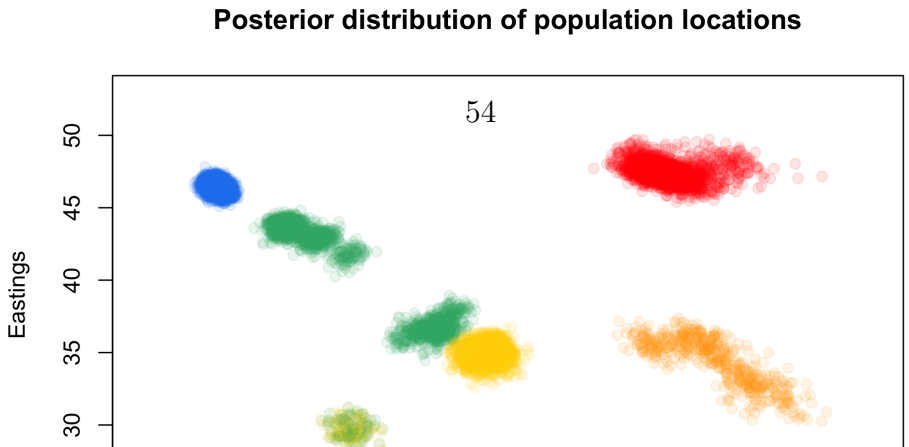
Figure 20: Credible intervals on estimated warbler population admixture proportion parameters.



(a) random location prior



(b) random location prior



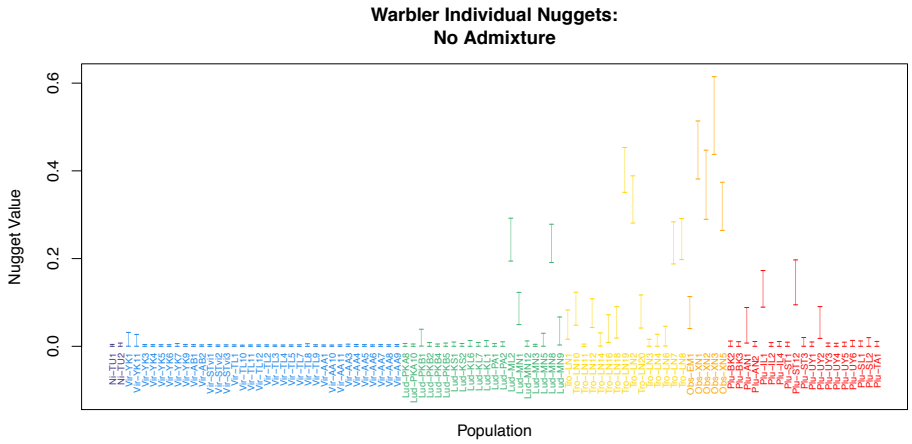


Figure 22: Credible intervals on estimated warbler individual nugget parameters in an analysis without admixture.

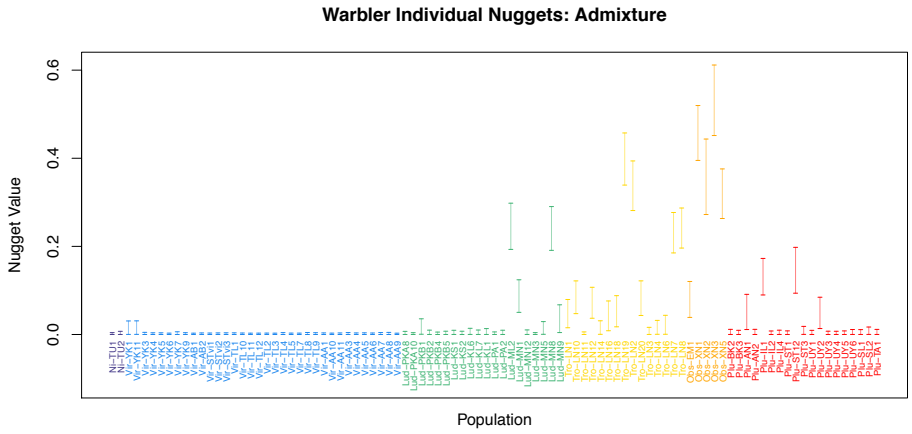


Figure 23: Credible intervals on estimated warbler individual nugget parameters in an analysis with admixture.

