Inferring Continuous and Discrete Population Genetic Structure Across Space

Gideon S. Bradburd[1,a], Peter L. Ralph[2,b], Graham M. Coop[3,c]

[1] Department of Integrative Biology, Ecology, Evolutionary Biology, and Behavior Graduate Group, Michigan State University, MI 48824

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

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

[a]bradburd@msu.edu; [b]plr@uoregon.edu; [c]gmcoop@ucdavis.edu

Abstract

One of the classic problems in population genetics is the characterization of discrete population structure when the genotyped samples also show continuous patterns of genetic differentiation. Especially when sampling is discontinuous, clustering or assignment methods may incorrectly ascribe differentiation due to continuous processes (e.g., isolation by distance) to discrete processes, such as geographic, ecological, or reproductive barriers between populations. This is partly a result of the difficulty of sampling uniformly and continuously across the range of a population or species, but more, it reflects a shortcoming of current methods for inferring and visualizing population structure from genetic data in the face of data that are characterized by both continuous and discrete population structure. Here, we present a novel statistical framework for the simultaneous inference of continuous and discrete patterns of population structure. The method estimates ancestry proportions for each sample from a set of discrete population clusters, and, within each cluster, estimates a rate at which relatedness decays with distance. This model explicitly addresses the "clines vs. clusters" problem in modeling population genetic variation by jointly accommodating both continuous and discrete patterns of differentiation. We demonstrate the utility of this approach using both simulations and an empirical application.

Introduction

The fundamental goal of evolutionary genetics is to link patterns of genetic variation to the processes that have generated and maintained them. Often, the first step in the analysis of genetic data is to define evolutionary units of analysis by delineating discrete populations from which individuals or alleles are sampled **??**. These units are useful for defining a scope of study in research on selection or local adaptation **??**, or on identifying either locally adapted alleles (human skin color, ralph & coop, gartersnakes) or alleles involved in disease (cystic fibrosis). Delineating populations can also be useful for informing conservation priorities **??**.

There have been many methods proposed to characterize populations, including generating population phylogenies **??**, k-means clustering analyses performed on dimensionality-reduction approaches **??**, such as principal components analysis **??**, and model-based clustering approaches **??**, among others. These methods each perform best under different specific demographic histories of samples being analyzed, but all can give misleading results when applied to data that show a continuous pattern of isolation by distance **??**.

Isolation by distance (IBD) can

Patterns of population genetic differentiation are characterized by both discrete and continuous structure. Identification of genetically discrete populations is important not just for understanding the distribution of genetic variation on a landscape, but also for understanding the demographic history of a species, or, from a conservation perspective, for characterizing discrete management units. However, it is frequently difficult to disentangle continuous and discrete patterns of genetic differentiation. By introducing a statistical framework that simultaneously models discrete and continuous patterns of population structure, we can do a more accurate job describing and quantifying both.

Methods

Data The statistical framework of our approach is conceptually similar to **?** and **?**. The genetic data modeled consist of allele frequencies F at L unlinked, bi-allelic single nucleotide polymorphisms (SNPs) genotyped across N samples. The sample frequency at locus ℓ in sample n, $f_{n,\ell}$, is calculated by first arbitrarily choosing an

allele segregating at locus ℓ to count, then dividing the total number of observations of that counted allele by the total number of chromosomes genotyped at that locus in sample n. We then calculate the sample covariance in allele frequencies, $\widehat{\Omega}$, as

$$\widehat{\Omega} = \frac{1}{L} F F^T \tag{1}$$

Continuous and discrete differentiation We wish to describe the observed patterns of genetic variation as a combination of discrete population clusters, within which genetic variation is continuously distributed, and between which sampled individuals or populations can be admixed. Each of these discrete clusters is modeled as a spatial process in which migration between neighboring demes acts to homogenize allele frequency changes that arise locally due to drift; this interplay between drift and migration results in a continuous pattern of isolation by distance within each cluster. Multiple population cluster processes can co-occur in space, leading to discrete jumps in allele frequencies across many loci over small geographic distances, and the genotyped samples can be admixed between different clusters.

