Supplemental Information for:

Interpreting principal components analyses of spatial population genetic variation

John Novembre[1] and Matthew Stephens[1,2]

[1]Department of Human Genetics, University of Chicago, IL 60637
[2]Department of Statistics, University of Chicago, IL 60637

Supplementary Methods

Principal components analysis of genetic data

PCA can be conducted on population genetic data in at least two major ways; either in a population-based or individual-based manner. The two approaches differ in how the input data matrix for PCA is defined, but once this matrix is defined, the steps are identical. Let G represent the input data matrix, and let it have n rows and m columns and let M represent a renormalized version of G, also with n rows and m columns.

For our population-based results, we used a data matrix with one row for each of n pre-defined populations, and one column for each of m bi-allelic loci. The element $G(i, j)$ is initially set to the frequency of the derived allele at the jth locus in the ith population. Following Cavalli-Sforza et al, we normalize G only by the column means. In this case, the matrix M is defined by

$$M(i, j) = G(i, j) - \mu(j)$$

where $\mu(j)$ is the mean of the jth column of G.

For the individual-based approach, there are no pre-defined populations, and the data matrix G has one row for each of n individuals and one column for each of m loci. The element $G(i, j)$ is then set to an integer representing the number of copies of the derived allele found in individual i at locus j (so that for autosomal data, the entries are 0,1, or 2). Let $\mu(j)$ and $\sigma(j)$ be the mean and standard deviation, respectively, of the jth column of G. The normalized matrix M is then defined by:

$$M(i, j) = \frac{G(i, j) - \mu(j)}{\sigma(j)}.$$

This approach is similar to that of Patterson et al [1] except they divide by an alternative estimator of the standard deviation,$\sqrt{p(j)(1 - p(j)}$, where $p(j)$ is the maximum likelihood estimate of the allele frequency for the jth SNP (i.e. $p(j) = \mu(j)/2$).

We make a slight variation to the individual-based approach for autosomal dominant markers (e.g. AFLPs). For autosomal dominant markers, we define an indicator variable $I(i, j)$, that takes the value 0 (or 1) if an AFLP band is absent (or present) for marker j in individual i. We initially set $G(i, j)$ equal to $I(i, j)$. We then compute the column means $\mu(j)$ and column standard deviations $\sigma(j)$. The normalized matrix M is then defined by:

$$M(i, j) = \frac{G(i, j) - \mu(j)}{\sigma(j)}.$$

Given the renormalized $n \times m$ data matrix M, the first step of PCA is to compute the $n \times n$ sample covariance matrix X among the units of interest (i.e. populations or individuals):

$$X = \frac{1}{n} M M'$$

where M denotes the transpose of M. Some examples of covariance matrices from idealized and simulated data are given in fig. S3.

The second step is to compute the eigenvectors of X. The kth eigenvector will be of length n with one entry for each individual/population. When geographical coordinates are available for each individual/population, each eigenvector entry is then naturally associated with a particular geographical coordinate, and a contour plot or heat map can be made to show how the eigenvector values vary across geographical space (Specifically, the kth PC-map is a heat map showing how the entries in the kth eigenvector vary across geographical space). When geographical coordinates are not available, a common visualization strategy is to plot the corresponding elements of one eigenvector against another, producing biplots as in fig. 2 for example. In practice, we computed the eigenvectors by applying the `prcomp` function of the `R` statistical package[2] with the appropriate `center` and `scale` arguments to achieve the normalizations described above.

Simulation details

To generate both individual and population-based data (as described in the Methods Summary), we used Hudson's `ms` software [3]. To simulate L polymorphic loci, we independently simulate L loci with the number of segregating sites per locus fixed to 1.

For the results of the two-dimensional population-based simulations shown in fig. 1, $n = 100$, $D_s = 15 \times 15$, $D = 31 \times 31$, $L = 500$, $4Nm = 0.1$. For the one-dimensional individual-based results of fig. 2, the parameters used were $n = 50$, $D = 100$, $L = 1000$, $4Nm = 1$.

