A Novel Spatial Framework for Understanding Genetic Admixture

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

The patterns of genetic variation observed in modern populations are the product of a complex demographic and evolutionary history. Genetic data can be used to illuminate that history, providing information about when and how populations have diverged, how migration connects populations, and how population sizes have fluctuated over time. Work in this area has largely focused on estimating a population phylogeny, in which shared branch length on a tree represents shared evolutionary history between a pair of populations sampled in the modern day. However, patterns of population differentiation are rarely tree-like, as migration and colonization will continuously re-shape patterns of relatedness between populations. Isolation by distance (IBD), in which population differentiation increases with the distance between them, 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.

Introduction

Population-level demographic processes leave their mark on patterns of genetic variation and differentiation within and between populations. Recent population genetics methods have focused on ways of learning about those processes from genetic observations, including demographic inference from the allele frequency spectrum (Song) and inference of the ancestral recombination graph from orthologous sequence data (Song, Rasmussen&Siepel).

Further work has focused on joint inference and visualization of whole-genome patterns of relatedness, specifically on estimating a population phylogeny describing the evolutionary history shared between sampled individuals or populations (e.g. Pickrell & Pritchard, Patterson & Moorjani, Reich?). The approach of modeling relatedness between populations within species as a tree-like structure was pioneered by Cavalli-Sforza (&Thompson?) in the 1960s *?*, when genetic character data (blood types) were first becoming available in humans.

Their work has been greatly extended by Pickrell and Pritchard (2012), whose method, TreeMix, models covariance in allele frequencies across loci as a directed acyclic graph between populations. TreeMix then accommodates reticulate population structure by allowing branches on the population tree tree to be connected by arrows of admixture that explain excess residual covariance between populations or population 'clades.'

In parallel, Patterson and Reich and Moorjani and Other Folks have developed a suite of tests for genetic admixture between a set of populations assuming a tree-like structure to their evolutionary history. These tests model shared genetic drift as shared branch length on a hypothesized population tree, and take covariance in excess of that stipulated by the tree to be evidence for admixture.

These tree-based methods are both valuable as genetic inference and visualization tools. However, a more natural framework for modeling genetic differentiation between sampled populations or individuals may be that of isolation by distance (IBD) (Wright). The pattern of IBD, in which genetic differentiation increases with the geographic distance between populations, is ubiquitous in nature (Meirmans 2012, Ralph & Coop 2012), and relies only on the weak assumption that individual's mating opportunities are geographically limited by dispersal.

transition section on visualization of population structure and patterns of relatedness

There is a rich history of this type of data visualization, especially using reduced dimensionality representations of the data such as Principal Components Analysis (e.g. Menozzi, Piazza, and Cavalli-Sforza 1978, Novembre et al 2008, Novembre and Stephens 2008). Our method offers an improvement over PCA-based visualization methods in that it offers an explicit model of spatial genetic variation, and can therefore be used as a robust framework for inference and hypothesis testing (see also Yang et al 2012, Yang et al 2014).

Here, we present an analytical framework and develop an inference algorithm for studying the spatial distribution of genetic variation based on a model of isolation by distance, and we demonstrate the utility of this approach with two empirical applications.

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 the method can also be applied to individuals. We summarize these genetic data as a set of allelic count data and sample size data, 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. 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:

$$\epsilon_B \sim N(\epsilon_A, \delta_B(\epsilon_A)(1-\epsilon_A)) \tag{1}$$

where the parameter δ_B is the amount of 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 sufficiently limited (short time-scales or large population sizes).

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 deviation around the ancestral (or global) mean. We model the population frequencies at a locus 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, which may be straightforwardly modeled as a parametric function of any pairwise distance variable (see, e.g., 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 between distant ones, so that a population has higher covariance with a close neighbor than with a more distant population. We choose as the form of our parametric covariance matrix a simple and flexible model, with an exponential decay of allele frequency covariance with geographic distance (Wasser et al 2004, Bradburd, Ralph, and Coop 2013). *The covariance between population allele frequencies* in populations i and j is

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

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 combined with binomial sampling to model the sample frequencies (**??**), here we chose to treat these standardized frequencies $\hat{f}_\ell$, 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.

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

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, 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

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

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

Given the assumption of multivariate normality for our sample frequencies, it follows that the sample covariance of our standardized sample frequencies calculated across loci ($\widehat{\Omega} = ff^T$) is Wishart distributed with degrees of freedom equal to the number of loci (L) across which the covariance is calculated. That is,

$$\widehat{\Omega} \sim \mathcal{W}\left(L^{-1}\Omega, L\right) \tag{5}$$

should the L^{-1} be there? I'm pretty sure it goes there, b/c the mean of the Wishart is $n\boldsymbol{V}$, where n is the degrees of freedom and $\boldsymbol{V}$ is the scale matrix. This seems fairly important, so maybe P *or* G *could take a look at the wiki page and confirm this?*

this section was repeated below so I took this one out and just kept it below

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

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

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

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

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 ($\widehat{\Omega} = XX^T$) is Wishart distributed with degrees of freedom equal to the number of loci (L) across which the covariance is calculated. That is,

$$P(\widehat{\Omega} \mid \Omega) = \mathcal{W}\left(\widehat{\Omega} \mid L^{-1}\Omega, L\right) \tag{8}$$

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

We can estimate the parameters of our simple isolation by distance model, $\alpha_0, \alpha_1, \alpha_2$, by treating Eqn. (8) as the likelihood of our 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: $\alpha_0 \sim Exp(0.01)$; $\alpha_1 \sim Exp(1)$; $\alpha_0 \sim U(0.1, 2)$, as well as independent priors on the nugget in each population: $\eta_k \sim Exp(1)$.* For more details on our inference procedure, please see Appendix NNN. *Have some link to an appendix describing what to do with linked loci?*

Accommodating non-equilibrium processes This 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. Barriers to dispersal on the landscape can also change the relationship between allele frequency covariance and distance.

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. 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 a random variable, 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) 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 $\Omega(\vec{\alpha}, D(G'))$. We write the posterior probability as

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

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. For the prior on population k's location parameter, G'_k, we use 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, $D(G)$:* $G'_k \sim \mathcal{N}(\mu = G_k, \sigma = \frac{1}{2}\bar{D}(G))$. We perform MCMC on all of the random variables; for more details on our Bayesian inference procedure see Appendix B. This method is implemented in a program called SpaceMix.