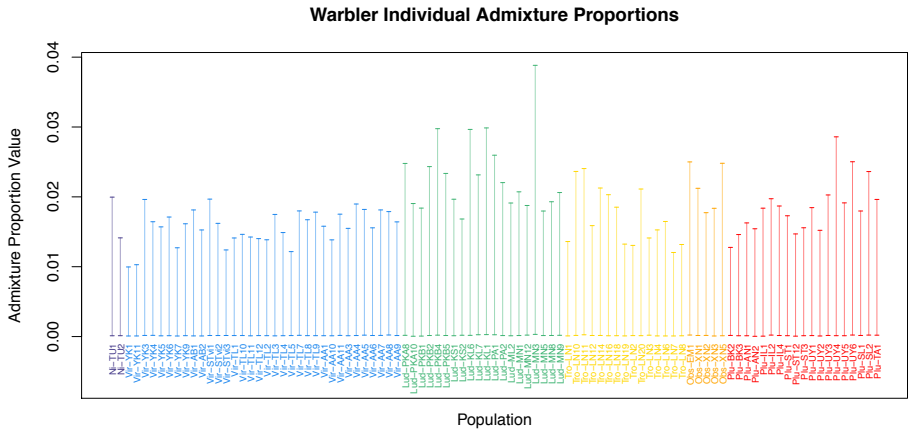
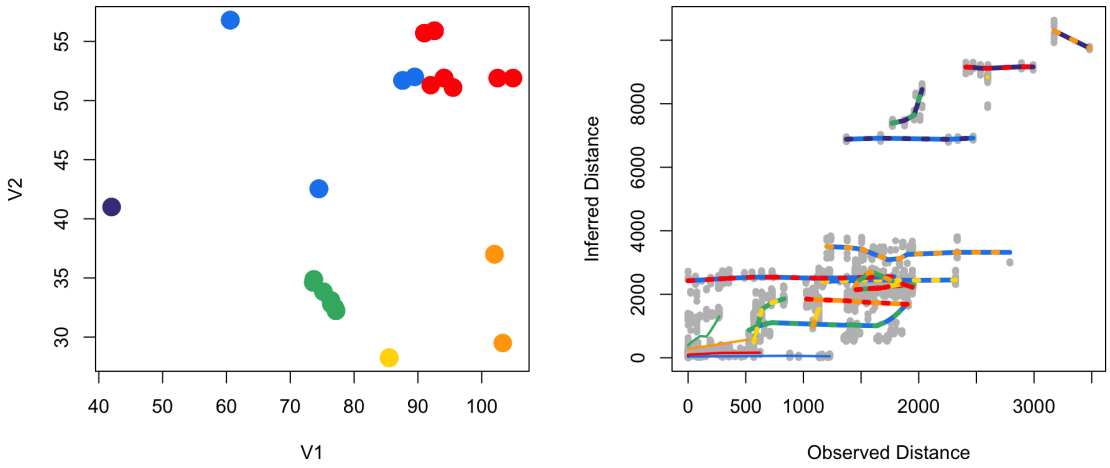
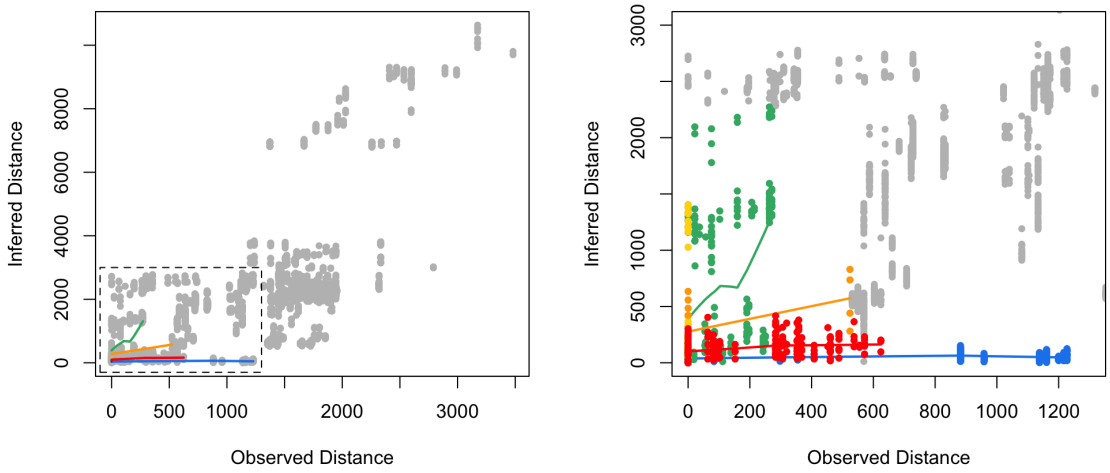


Figure 24: Credible intervals on estimated warbler individual admixture proportion parameters.