The continuous decay of allele frequency covariance with geographic distance within a cluster is described using a simple and flexible powered exponential function; the within-cluster covariance between samples i and j is given by:

$$G_{i,j}^{(k)} = \alpha_0^{(k)} \left(\exp\left(-(\alpha_D^{(k)} D_{i,j})^{\alpha_2^{(k)}} \right) \right) + \mu^{(k)} \tag{2}$$

where $G_{i,j}$ is the covariance function between samples i and j within a cluster, and the superscript (k) denotes that this function is specific to the kth cluster. The quantity $D_{i,j}$ is the observed geographic distance between samples i and j, the α parameters control the shape of the decay of covariance with distance, and μ is a parameter that describes the amount of shared drift within a cluster. The shared drift parameter $\mu^{(k)}$ can be interpreted as the branch length connecting the kth population to the population ancestral to all modeled clusters.

Genotyped samples can be admixed between these different clusters. The admixture proportion of the ith sample in the kth cluster, $w_i^{(k)}$, gives the probability that an allele in sample i was derived from cluster k; each sample's admixture proportions across all clusters must sum to one.

We can then describe the covariance between samples i and j across all clusters, $\Omega_{i,j}$, by summing their within-cluster spatial covariances ($G_{i,j}^{(k)}$ in cluster k) across all K clusters and weighting by the product of those samples' admixture proportions in each cluster.

$$\Omega_{i,j} = \gamma + \sum_{K} w_i^{(k)} w_j^{(k)} G_{i,j}^{(k)} + \delta_{i,j}\eta_i \tag{3}$$

In addition to the admixture-weighted sum of the within-cluster spatial covariances, this function contains two terms, γ and $\delta_{i,j}\eta_i$. The first, γ, describes the global covariance between all samples, and arises because all samples share an ancestral mean allele frequency at each locus. In the second term, $\delta_{i,j}$ is an indicator variable that takes a value of 1 when i equals j and 0 otherwise. and η_i gives the variance specific to sample i, This term on the diagonal of the parametric covariance matrix is meant to capture processes shaping variance within the sampled deme, such as inbreeding.

Likelihood and inference If we assume that the allele frequencies $\widehat{F}$ are independent and multivariate normally distributed, their sample covariance $\widehat{\Omega}$ will be Wishart distributed with degrees of freedom equal to L, the number of loci genotyped. Linkage disequilibrium (LD) between loci will decrease the effective number of degrees of freedom, which we discuss further below. The likelihood of the sample standardized allele frequency covariance is therefore given by

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

We estimate the values of the parameters of the model using a Bayesian approach. Acknowledging the dependence of the parametric covariance matrix Ω on its constituent parameters $w, \alpha, \mu, \eta, \gamma$ and on observed quantiy D with the notation $\Omega(w, \alpha, \mu, \eta, \gamma, D)$, we denote the posterior probability of the parameters as:

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

The priors, $P(w), P(\alpha), P(\mu), P(\eta), P(\gamma)$, are detailed in the Appendix, and the constant of proportionality is the normalization constant. We use a Hamiltonian

Monte Carlo sampling algorithm implemented in the statistical language STAN (**????**) to estimate the posterior distribution on the parameters. We also present an R package (**?**) called `conStruct` that functions as a wrapper around this inference machinery.

Results

Simulations

Empirical Applications

Discussion

Acknowledgements

This work was supported in part by the National Science Foundation under award number NSF #1262645 (DBI) to PR and GC, the National Institute of General Medical Sciences of the National Institutes of Health under award numbers NIH RO1GM83098 and RO1GM107374 to GC, and the National Science Foundation under award numbers NSF # 1148897 and # 1402725 to GB.

Appendix

1 Model rationale

1.1 Drift, admixture, and allele frequencies

Drift We imagine that the allele frequencies at each locus in each sample are the sum of 3 components: the ancestral allele frequency ϵ shared by all samples, the deviation from that ancestral mean in the kth population, $\Delta^{(k)}$, which is shared by all samples with 100% ancestry in that population, the deviation specific to the ith sample, $\Delta^{(i)}$, which captures drift not shared by all samples at the population level (i.e., subpopulation-specific drift due to, e.g., inbreeding).