The sampled locations are a natural prior on where the populations are positioned under the null of isolation by distance (as well as a natural starting position for the MCMC). 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 this concern, 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.

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 Appendix NNN). 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.

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. 1c), 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 1c, 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. 1d), 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 (referring to Fig. 1a, 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. 1d), the daughter populations of the expansion event cluster with their parent populations, reflecting the higher relatedness (per unit geographic separation) between them. *In Supp. Figs XXX. 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. 1a 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 analysis in which the locations of these populations were estimated is shown in Figure 2. 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 2, 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 lineages. However, this warping of space is difficult to interpret, especially 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.

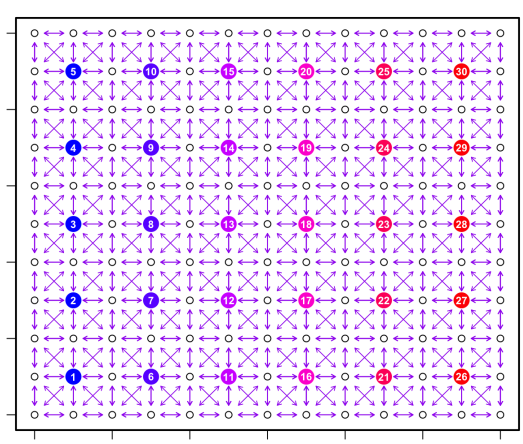
Inference of Spatial Admixture

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 a proportion p_k of its ancestry comes from another location $G_k^{(A)}$, 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 and those at the coordinates of the source from which the observed population draws admixture:

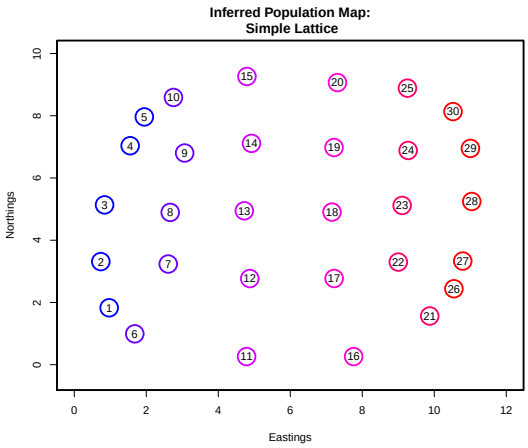
$$p_k f_{\ell,k} + (1-p) f_{\ell,k*} \tag{10}$$

where $f_{\ell,k}$ are the model-estimated allele frequencies at locus ℓ at the spatial location of population G_k and $f_{\ell,k*}$ are the model-estimated allele frequencies at the spatial location of the source of admixture $G_k^{(A)}$. We can allow each of our populations to have this setup, each with an independent spatial source of admixture.

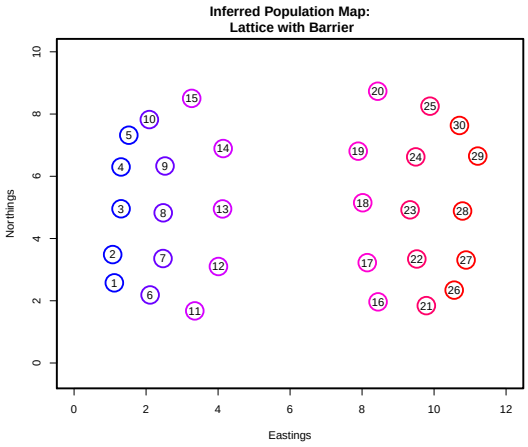
Following from the form of eqn. (10) the covariance between the standardized



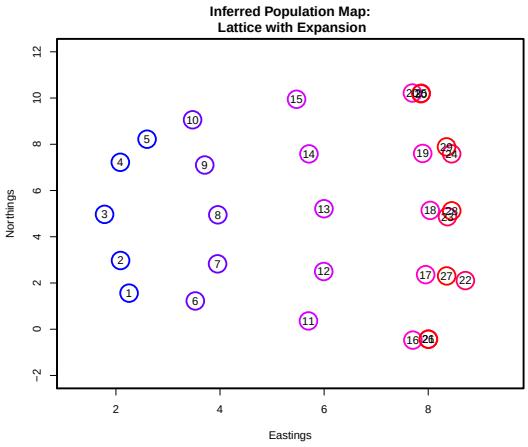
(a) Simulated Lattice



(b) Lattice



(c) Barrier



(d) Expansion

Figure 1: Population maps inferred using SpaceMix under three different scenarios: a) configuration of simulated populations; (c) simple lattice at equilibrium; c) a lattice with a barrier across the center line of longitude; d) a lattice with recent expansion on the eastern margin.

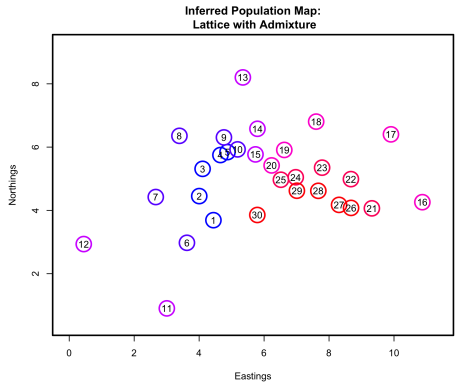


Figure 2: Inference of population locations in the scenario depicted in Figure **??**. Population 30 has received half of its lineages from population 1, to simulate a long distance admixture event in the very recent past.

allele frequencies of population i and j can be modeled as

$$\begin{aligned}\Omega_{i,j}^{(A)} = (1-p_i)(1-p_j)\Omega_{i\,,\,j} \;\times \\ (p_i)(1-p_j)\Omega_{i^{(A)},\,j} \;\times \\ (p_j)(1-p_i)\Omega_{i\,,\,j^{(A)}} \;\times \\ (p_i)(p_j)\Omega_{i^{(A)},\,j^{(A)}} \;+ \\ \delta_{i,j}(\eta_i + \bar{S}^{-1})\end{aligned} \tag{11}$$

where $i^{(A)}$ and $j^{(A)}$ 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, Ω, is parameterized by the pairwise geographic distances between each pair of populations, *without* the population-specific variance terms on the diagonal (i.e., in Eqn. (11), i.e. $\Omega_{i,j} = \frac{1}{\alpha_0}\exp\left((\alpha_1 D_{i,j})^{\alpha_2}\right)$). Note that we then reintroduce the nugget, η_i, and the sample size effect, $\bar{S}^{-1}$, for each population, to model drift or excess variance in population i on top of that predicted by the mixture of frequencies predicted by our spatial model. As in Eqn. (3), $\delta_{i,j}$ is the indicator function 1 if $i = j$, and 0 otherwise.