(a) All population pairs



(b) Just within population comparisons

Figure 25: Comparing observed to estimated pairwise distance between warbler individuals, (a) between and (b) within subspecies populations.

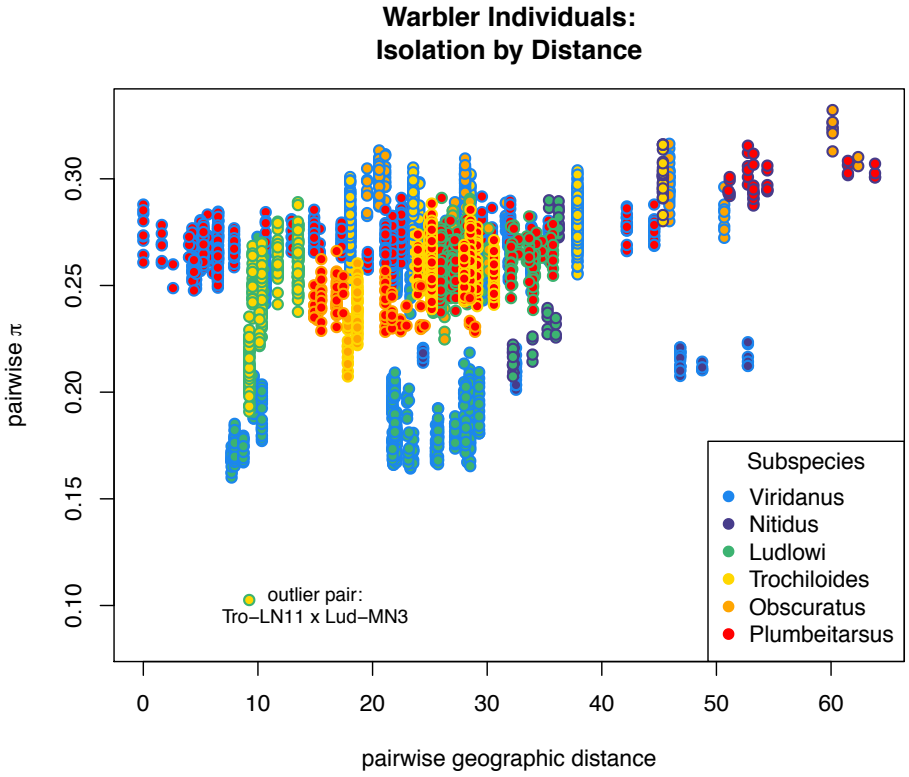
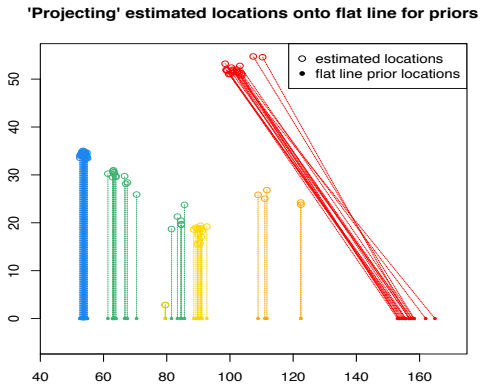
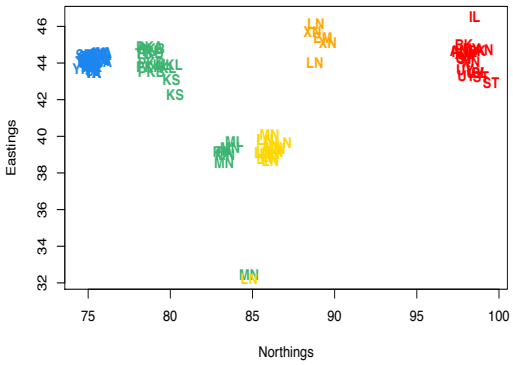