We also simulated data using an alternative Gaussian-process-based spatial model for allele frequencies (originally described in [4]). We observed similar sinusoidal patterns in PCs computed from these data to those we observed in the explicit population genetic simulations using `ms` (results not shown). This is as predicted by theory (see main paper) as both models induce a spatial covariance structure among sampled individuals, with genetic similarity tending to decay with distance.

Analysis of *Phylloscopus trochiloides* (Greenish warblers) data

To examine the behavior of PCA of spatial data in an empirical context, we applied PCA to a previously published dataset [5] of amplified fragment length polymorphism (AFLP) data from greenish warblers (*Phylloscopus trochiloides*). Greenish warblers are of broader interest because they are a well-documented example of a ring species complex [6, 5]. Greenish warblers are most abundant in western and eastern Siberia. Where these two main populations overlap geographically, there is no mating between the two, yet the two populations are connected by gene flow via a narrow band of populations to the south that are arranged in a ring around the Tibetan plateau. While this species is distributed along a ring, because the warblers do not interbreed across the top of the ring, greenish warblers can be thought of as inhabiting a one-dimensional habitat. Thus for our purposes, greenish warblers are an interesting test case for our results regarding PCA in one-dimensional habitats.

The data collected by Irwin et al [5] consist of 62 AFLP markers typed on 105 individuals from 26 geographic sites. AFLPs are dominant markers, so each marker is typed for presence or absence. Irwin et al also conducted PCA on this data and plotted PC1 against distance along the ring; however our analysis differs in a few ways. We normalized each AFLP variable to have a standard unit variance before applying PCA (similar to [7]), we excluded five sites that are outside of the central ring (GT,FN,NZ,TU, and YK), and we calculated position along the ring in a different manner. To calculate position of each individual along the ring we fit an ellipse to the geographic distribution of sampling sites and then mapped each site onto the ellipse, and took the distance from an arbitrary point on the ellipse as an indicator of position (see below for more detail). Positive distances correspond to sites on the east side of the ring-shaped habitat, and negative distances to the distances on the west side of the ring-shaped habitat.

If covariance between each individual's AFLP markers decays with distance and sampling error is small, we expect sinusoidal patterns would emerge in the PCA results. Indeed, biplots for PC1 and PC2 (fig. 9) revealed the horseshoe-shaped Lissajous pattern that is expected when plotting a roughly linear gradient for PC1 against a quadratic form for PC2 (as in fig. 2). In agreement with our simulation results (eg, fig. 2) PC1 is directly related to location within the one-dimensional habitat and PC2 is related to distance from the center of the 1-dimensional habitat (fig. 9). These patterns were also observed if we treated each sampling location as a population and used population-based PCA on the data (results not shown).

These results are consistent with arguments made by Irwin et al. regarding the presence of isolation-by-distance in this system. PC3 (fig. 9) does not have a clear relationship to geography, rather it appears to account mainly for variation among individuals sampled from Eastern Siberia. Subsequent PCs (data not shown) appear noisy with no clear geographical relationship. This is consistent with a result we found in simulations, that for smaller datasets (in terms of both number of loci and individuals) the higher principal components are typically too noisy to recognize the sinusoidal-like patterns.

Details of fitting an ellipse to the *Phylloscopus trochiloides* (Greenish warblers) sample site data

