Inferring Continuous and Discrete Population Genetic Structure Across Space and Time

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

[1] Museum of Vertebrate Zoology, Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720

[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@berkeley.edu; [b]pralph@usc.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., drift across space or time) to discrete processes, such as geographic, reproductive, or behavioral barriers between populations. This is partly a result of the difficulty of sampling uniformly and continuously across the spatiotemporal 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 spatial and/or temporal distance. This model explicitly addresses the "clines vs. clusters" problem in quantifying population structure by jointly accommodating both continuous and discrete patterns of differentiation. The model also naturally captures population replacement, a phenomenon for which there is substantial evidence in humans from archaeological evidence and ancient DNA. We demonstrate the utility of this approach using a combination of ancient and modern human individuals sampled throughout Europe, and find compelling evidence for space aliens.

Argument

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.

This difficulty arises for two reasons. First, both patterns are frequently present in empirical systems, and, with incomplete spatial or temporal sampling, it can be difficult to differentiate continuous and discrete patterns. For example, if a species is dispersal limited, it may exhibit a pattern of isolation by distance (IBD), in which samples taken from populations separated by greater geographic distance are more genetically differentiated than samples taken from neighboring populations. If only two samples are taken from that species, from a pair of populations separated by a large geographic distance, it would be impossible to determine whether the differentiation between them conformed to a continuous pattern, or some discrete pattern (perhaps due to a physical barrier to dispersal between them, or a behavioral or phenological or barrier to reproduction between them).

The second reason it may be difficult to disentangle continuous and discrete patterns of population genetic differentiation is that such a classification ("continuous" vs. "discrete") is artificial, and can rarely be applied cleanly in empirical systems. Even a group of isolated populations cannot be considered truly discrete units, as they will have diverged from shared ancestral populations, and stopped exchanging migrants, at different points in the past, and should therefore be characterized along a continuum of pairwise genetic relatedness, even though they represent 'discrete' evolutionary or management units.

However, there are many empirical systems in which the different, overlaid patterns of genetic differentiation can be well characterized as continuous or discrete, and, in such systems, it is vital to model both patterns simultaneously, to avoid confusing one for the other. Currently, there are no methods that can probabilistically place individuals into discrete population clusters while accounting for a background rate at which genetic differentiation accrues with distance between samples in space or time. This gap in methods for the study of population genetic structure frequently results in the mischaracterization of continuous differentiation as discrete (and vice versa), potentially inflating the number of inferred discrete population clusters

above what is biologically accurate. 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.

Introduction

Methods

Data The statistical framework of our approach is conceptually similar to Wasser et al. [2004] and Bradburd et al. [2016]. The genetic data modeled consist of allele frequencies at L unlinked, bi-allelic single nucleotide polymorphisms (SNPs) genotyped across M samples. The sample frequency at locus ℓ in sample m, $f_{m,\ell}$, 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 m. Following Bradburd et al. [2016], we standardize allele frequencies by subtracting off the mean and normalizing the variance. Specifically, we compute the standardized allele frequency at locus ℓ in sample m, $X_{m,\ell}$, as follows:

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

where $\hat{f}_{m,\ell}$ is the sample allele frequency at locus ℓ in sample m, $\bar{f}_{\ell}$ is the mean allele frequency across all samples at locus ℓ. With $\bar{f}'_{\ell}$ we make a slight adjustment to the mean to avoid extreme values:

$$\bar{f}_{\ell} = \frac{1}{M} \sum_{M} \hat{f}_{m,\ell} \tag{2}$$

$$\bar{f}'_{\ell} = \frac{1}{M+1} \left(\sum_{M} \hat{f}_{m,\ell} + 0.5 \right) \tag{3}$$

This standardization procedure gives each locus roughly unit variance, and makes the result invariant to which allele (major or minor) is chosen to be counted at each locus. We then calculate the sample standardized allele frequency covariance between all samples:

$$\widehat{\Omega} = \frac{1}{L} \hat{X} \hat{X}^T \tag{4}$$

Continuous and discrete differentiation We 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 draw admixture. 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:

$$F_{i,j} = \alpha_0 \left(\exp\left(-(\alpha_D D_{i,j})^{\alpha_2}\right) + \mu\right) \tag{5}$$

where $F_{i,j}$ is the covariance function within a cluster, $D_{i,j}$ 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 μ can be interpreted as the branch length connecting the population described by the spatial cluster 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, $w_i^{(k)}$, is the probability that an allele in sample i was derived from cluster k, and that sample's admixture proportions in each cluster must sum to one.

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

$$\Omega_{i,j} = \sum_K w_i^{(k)} w_j^{(k)} F_{i,j}^{(k)} \tag{6}$$

The within-sample variance, $\Omega_{i,i}$, follows the same form as Eqn. 6, with an additional component to describe both sampling noise from the finite number of chromosomes genotyped and sample-specific drift due to, e.g., inbreeding. Thus,

the form of the covariance generalized across all sample pairs is given by:

$$\Omega_{i,j} = \sum_{K} w_i^{(k)} w_j^{(k)} F_{i,j}^{(k)} + \delta_{i=j} \left(\frac{1}{S_i} \left(1 - \sum_{K} w_i^{(k)} w_j^{(k)} F_{i,j}^{(k)} \right) + \eta_i \right) \tag{7}$$

where $\delta_{i=j}$ is an indictor variable that evaluates to 1 when i is equal to j, and zero otherwise, S_i is the number of chromosomes genotyped in the ith sample (e.g., S = 2 for a diploid), and η_i describes the sample-specific drift or inbreeding.

Likelihood and inference If we assume that the standardized allele frequencies $\widehat{X}$ are multivariate normally distributed, their sample covariance will be Wishart distributed with degrees of freedom equal to L, the number of loci genotyped. We also assume that the observed loci are independent; linkage disequilibrium (LD) between loci will decrease the effective number of degrees of freedom. The likelihood of the sample standardized allele frequency covariance is given by

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

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, α, μ, η and on observed quantities D and S with the notation $\Omega(w, \alpha, \mu, \eta, D, S)$, we denote the posterior probability of the parameters as:

$$P\left(w, \alpha, \mu, \eta \mid \widehat{\Omega}, L\right) \propto P\left(\widehat{\Omega} \mid \Omega(w, \alpha, \mu, \eta, D, S)\right) P(w)P(\alpha)P(\mu)P(\eta) \tag{9}$$

The priors, $P(w), P(\alpha), P(\mu), P(\eta)$, 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 [Carpenter, 2015, Hoffman and Gelman, 2014, Stan Development Team, 2015, 2016] to estimate the posterior distribution on the parameters. We also present an R package [R Core Team, 2015] called `geoStructure` 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.

References

Gideon S. Bradburd, Peter L. Ralph, and Graham M. Coop. A spatial framework for understanding population structure and admixture. *PLoS Genet*, 12(1):1–38, 01 2016.

Bob Carpenter. Stan: A probabilistic programming language. *Journal of Statistical Software*, 2015.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

Stan Development Team. Stan: A c++ library for probability and sampling, version 2.10.0, 2015.

Stan Development Team. Rstan: the r interface to stan, version 2.10.1, 2016.

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

Supplementary Materials

1 Appendix

1.1 Priors

1.2 Mean-centering procedure and implications