The admixed covariance between populations i and j, $\Omega_{i,j}^{(A)}$ is then a function of all the pairwise spatial covariances between populations i and j and the points from which they draw admixture, $i^{(A)}$ and $j^{(A)}$. Those spatial covariances in turn are a function of all combinations of pairwise distances between their locations: G_i, G_i, $G_i^{(A)}$, and $G_j^{(A)}$. This parametric covariance form is illustrated in Figure 3.

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 XXXX), and we again treat the likelihood of our sample covariance matrix as Wishart:

$$P(\widehat{\Omega} \mid \Omega^{(A)}) = \mathcal{W}\left(\widehat{\Omega} \mid L^{-1}\Omega^{(A)}\left(G^{(A)}, \vec{p}, \vec{\alpha}\right), L\right). \tag{12}$$

Now, the posterior probability can be expressed as a function of this parametric admixed covariance, $\Omega^{(A)}$,

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

as specified by the parameters p, $G^{(A)}$, $\vec{\alpha}$, and η, and the observed locations, G. We treat the location of the source of admixture for population k, $G_k^{(A)}$, and the population's admixture proportion, p_k, as random variables and jointly estimate them as part of our inference procedure.

We place the same priors as stated above on $\vec{\alpha}$, η, and we now specify priors on p and $G^{(A)}$. The admixture proportions, p, are capped at 0.5, to prevent populations from swapping identities with their source of admixture, and are heavily weighted at small values to be conservative with respect to admixture inference. They are independently beta-distributed: $2p_k \sim \beta(\alpha = 1, \beta = 100)$. The priors on the sources of admixture, $G^{(A)}$ 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: $G_k^{(A)} \sim \mathcal{N}(\mu = \bar{G}, \sigma = 2\bar{D}(G))$.

Using the example admixture scenario described above and used in the analysis depicted in Figure 2, we demonstrate the inference of populations' sources and strengths of admixture and illustrate the results in Figure 4. 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.

Models

The models described above may be used in various combinations. In the simplest model, populations may not choose their own locations, nor are they allowed to

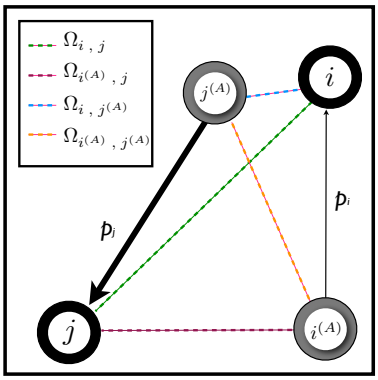


Figure 3: An illustration of the form of the admixed covariance given in Eqns. (11) and (20). Populations i and j are drawing admixture in proportions p_i and p_j from their respective sources of admixture, $i^{(A)}$ and $j^{(A)}$, and all pairwise spatial covariances (the Ω's) are shown. In this cartoon example, population j is drawing more admixture from its source $j^{(A)}$ than i is from its source $i^{(A)}$ (i.e., $p_j > p_i$).

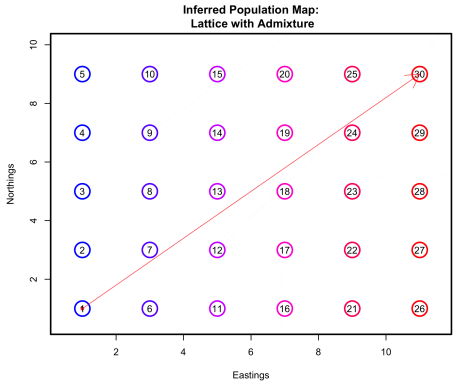


Figure 4: Posterior distribution of inference of the sources and strengths of admixture for the sampled populations. The admixed population (Population 30) is drawing admixture from the location of its source of admixture that was used to simulate the data (the location of Population 1).

draw admixture, and the only parameters to be estimated are those of the spatial covariance function given in Eqn (2), and the population-specific variance terms (η_k). 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. The full likelihood of this most complex parameterization is given by,

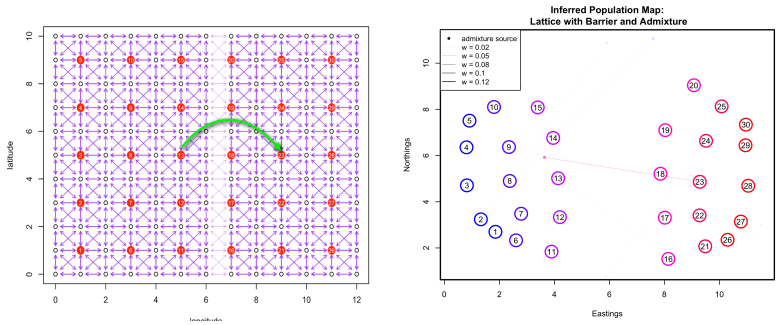
$$P(\widehat{\Omega} \mid \Omega^{(A)}) = \mathcal{W}\left(\widehat{\Omega} \mid L^{-1}\Omega^{(A)}\left(G', G^{(A)}, \vec{p}, \vec{\alpha}\right), L\right). \tag{14}$$

Now, the posterior probability can be expressed as a function of this parametric admixed covariance, $\Omega^{(A)}$,

$$P(G', G^{(A)}, \vec{p}, \vec{\alpha}, \eta \mid \widehat{\Omega}, L, G) \propto P(\widehat{\Omega} \mid \Omega^{(A)})P(\vec{\alpha})P(G')P(G^{(A)})P(\vec{p})P(\eta) \tag{15}$$

A note here about how this complex model is still identifiable

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. 5a). In the SpaceMix analysis (Fig. 5b), 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 ($\bar{p} = 0.05; 95\%$ credible interval $= 0.02 - 0.08$). *discuss how prior forces amount down, so it's cheaper to choose more westward pop as source*.