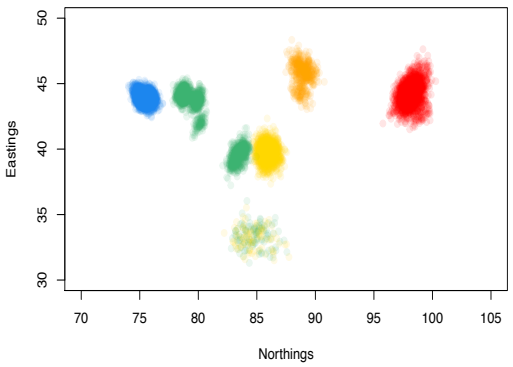
Figure 26: Pairwise π at polymorphic sites calculated between all pairs of individuals from different subspecies, and colored by the subspecies to which each individual in the comparison is drawn. Note that individuals Tro-LN11 and Lud-MN3 have sequence divergence that is unusually low relative to that of other comparisons between individuals from the same two subspecies.



(a) warbler individuals projected onto a line



(b) MAP estimate from SpaceMix



(c) Posterior distribution of locations from SpaceMix

Figure 27: SpaceMix analyses performed on a warbler individuals for which prior locations are projected onto a flat line in an order that approximately corresponds to their order around the ring: a) the setup of the analysis, in which warbler individuals are projected from their estimated locations in a previous SpaceMix run onto a flat line; b) the MAP estimate of individual locations from a SpaceMix analysis with no admixture, using the flat line locations from (a) as priors on G'; c) posterior distribution of individual locations from the same analysis. As with the other inferred maps, here, for clarity, the inferred locations have been rotated via a Procrustes superimposition around their true sampling coordinates.

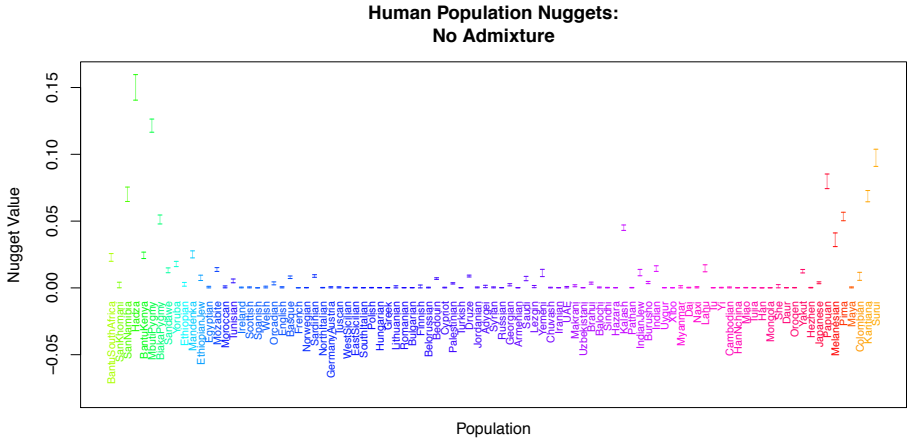


Figure 28: Credible intervals on estimated human sample nugget parameters in an analysis without admixture.

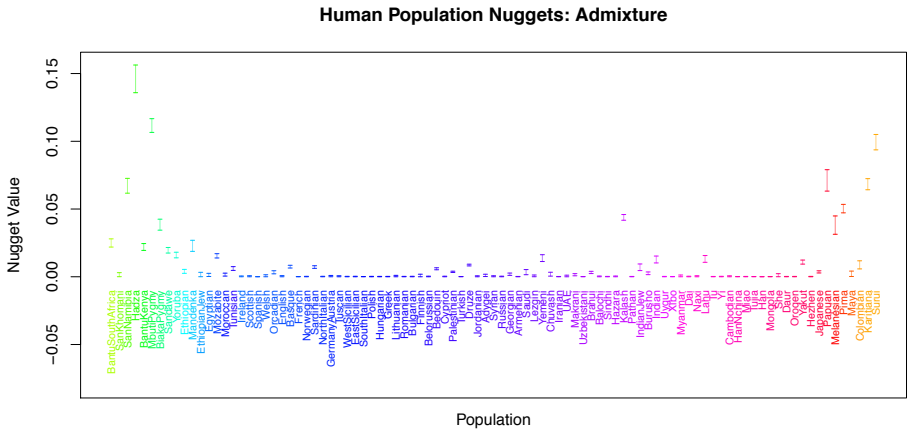


Figure 29: Credible intervals on estimated human sample nugget parameters in an analysis with admixture.