If all samples drew all of their ancestry from any one of the K clusters, the allele frequency in the ith sample at the ℓth locus would be given by:

$$F_{i,\ell} = \epsilon_\ell + \Delta_\ell^{(k)} + \Delta_\ell^{(i)} \tag{A1}$$

Admixture The model above describes the simple case in which samples draw 100% of their ancestry from only a single cluster each. To accommodate admixture between clusters, we can model allele frequencies within samples as linear combinations of the population-specific drift terms across the clusters from which they draw ancestry. To do so, we introduce a term $w_i^{(k)}$ that describes the admixture proportion of sample i in cluster k, which can be interpreted as the probability that an allele sampled in the ith sample came from the kth cluster. The allele frequency in the ith sample at the ℓth locus can therefore be written as:

$$F_{i,\ell} = \epsilon_\ell + \sum_K \left(w_i^{(k)} \Delta_\ell^{(k)} \right) + \Delta_\ell^{(i)} \tag{A2}$$

where

$$\sum_K w_i^{(k)} = 1 \tag{A3}$$

1.2 The Normal approximation to drift

Genetic drift is serial binomial sampling, but over short timescales, and for intermediate ancestral allele frequencies, we can reasonably model the allele frequencies at each locus across populations as multivariate normal. This will be a good approximation if the amount of elapsed drift in each sample is relatively small, as would be expected if time since divergence from a common ancestral population is short, or effective population sizes are large.

Beginning with the simplest case of two clusters ($K = 2$), with all samples drawing 100% of their ancestry from either one or the other cluster, we can model allele frequencies at locus ℓ as:

$$F_\ell \sim MVN\left(\mu = \epsilon_\ell + \iota^{(k)}\Delta_\ell^{(k)}, \Sigma = \Omega\right) \tag{A4}$$

where $\iota^{(k)}$ is an indicator variable that equals 1 for samples with 100% membership in cluster k and 0 otherwise. In Eqn (**??**), all entries of the covariance Ω have value 0 except the diagonals, which describe sample-specific drift, and have the following value:

$$\Omega_{i,i} = \mathrm{Var}\left(\Delta^{(i)}\right) \tag{A5}$$

1.3 Describing relatedness as a covariance

However, our data can consist of many loci, so it may be computationally impractical to model the ancestral, cluster, and sample components of each sampled allele frequency. Instead, we can describe the covariance induced between samples due to the sharing of the components. All samples share the same ancestral mean at all loci, which leads to a global covariance equal to the variance of the ancestral frequencies across loci. Likewise, samples that are in the same cluster will have a covariance equal to the variance of the cluster-specific deviates from the ancestral frequencies across loci.

In Eqn **??** above, we described the relatedness between samples due to the shared ancestral allele frequency and shared population-level drift via the mean of the

multivariate normal. We can describe this relatedness as a covariance as follows:

$$\Omega_{i,j} = \begin{cases} \text{Var}(\epsilon) + \text{Var}\left(\Delta^{(k)}\right) + \text{Var}\left(\Delta^{(i)}\right), & \text{if } i = j \\ \text{Var}(\epsilon) + \text{Var}\left(\Delta^{(k)}\right), & \text{if } \iota_i^{(k)} = \iota_j^{(k)} \text{ and } i \neq j \\ \text{Var}(\epsilon), & \text{otherwise} \end{cases} \tag{A6}$$

Generalizing this result to any number of clusters and incorporating admixture as in Eqn **??**, we can write the form of the covariance of samples that are admixed between K clusters as:

$$\Omega_{i,j} = \text{Var}(\epsilon) + \sum_K \left(w_i^{(k)} w_j^{(k)} \text{Var}(\Delta^{(k)}) \right) + \delta_{i=j} \text{Var}(\Delta^{(i)}) \tag{A7}$$

where $\delta_{i=j}$ is an indicator variable that equals 1 when i is equal to j and 0 otherwise, as in Eqn **??**.