(a) Lattice (b) Barrier

Figure 5: Population maps inferred using SpaceMix under three different scenarios: a) simple lattice at equilibrium; b) a lattice with a barrier across the center line of longitude; c) a lattice with recent expansion on the eastern margin.

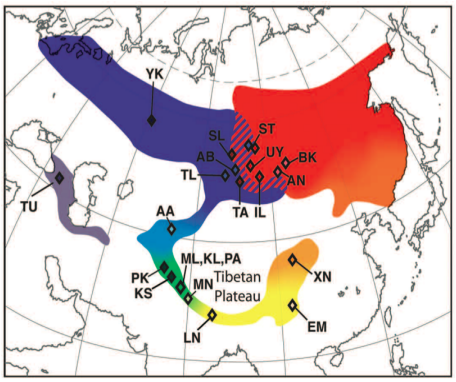
Empirical Applications

To demonstrate the applications of this novel method, we employed it in two canonical empirical systems: the greenish warbler ring species complex, and a global sampling of contemporary human populations.

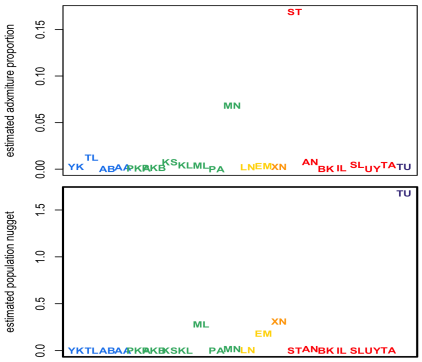
Greenish Warblers

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 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 a genetic discontinuity and reproductive isolation (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 (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 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 , for discussion).

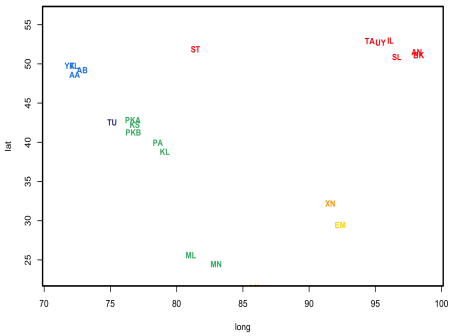
Based on a larger SNP dataset Alcaide et al (2014), have suggested that the greenish warbler species complex constitutes a 'broken' ring species, in which there have been historical discontinuities in gene flow that facilitated the evolution of reproductive isolation between adjacent forms. Because the questions in this system are fundamentally both geographic and genetic in nature, it is eminently SpaceMixable, 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.



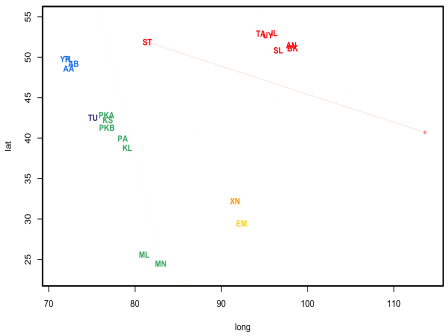
(a) Warbler subspecies distribution map



(b) Inferred admixture proportions and population nuggets



(c) Map inferred using SpaceMix



(d) Map inferred using SpaceMix with admixture arrows shown

Figure 6: Greenish warbler subspecies distributions contrasted with maps of all 22 sampled populations inferred using SpaceMix. For clarity of presentation, the inferred coordinates and parameter values are taken from the single draw of the MCMC with the highest posterior probability and have have been Procrustes transformed around the coordinates of the lattice used to simulate the data. (a) Greenish warbler subspecies distribution map (from Irwin et al NNN). (b) Inferred population admixture proportions and nugget parameters. (c) Inferred population map without admixture arrows shown, with population labels colored as in (a). (d) Inferred population map with admixture arrows (with thickness proportional to inferred admixture proportion) shown, with population labels colored as in (a).

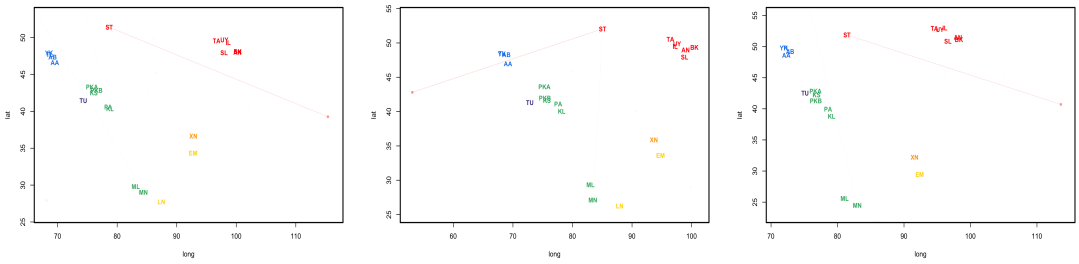
run method w. no admixture. I think we should define the random priors in the methods above, as it's not specific to any of the analyses

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 6) 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, 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 not commensurate 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 **POINT to graph showing this in supp**. *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* **?**. *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 random priors results. The Stolby population chooses the highest admixture proportion, with a mean of 0.19 **95% credible intervals**. Multiple runs agreed well on the level of admixture of the Stolby (see caption of Supplementary Figure 7). 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 (Supplementary Figure 7).

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 on an individual level. In these analyses, the sample size in each 'population' was 2 (for the two alleles in a diploid), and each individual chose its



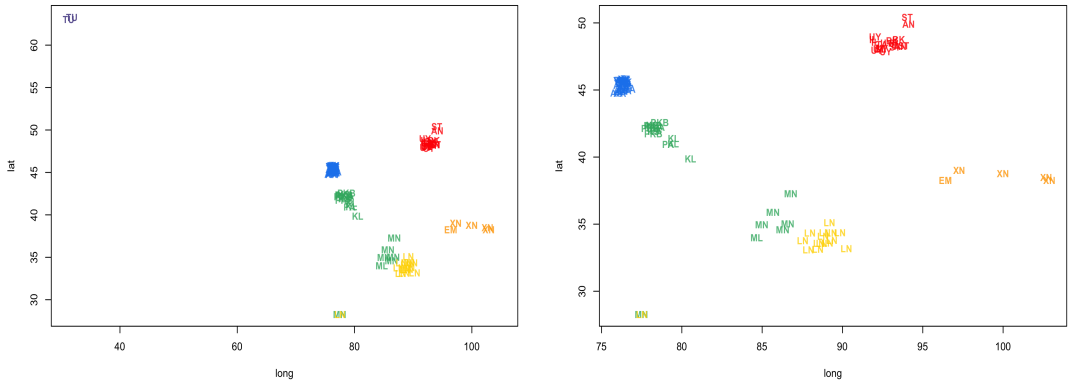
(a) (b) (c)