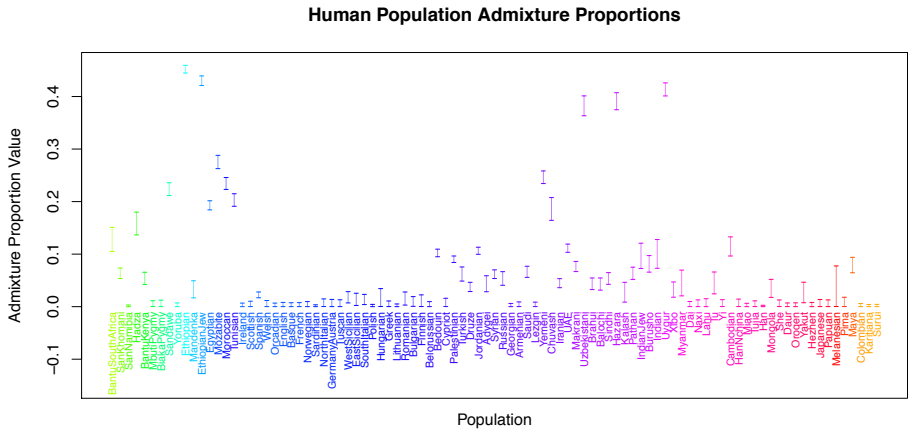


Figure 30: Credible intervals on estimated human sample admixture proportion parameters.

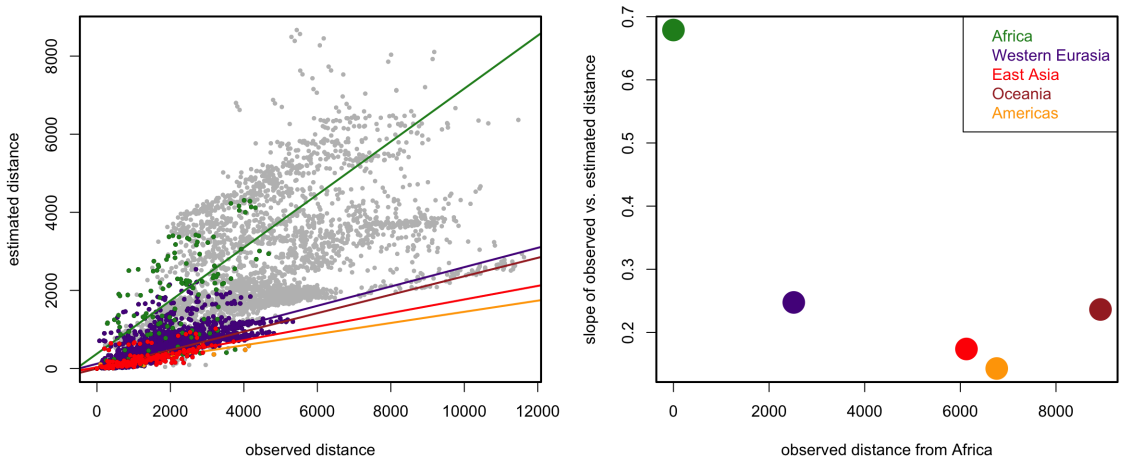
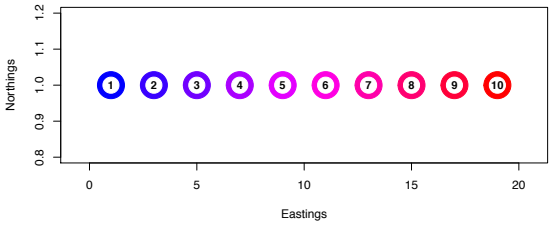
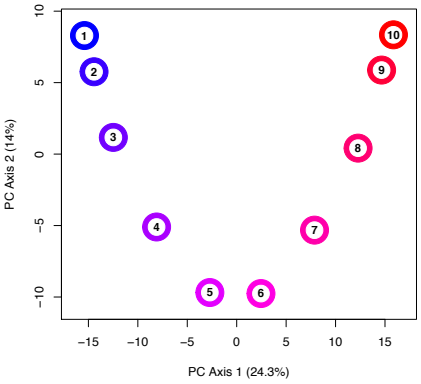


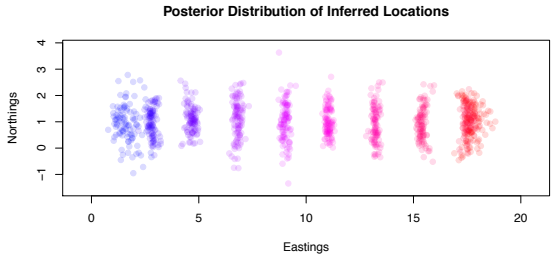
Figure 31: Comparison of observed distance to estimated distance between human populations, colored by continent from which populations were sampled (i.e. - two populations sampled from Africa are green). Eurasia is divided into Western Eurasia and East Asia.