1.4 Incorporating spatial differentiation

Equation **??** describes a model in which samples can be continuously admixed between a set of K discrete clusters. In this model, any pair of samples with 100% ancestry in a cluster have exactly the same covariance with each other (namely $\text{Var}(\epsilon)+\text{Var}(\Delta^{(k)})$). However, we also wish to model continuous decay of covariance with spatial separation between samples. That is, we expect (and want our model to reflect) that samples within the same cluster will have higher covariance if they are sampled closer together than if they are sampled farther apart. To describe this spatial pattern, we build in a spatial component to our covariance model.

Specifically, we write that the covariance between a pair of samples i and j that draw all their ancestry from a single cluster k decays exponentially as a function of the distance between their sampling locations as follows:

$$G_{i,j}^{(k)} = \alpha_0^{(k)} \times \left(\exp\left(-(\alpha_D^{(k)} D_{i,j})^{\alpha_2^{(k)}} \right) \right) + \mu^{(k)} \tag{A8}$$

where the α parameters control: the sill of the covariance (α_0), the rate at which covariance decays with distance D (α_D), and the shape of that decay (α_2) in the kth cluster. The parameter $\mu^{(k)}$ describes the covariance shared by all samples

in the kth cluster; it and the spatial covariance function G together describe the quantity $\text{Var}(\Delta^{(k)})$ from Eqn (**??**).

Tying these ideas together, we can then construct a covariance that describes

1. continuous decay of covariance within a cluster
2. a discrete amount of covariance shared within a cluster
3. continuous admixture between clusters.

$$\Omega_{i,j} = \text{Var}(\epsilon) + \sum_{K} \left(w_i^{(k)} w_j^{(k)} G_{i,j}^{(k)}(\theta) \right) + \delta_{i=j} \text{Var}(\Delta_i) \tag{A9}$$

where $G_{i,j}^{(k)}(\theta)$ denotes the dependence of the spatial covariance between samples i and j within a cluster on other quantities θ: specifically, the parameters $\vec{\alpha}^{(k)}$ and $\mu^{(k)}$, which are specific to the kth cluster, as well as on the observed quantity $D_{i,j}$, the pairwise distance between samples i and j.

2 Likelihood

If the allele frequency data are well approximated by a Gaussian, their sample covariance is a sufficient statistic, so that calculating the likelihood of their sample covariance is the same as calculating the probability of the frequency data up to a constant. We can therefore model the covariance of the sample allele frequencies, $\widehat{\Omega}$, as a draw from a Wishart distribution with degrees of freedom equal to the number of loci L across which the sample covariance is calculated:

$$\begin{aligned} \widehat{\Omega} &= \frac{1}{L} F F^T \\ \widehat{\Omega} &\sim \mathcal{W}\left(L\Omega, L\right) \end{aligned} \tag{A10}$$

A benefit of directly modeling the sample allele frequency covariance is that, after the initial calculation of the sample covariance matrix, the computation time of the likelihood is not a function of the number of loci, so inference can theoretically be done using whole genome data.

3 Models, parameters, and priors

Spatial vs. nonspatial In this paper, we discuss two types of models, spatial and nonspatial. The spatial model is parameterized as in Eqn **??**, and the nonspatial model is described in Eqn **??**. The nonspatial model therefore has $3K$ fewer parameters than the spatial model, as there are three α parameters that describe the continuous differentiation effect of distance per cluster, and they are not included in the nonspatial model.

Single cluster Each of these models can be run with a single cluster ($K = 1$), which involves a slight re-parameterization, detailed below.

For the spatial model, the single-cluster parametric covariance is:

$$\Omega_{i,j} = \text{Var}(\epsilon) + \alpha_0^{(k)} \times \left(\exp\left(-(\alpha_D^{(k)} D_{i,j})^{\alpha_2^{(k)}}\right)\right) + \delta_{i=j}\text{Var}(\Delta_i) \tag{A11}$$

For the nonspatial model, the single-cluster parametric covariance is:

$$\Omega_{i,j} = \text{Var}(\epsilon) + \delta_{i=j}\text{Var}(\Delta^{(i)}) \tag{A12}$$

Priors We use a Bayesian approach to parameter inference,

A table of all parameters, their descriptions, and their priors is given in Table **??** below.