Figure 7: 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.

own location as well as the location of its source of admixture, the proportion of that admixture, and its nugget. As with the analysis on multi-sample populations, the results *approximately* mirror the geography of the individuals. *Individuals choose very low levels of admixture...*

There are 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 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 is truely independent (e.g. our populations could form along a line). 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 clouds NNN). Is this biologically meaningful? A somewhat similar horseshoe shape appears when a principal components (PC) analysis is conducted and individuals are ploted on the first two PCs (Alcaide et al (2014), see our Supp. Figure NNN). However, as discussed by Novembre & Stephens 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 Supp. Figure NNN) not because of any particular gene flow connection between the tips but rather because of the



(a) Inferred map of warbler individuals (b) Closeup of non-*nitidus* samples

Figure 8: Inferred maps for individual warbler individuals, colored by subspecies. (a) complete map including Turkish *nitidus* samples. (b) close up of all non-*nitidus* samples.

othogonality requirement of PCs (see Novembre & Stephens for more discussion). When spacemix is applied to one dimensional stepping stone data, the placement of samples is consistent with a line. 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 postions around the horseshoe, the posterior positons of the populations still curl up to form a horseshoe. 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 individuals, 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. Examing pairwise sequence difference shows that these two individuals show unusually recent common ancestry (see SuppMat Figure NNN), 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. *move the commented out bit below to the caption of the pairwise seq. plot.*

The SpaceMix map also diverges from the observed map in the distribution of

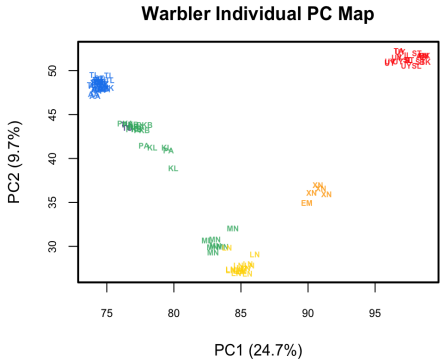


Figure 9: The map of warbler individuals derived from a Principal Components analysis.

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, 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 NNN).

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 set out to create a geogenetic map for a worldwide sample of modern human populations.

Specifically, our research questions were:

1. What is the geography of human population structure? Which populations cluster with which, and how does the geogenetic map contrast with the observed geography?
2. Can the map we estimate reconstruct the many of the important expansion events in human pre-history, including the out-of-Africa expansion, the colonizations of Oceania, and the colonization of the Americas via Beringia?
3. Which human populations are most greatly admixed, and from where? Will the SpaceMix results confirm the admixture findings of previous studies?

To answer these questions, we used a subset of 10,000 SNPs from the SNP dataset of Hellenthal et al (2014), which is comprised of 1490 sampled in 95 populations (see Fig. 10 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, the position of their sources of admixture, and the proportion of admixture they draw.

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 populations sampled on their continent, and the relative placement of populations is similar to that of their observed geography. Sub-Saharan African

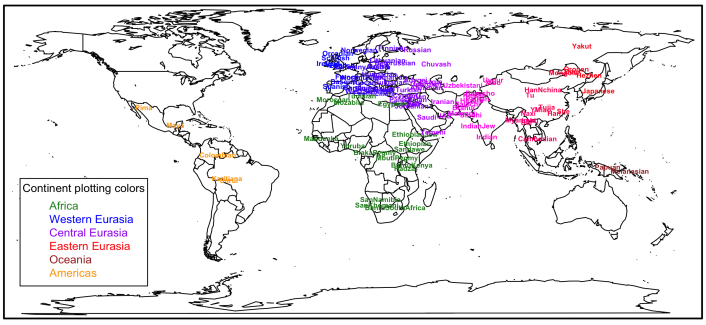
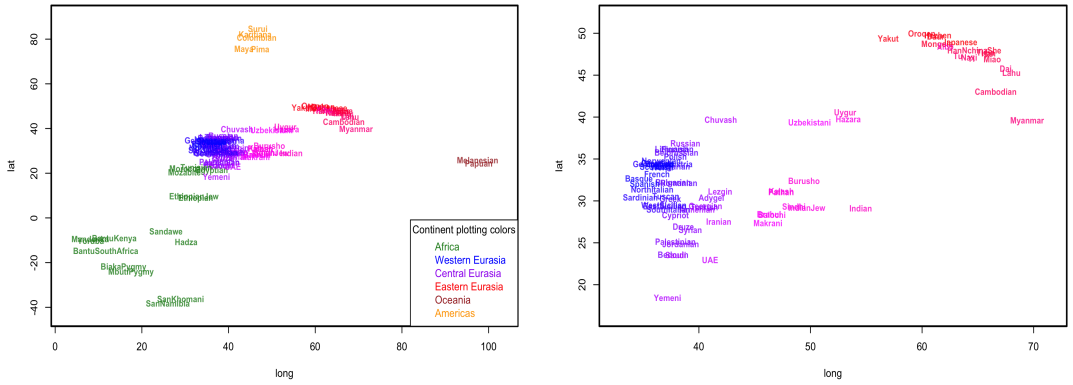


Figure 10: The map of sampled human populations

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. 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 populations from Oceania.

We also recover evidence for several major population expansions and colonization routes in human pre-history. First, the scale of inferred inter-population distance within sub-Saharan Africa is much greater than that between any other group (see Fig NNN), a pattern indicative of a bottleneck (or series of bottlenecks) during an out-of-Africa event. In addition, we see that both the populations in the Americas and 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 were ancestrally East Asian (REF?), but that



(a) Inferred map of human populations (b) Closeup of Eurasian populations

Figure 11: Map of human populations, inferred without admixture, colored by continent. (a) complete map; (b) close up African populations; (c) close up of all non-African populations; (d) close up of Eurasian populations.