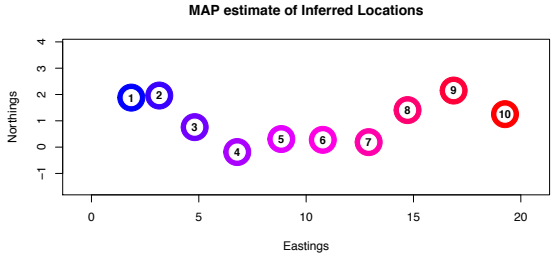
(a) populations on a line



(b) PCA map of line scenario



(c) posterior from SpaceMix map of line scenario



(d) MAP estimate from SpaceMix map of line scenario

Figure 32: 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.

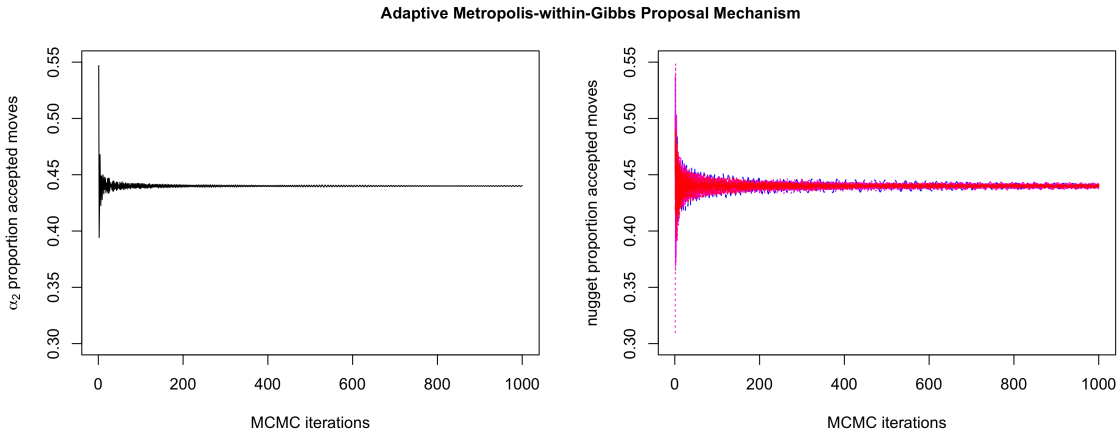


Figure 33: Example parameter acceptance proportions for the α_2 parameter and the nugget parameter, η, using the adaptive Metropolis-within-Gibbs proposal mechanism.