Parameter	Description	Prior
$\boldsymbol{\gamma}$	global covariance due to shared ancestral frequency	$\gamma \sim \mathcal{N}(\mu = \text{Var}(\bar{f}), \sigma = 0.5)$
$\boldsymbol{\alpha_0^{(k)}}$	controls the sill of the covariance matrix in cluster k	$\alpha_0^{(k)} \sim \mathcal{N}(\mu = 0, \sigma = 1)$
$\boldsymbol{\alpha_D^{(k)}}$	controls the rate of the decay of covariance with distance in cluster k	$\alpha_D^{(k)} \sim \mathcal{N}(\mu = 0, \sigma = 1)$
$\boldsymbol{\alpha_2^{(k)}}$	controls the shape of the decay of covariance with distance in cluster k	$\alpha_2^{(k)} \sim U(0, 2)$
$\boldsymbol{\eta_i}$	the nugget in population i (population specific drift parameter)	$\eta_i \sim \mathcal{N}(\mu = 0, \sigma = 1)$
$\boldsymbol{\mu^{(k)}}$	cluster-specific shared drift in cluster k	$\mu^{(k)} \sim \mathcal{N}(\mu = 0, \sigma = 1)$
$\boldsymbol{w_i}$	admixture proportions sample i draws across K clusters	$w_i \sim \text{Dir}(\alpha_1 ... \alpha_K = 0.1)$

Table A1: List of parameters used in the `conStruct` model, along with their descriptions and priors. The mean of the Normal prior on γ, $\text{Var}(\bar{f})$, is the variance of the sample mean allele frequencies across loci.

4 Cross validation

To perform model comparison, we employ a Monte Carlo cross-validation approach, also known as repeated random sub-sampling validation. Briefly, we follow the following procedure:

1. For each of X replicates:
   (a) partition the allele frequency data into a 90% "training" partition (F_1^x) and a 10 %"testing" partition (F_2^x)
   (b) run our inference procedure on the training partition to estimate model parameters θ_{mk} for:
      i. m: the spatial and the nonspatial model
      ii. k: the number of clusters 1 through K
   (c) calculate the mean log likelihood of the testing data partition, over the posterior distribution of training-estimated parameters for each model ($\bar{\mathcal{L}}(F_2^x \mid \theta_{mk})$, henceforth $\bar{\mathcal{L}}_{xmk}$)
   (d) generate standardized mean log likelihoods, $\mathcal{Z}_{xmk}$, across models:
      i. identify the highest mean log likelihood, $\bar{\mathcal{L}}_{xmk}^{\max}$
      ii. subtract $\bar{\mathcal{L}}_{xmk}^{\max}$ from $\bar{\mathcal{L}}_{xmk}$ across all models, such that the standardized log likelihood, $\mathcal{Z}_{xmk}$, of the best model is 0, and less than 0 for all inferior models.
2. For each model (i.e., each combination of m and k) calculate the mean standardized log likelihood of the testing data partition across X replicates, as well its standard error and 95% confidence interval:
   (a) mean

$$\bar{\mathcal{Z}}_{mk} = \frac{1}{X} \sum_{x=1}^{X} \mathcal{Z}_{xmk} \tag{A13}$$

   (b) standard error

$$SE_{\bar{\mathcal{Z}}_{mk}} = \frac{1}{\sqrt{X}} \sum_{x=1}^{X} \left[\mathcal{Z}_{xmk} - \bar{\mathcal{Z}}_{mk} \right] \tag{A14}$$

   (c) 95% confidence interval

$$95\%\text{CI} = \bar{\mathcal{Z}}_{mk} \pm 1.96 \times SE_{\bar{\mathcal{Z}}_{mk}} \tag{A15}$$

This cross-validation procedure generates a mean predictive accuracy for each model and each value of K, as well as a confidence interval around that mean, which can then be used for model comparison or selection.