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, as well as in the placement of the Moroccan, Mozabite, and Ethiopian populations, which are choosing locations between the western Eurasian cluster and the sub-Saharan Africans. To investigate possible patterns of admixture further, we ran a SpaceMix analysis in which we estimated, for each population, a location, a source of admixture, and the magnitude of that admixture.

Conclusion

In this paper we have presented blah blah blah. We believe this represents an advance over previous methods because blah blah blah. This method can be used to answer a variety of empirical questions, including blah, blah, and blah, and also serves as an intuitive data visualization tool.

Future Directions

spatiotemporal model

spatialSTRUCTURE

Methods

this is all just taken from the main text and put together in rough order.

Normal Approximation to Drift

The parameter δ_B, the amount of drift that has occurred in B since it split off from A, is approximately equal to $\frac{t}{2N_e}$, where t is time in generations, and N_e is the effective population size. Note that 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. Note also that this approximation works best for alleles at intermediate frequency, which are less likely to be lost or fixed during the interval separating A and B. However, if the extent of drift sufficiently limited (time scales are short or population sizes are large), the approximation of inter-generational binomial sampling to Brownian motion is good (*see figure comparing fit of binomial sampling to fit of normal approx?*).

Unlike multivariate normal observations however, allele frequencies at variable loci are constrained to have a mean between 0 and 1, and may have heterogeneous variance across loci. To more closely meet the assumptions of our statistical inference model, we mean-center and normalize our observations by the following procedure.

Sample Mean-Centering Procedure

However, we do not have the population frequencies nor 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 population sample size in population k, as $\bar{S}_k = \frac{1}{L}\sum_L S_{\ell,k}$. We then calculate the weighted sample mean frequency at locus ℓ as follows:

$$\bar{f}_\ell = \frac{1}{\sum_K S_{\ell,k}} \sum_K \hat{f}_{\ell,k} S_{\ell,k} \tag{16}$$

We approximate the binomial variance at each locus by $\bar{f}_\ell(1 - \bar{f}_\ell)$. To avoid modeling the heterogeneous effect of this variance across loci, we standardize by this variance. We call the standardized allele frequencies $X_{\ell,k}$, and calculate them as follows:

$$X_{\ell,k} = \frac{\hat{f}_{\ell,k} - \bar{f}_\ell}{\bar{f}_\ell(1 - \bar{f}_\ell)} \tag{17}$$

Note, by using the sample mean frequency to mean-center our observations, we lose a degree of freedom, and reduce the covariance across loci between populations (sometimes inducing negative covariance among distant populations). We accommodate the extra sampling noise distortion, and reduced rank of the covariance matrix by assuming that our

$$X_\ell \sim MVN(0, \Omega') \tag{18}$$

where Ω' is a simple transform of Ω.

Between 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. Across loci, populations with highly shared drift will tend to covary more strongly in the deviations their allele frequencies take from some ancestral (or global) mean.

Modeling Admixture

where $f_{\ell,i}$ is the allele frequency in population i, $f_{\ell,j}$ is the allele frequency in population j, and p is the admixture proportion, which varies between 0 and 1 and describes the extent to which populations i and j are contributing to the genetic make-up of population k.

To infer the spatial context of this admixture, we allow each population a point in space, which we refer to as its source of admixture, from which it draws its admixture, and we model both the location of that source and the extent (proportion) of that admixture. The observed allele frequencies in sampled populations are therefore a weighted average of the model-estimated allele frequencies at the geographic location of the sampled population and those at the coordinates of the source from which the observed population draws admixture. That is, the observed allele frequencies in population k are modeled as follows:

$$f_k = pf_{k'} + (1-p)f_j, \tag{19}$$

where $f_{k'}$ are the model-estimated allele frequencies across loci at the spatial location of population k and f_j are the model-estimated allele frequencies at the spatial location of the source of admixture j, from which population k is drawing admixture in proportion p. The admixture proportion p is constrained to vary between 0 and 0.5, such that at least half of a population's genetic make-up must be determined by its geographic location.

We re-introduce the population-specific variance terms on each diagonal element of this admixed covariance matrix. The full expression for our admixed covariance function is below.

$$\begin{aligned} \Omega_{i,j}^{(A)} = (1-p_i)(1-p_j)\Omega_{i\,,\,j} \;\times \\ (p_i)(1-p_j)\Omega_{i^{(A)},\,j} \;\times \\ (p_j)(1-p_i)\Omega_{i\,,\,j^{(A)}} \;\times \\ (p_i)(p_j)\Omega_{i^{(A)},\,j^{(A)}} \;+ \\ \delta_{i,j}\bar{S_k}^{-1} + \delta_{i,j}\eta_k \end{aligned} \tag{20}$$

where I is the identity matrix, $\bar{S_k}$ is the mean sample size in population k across all loci, and η_k is the nugget estimated in population k.

Empirical Applications

The analysis procedure is detailed in Appendix XXX. 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. We then repeated these analyses, but treated 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 ($G^{(A)}$), 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 of the warbler individuals (Figure NNN). The concordance between the PC map and the SpaceMix map is generally quite good, and we discuss the interpretation of the geography implied by this map further below.

To investigate the potential reason for this behavior, we calculated average pairwise sequence divergence at the 2,247 polymorphic loci in the dataset between all 95 individuals and plotted it against the pairwise geographic distance between the individuals (see SuppMat Figure NNN). The pairwise sequence divergence (0.103) at polymorphic loci between Lud-MN3 and Tro-LN11 is significantly lower than that between any other pair of individuals separated by a comparable distance - lower, in fact, than any comparison between individuals that were not co-located, and lower than any pairwise divergence between any pair of individuals save that between the two Turkish *nitidus* samples.

Inference

For details on our Bayesian inference framework and Markov chain Monte Carlo inference procedure, please see the Section: How I spent the past year!