Sample	Subspecies	Longitude	Latitude
Vir-YK1	Viridanus	60.60	60.60
Vir-YK11	Viridanus	60.60	60.60
Vir-YK3	Viridanus	60.60	60.60
Vir-YK4	Viridanus	60.60	60.60
Vir-YK5	Viridanus	60.60	60.60
Vir-YK6	Viridanus	60.60	60.60
Vir-YK7	Viridanus	60.60	60.60
Vir-YK9	Viridanus	60.60	60.60
Vir-AB1	Viridanus	89.50	89.50
Vir-AB2	Viridanus	89.50	89.50
Vir-STvi1	Viridanus	92.60	92.60
Vir-STvi2	Viridanus	92.60	92.60
Vir-STvi3	Viridanus	92.60	92.60
Vir-TL1	Viridanus	87.60	87.60
Vir-TL10	Viridanus	87.60	87.60
Vir-TL11	Viridanus	87.60	87.60
Vir-TL12	Viridanus	87.60	87.60
Vir-TL2	Viridanus	87.60	87.60
Vir-TL3	Viridanus	87.60	87.60
Vir-TL4	Viridanus	87.60	87.60
Vir-TL5	Viridanus	87.60	87.60
Vir-TL7	Viridanus	87.60	87.60
Vir-TL8	Viridanus	87.60	87.60
Vir-TL9	Viridanus	87.60	87.60
Vir-AA1	Viridanus	74.48	74.48
Vir-AA10	Viridanus	74.48	74.48
Vir-AA11	Viridanus	74.48	74.48
Vir-AA3	Viridanus	74.48	74.48
Vir-AA4	Viridanus	74.48	74.48
Vir-AA5	Viridanus	74.48	74.48
Vir-AA6	Viridanus	74.48	74.48
Vir-AA7	Viridanus	74.48	74.48
Vir-AA8	Viridanus	74.48	74.48
Vir-AA9	Viridanus	74.48	74.48
Ni-TU1	Nitidus	42.00	42.00
Ni-TU2	Nitidus	42.00	42.00
Lud-PKA8	Ludlowi	73.69	73.69
Lud-PKA10	Ludlowi	73.69	73.69
Lud-PKB1	Ludlowi	73.61	73.61
Lud-PKB2	Ludlowi	73.61	73.61
Lud-PKB4	Ludlowi	73.61	73.61
Lud-PKB5	Ludlowi	73.61	73.61
Lud-KS1	Ludlowi	75.19	75.19
Lud-KS2	Ludlowi	75.19	75.19
Lud-KL6	Ludlowi	76.37	76.37
Lud-KL7	Ludlowi	76.37	76.37
Lud-KL1	Ludlowi	76.37	76.37
Lud-PA1	Ludlowi	76.97	76.97
Lud-PA2	Ludlowi	76.97	76.97
Lud-ML2	Ludlowi	76.43	76.43
Lud-MN1	Ludlowi	77.16	77.16
Lud-MN12	Ludlowi	77.16	77.16
Lud-MN3	Ludlowi	77.16	77.16
Lud-MN5	Ludlowi	77.16	77.16
Lud-MN8	Ludlowi	77.16	77.16
Lud-MN9	Ludlowi	77.16	77.16
Tro-LN1	Trochiloides	85.50	85.50
Tro-LN10	Trochiloides	85.50	85.50
Tro-LN11	Trochiloides	85.50	85.50
Tro-LN12	Trochiloides	85.50	85.50
Tro-LN14	Trochiloides	85.50	85.50
Tro-LN16	Trochiloides	85.50	85.50
Tro-LN18	Trochiloides	85.50	85.50
Tro-LN19	Trochiloides	85.50	85.50
Tro-LN2	Trochiloides	85.50	85.50
Tro-LN20	Trochiloides	85.50	85.50
Tro-LN3	Trochiloides	85.50	85.50
Tro-LN4	Trochiloides	85.50	85.50
Tro-LN6	Trochiloides	85.50	85.50
Tro-LN7	Trochiloides	85.50	85.50
Tro-LN8	Trochiloides	85.50	85.50
Obs-EM1	Obscuratus	103.30	103.30
Obs-XN1	Obscuratus	102.00	102.00
Obs-XN2	Obscuratus	102.00	102.00
Obs-XN3	Obscuratus	102.00	102.00
Obs-XN5	Obscuratus	102.00	102.00
Plu-BK2	Plumbeitarsus	104.90	104.90
Plu-BK3	Plumbeitarsus	104.90	104.90
Plu-AN1	Plumbeitarsus	102.50	102.50
Plu-AN2	Plumbeitarsus	102.50	102.50
Plu-IL1	Plumbeitarsus	95.50	95.50
Plu-IL2	Plumbeitarsus	95.50	95.50
Plu-IL4	Plumbeitarsus	95.50	95.50
Plu-ST1	Plumbeitarsus	92.60	92.60
Plu-ST12	Plumbeitarsus	92.60	92.60
Plu-ST3	Plumbeitarsus	92.60	92.60
Plu-UY1	Plumbeitarsus	94.10	94.10
Plu-UY2	Plumbeitarsus	94.10	94.10
Plu-UY3	Plumbeitarsus	94.10	94.10
Plu-UY4	Plumbeitarsus	94.10	94.10
Plu-UY5	Plumbeitarsus	94.10	94.10
Plu-UY6	Plumbeitarsus	94.10	94.10
Plu-SL1	Plumbeitarsus	91.00	91.00
Plu-SL2	Plumbeitarsus	91.00	91.00
Plu-TA1	Plumbeitarsus	92.00	92.00

Table 3: Subspecies and geographic meta-data for greenish warbler individuals included in analysis