Appendix

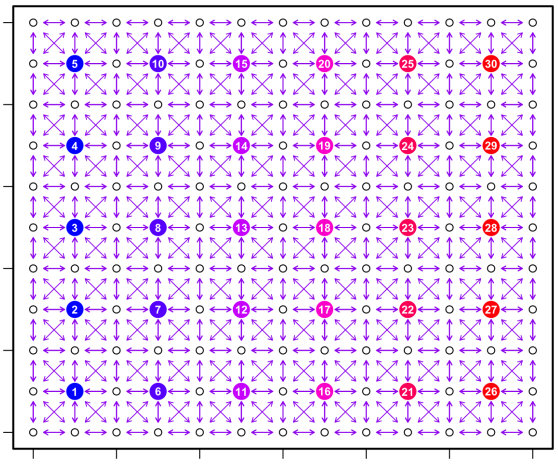
The analysis procedure went as follows:

1. five independent chains were run for 5e6 generations each in which populations were allowed to choose their own locations (but no admixture). Population locations were initiated at the origin (i.e. - at generation 1 of the MCMC, $G_i' = (0, 0)$), and all other parameters were drawn randomly from their priors at the start of each chain.

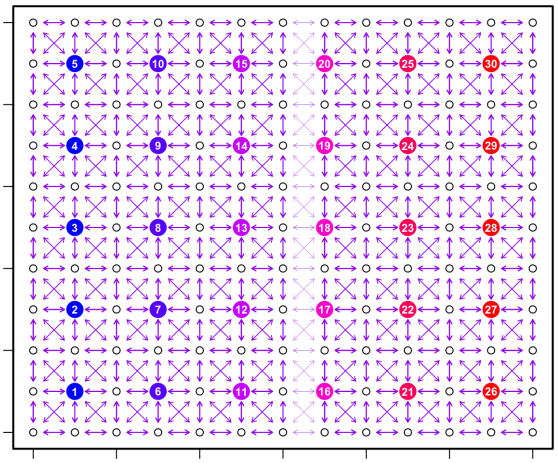
2. The chain with the highest posterior probability at the end of the analysis was selected and identified as the"Best Short Run".

3. A chain was initiated from the parameter values in the last generation 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, $G^{(A)}$ were initiated at the origin. This chain (the "Long Run") was run for 1e8 generations, and sampled every 1e5 generations for a total of 1e3 draws from the posterior.

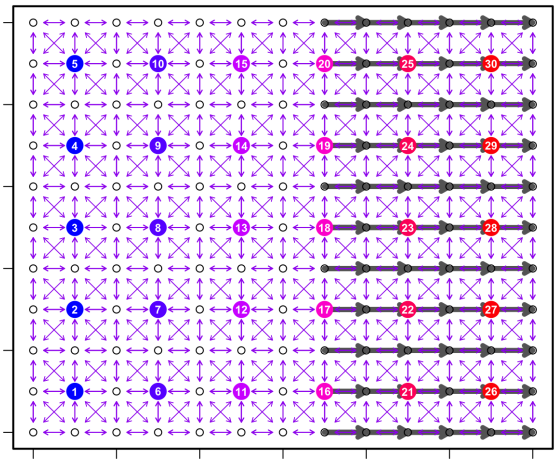
Supplementary Materials



(a) Likelihood



(b) Posterior probability



(c) Posterior probability

Figure 12: Different simulation scenarios: (a) basic lattice; (b) lattice with a longitudinal barrier; (c) lattice with expansion event.

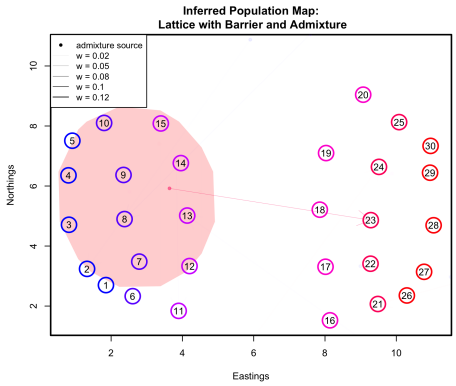
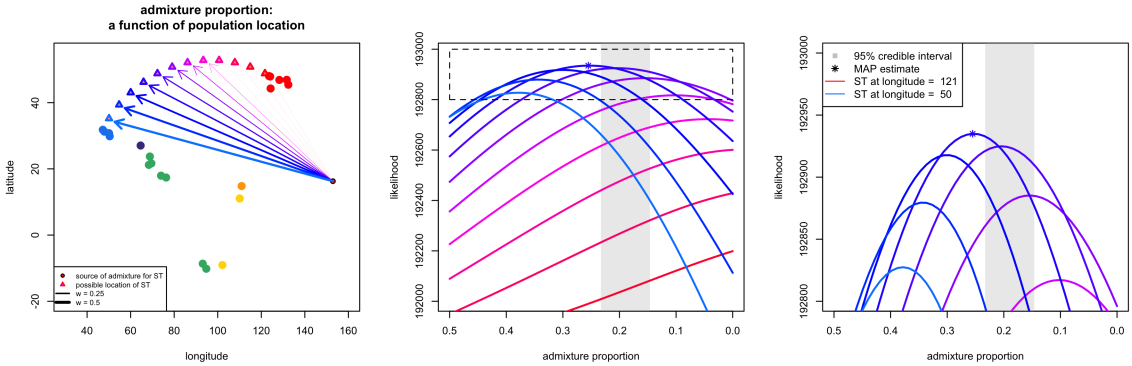
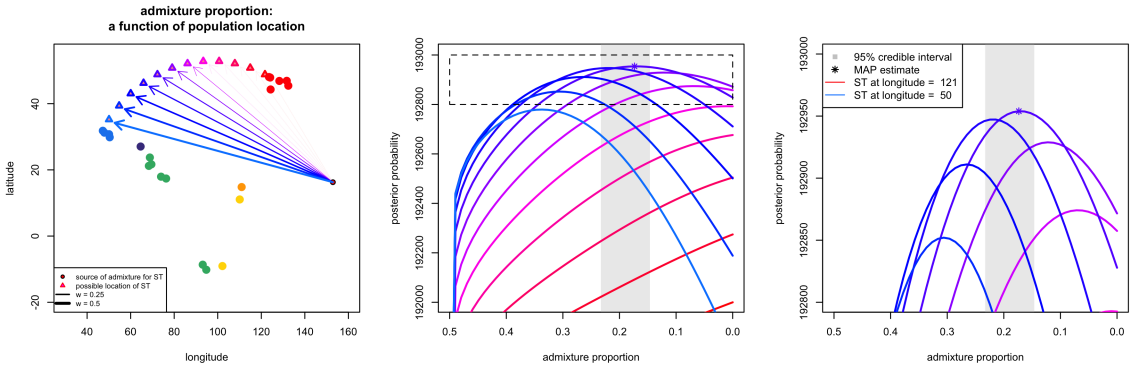


Figure 13: credible interval of where population 23 draws admixture from.

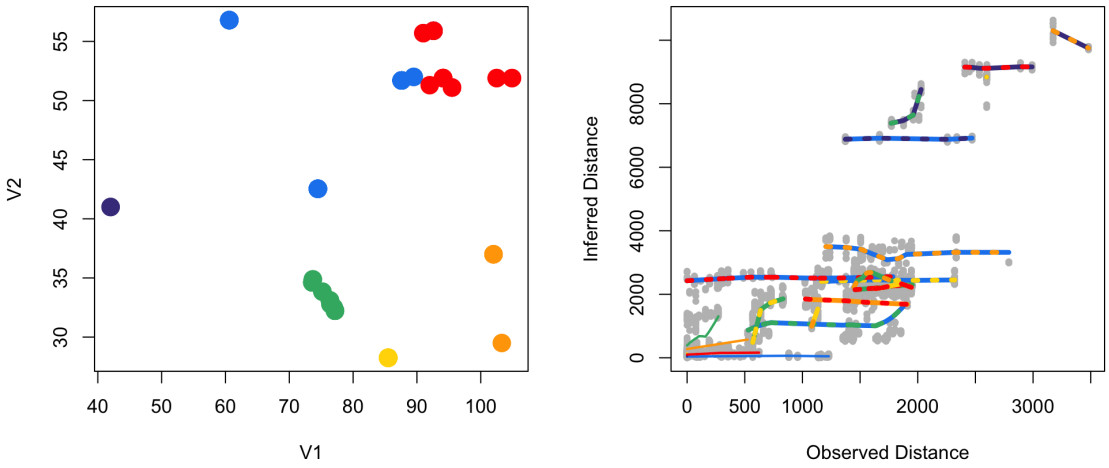


(a) Likelihood

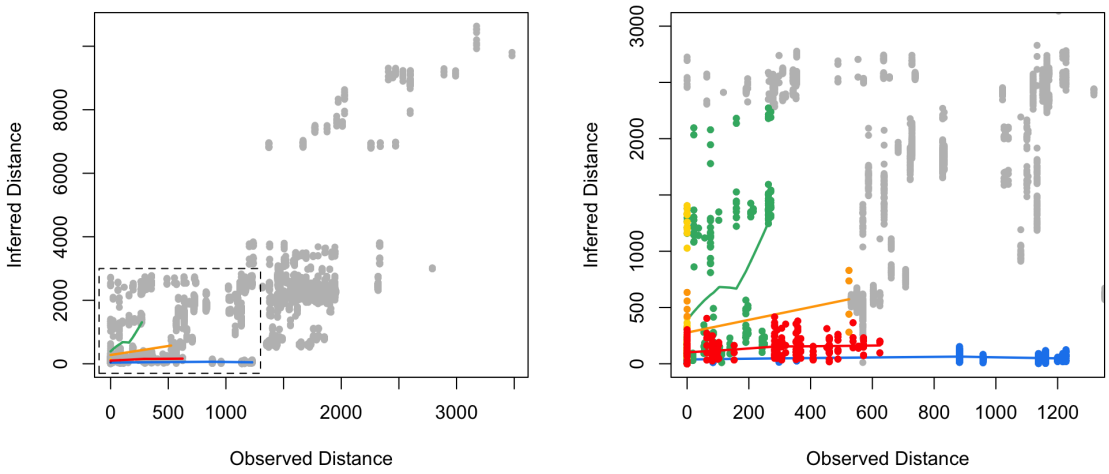


(b) Posterior probability

Figure 14: Likelihood surfaces for different placements of population ST between *plumbeitarsus* and *viridanus* clusters: (a) log likelihood surface; (b) posterior probability surface, incorporating the priors.



(a) All population pairs



(b) Just within population comparisons

Figure 15: Comparing observed to estimated pairwise distance between warbler individuals, (a) between and (b) within subspecies populations.

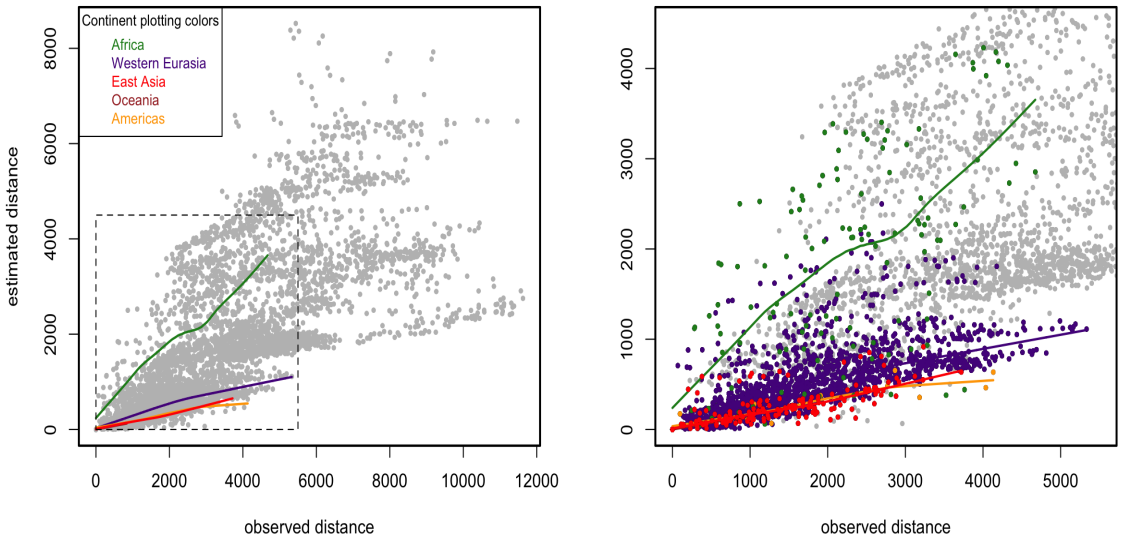


Figure 16: Comparison of observed distance to estimated distance between human populations, colored by continent from which populations were sampled (i.e. - two populations sampled from Africa are green). Eurasia is divided into Western Eurasia and East Asia.