	Population	Longitude	Latitude	Mean Sample Size
1	BantuSouthAfrica	28.00	-26.00	15.99
2	SanKhomani	18.10	-24.60	59.96
3	SanNamibia	20.00	-21.50	9.99
4	Hadza	33.10	-4.50	5.93
5	BantuKenya	37.00	-3.00	21.99
6	MbutiPygmy	29.00	1.00	25.98
7	BiakaPygmy	17.00	4.00	41.97
8	Sandawe	35.70	6.20	55.94
9	Yoruba	5.00	8.00	41.98
10	Ethiopian	38.70	9.00	37.70
11	Mandenka	-12.00	12.00	43.98
12	EthiopianJew	38.70	14.10	22.00
13	Egyptian	26.80	30.80	24.00
14	Mozabite	3.00	32.00	57.98
15	Moroccan	-5.50	33.60	49.97
16	Tunisian	9.80	35.60	24.00
17	Ireland	-8.20	53.40	14.00
18	Scottish	-4.20	56.50	12.00
19	Spanish	-3.70	40.50	67.95
20	Welsh	-3.70	52.60	8.00
21	Orcadian	-3.00	59.00	29.99
22	English	-0.80	52.00	12.00
23	Basque	0.00	43.00	47.99
24	French	2.00	46.00	55.97
25	Norwegian	8.50	60.50	35.99
26	Sardinian	9.00	40.00	55.98
27	NorthItalian	9.70	45.70	23.99
28	GermanyAustria	10.50	51.20	8.00
29	Tuscan	11.00	43.00	16.00
30	WestSicilian	12.50	38.00	20.00
31	EastSicilian	16.10	37.00	20.00
32	SouthItalian	16.90	39.50	35.96
33	Polish	19.10	51.90	31.99
34	Hungarian	19.50	47.20	40.00
35	Greek	21.80	39.10	39.99
36	Lithuanian	23.90	55.20	20.00
37	Romanian	25.00	45.90	28.00
38	Bulgarian	25.50	42.70	35.99
39	Finnish	25.70	61.90	4.00
40	Belorussian	28.00	53.70	16.00
41	Bedouin	33.50	31.00	89.98
42	Cypriot	33.50	35.50	24.00
43	Palestinian	35.00	33.50	91.95
44	Turkish	35.20	39.00	34.00
45	Druze	37.00	32.00	83.96
46	Jordanian	37.00	30.00	40.00
47	Adygei	39.00	44.00	33.99
48	Syrian	39.00	34.80	32.00
49	Russian	40.00	61.00	49.98
50	Georgian	44.60	41.80	39.99
51	Armenian	45.00	40.10	31.99
52	Saudi	45.10	23.90	20.00
53	Lezgin	47.50	43.00	35.96
54	Yemeni	48.50	15.60	13.99
55	Chuvash	50.20	53.20	34.00
56	Iranian	53.70	32.40	39.99
57	UAE	54.40	24.50	27.98
58	Makrani	64.00	26.00	49.99
59	Uzbekistani	64.60	41.40	29.99
60	Brahui	65.00	29.00	49.98
61	Balochi	67.00	31.00	47.99
62	Sindhi	69.00	25.00	47.99
63	Hazara	69.50	33.00	43.98
64	Kalash	71.00	36.00	45.99
65	Pathan	72.50	34.00	43.99
66	IndianJew	72.90	19.00	16.00
67	Burusho	74.00	37.00	49.98
68	Indian	77.60	13.00	25.97
69	Uygur	81.00	44.00	20.00
70	Xibo	81.00	43.00	17.99
71	Myanmar	96.00	21.90	5.99
72	Dai	99.00	21.00	19.98
73	Naxi	100.00	26.00	15.99
74	Lahu	101.00	22.00	16.00
75	Tu	101.00	36.00	20.00
76	Yi	103.00	28.00	20.00
77	Cambodian	105.00	12.00	19.98
78	HanNchina	108.00	39.00	20.00
79	Miao	108.00	28.00	19.99
80	Tujia	110.00	29.00	20.00
81	Han	114.00	26.00	67.96
82	Mongola	119.00	48.00	20.00
83	She	119.00	27.00	19.99
84	Daur	124.00	49.00	17.99
85	Oroqen	126.00	50.00	18.00
86	Yakut	129.00	63.00	49.98
87	Hezhen	133.00	47.00	16.00
88	Japanese	138.00	38.00	55.97
89	Papuan	143.00	-4.00	33.97
90	Melanesian	155.00	-6.00	19.99
91	Pima	-108.00	29.00	27.99
92	Maya	-91.00	19.00	41.97
93	Colombian	-68.00	3.00	13.99
94	Karitiana	-63.00	-10.00	27.99
95	Surui	-62.00	-11.00	16.00

Table 4: sample size and geographic meta-data for human samples included in analysis