To fit the ellipse, recall that for an ellipse whose axes are perpendicular to the coordinate axes, any point (x, y) on the ellipse satisfies the equation

$$(x-a)^2/b^2 + (y-c)^2/d^2 = 1.$$

Let $\phi \in (-\frac{\pi}{2}, \frac{\pi}{2})$ define the minimum angle between the axes of an observed ellipse and one that has its axes perpendicular to the coordinate axes. We fit the ellipse by finding the values of a,b,c,d, and ϕ that jointly minimize the equation

$$\sum_{i=1}^{21} [(x_i^{(r)} - a)^2/b^2 + (y_i^{(r)} - c)^2/d^2 - 1]^2$$

where $x_i^{(r)} = x_i \cos(\phi) - y_i \sin(\phi)$, $y_i^{(r)} = x_i \sin(\phi) + y_i \cos(\phi)$, and x_i and y_i are the latitude and longitude of the ith sampled site. Points on the resulting ellipse are given by:

$$\hat{x}(t) = (a + b\cos(t))\cos(-\phi) - (c + d\sin(t))\sin(-\phi)$$

and

$$\hat{y}(t) = (a + b\cos(t))\sin(-\phi) + (c + d\sin(t))\cos(-\phi)$$

where $t \in (0, 2\pi)$. We then mapped each sampling site onto the ellipse by finding the nearest point on the ellipse to each sampling site [i.e. for the ith sampled site with raw coordinates (x_i, y_i), we assign it to the transformed coordinates $(\hat{x}(t_{min}), \hat{y}(t_{min}))$ where $t_{min} = \text{argmin}_{t\in(0,2\pi)}[(x_i - \hat{x}(t))^2 + (y_i - \hat{y}(t))^2]$. Finally, to calculate a position along the ring, we calculate the arc length along the ellipse (i.e. the elliptic integral) between the transformed coordinates of each sample site and the arbitrary point $(\hat{x}(1.4\pi), \hat{y}(1.4\pi))$, which is a point on the ellipse in the southern part of the Greenish warbler range.

Color-coding of Cavalli-Sforza et al's original PCA maps

Figures 3.11.1-3.11.4, 4.17.1-4.17.5, and 5.11.1-5.11.4 from the "History and Geography of Human Genes" were scanned in using Adobe Photoshop software. Adobe Illustrator CS2's LiveTrace feature was used to create vector-based representations of each scanned image. Some minor errors in original plots are introduced by this step but they are only very fine-scale errors in small regions of the graphs. The hash marks that denote contour plot level intensities in the original images were deleted manually using the Selection tool. The LivePaint feature was used to fill each contour region with colors meant to represent the eight levels used in Cavalli-Sforza's original plots. Specifically, we used a CMYK color model with the C and K components set to 100%, K set to 100%, and values of M that vary along a uniform interval between 0 and 100%. In five cases to make the similarity among PC plots more clear, the ordering of the levels was reversed from that in the original Cavalli-Sforza plot (i.e., Africa PC1 & PC4, Asia PC1, PC2, and PC5). Because PCs are arbitrary with respect to having a positive/negative sign, reversing the order of the levels does not represent a distortion of the original PCA results.

References

[1] Patterson, N., Price, A. L., and Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

[2] R Development Core Team. *R: A Language Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2007).

[3] Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–8 (2002).

[4] Wasser, S. K., Shedlock, A. M., Comstock, K., Ostrander, E. A., Mutayoba, B., and Stephens, M. Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. *Proc Natl Acad Sci U S A* **101**, 14847–52 (2004).

[5] Irwin, D. E., Bensch, S., Irwin, J. H., and Price, T. D. Speciation by distance in a ring species. *Science* **307**, 414–6 (2005).

[6] Irwin, D. E., Bensch, S., and Price, T. D. Speciation in a ring. *Nature* **409**, 333–7 (2001).

[7] Patterson, N., Price, A., and Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

Supplementary Note

Examples of sinusoidal PCs from spatial data

Here we provide more detail and references regarding a few idealized scenarios where sinusoidal PCs arise from spatial data. Consider a situation where the covariance between two populations depends only on the geographic distance between them, and assume that sufficient genetic data (loci/alleles) are available to accurately estimate this covariance structure. Then:

1. If populations are regularly spaced on a line, then their covariance matrix has a "Toeplitz" structure. A Toeplitz matrix is a matrix whose (i, j)th element X_{ij} depends only on $(j - i)$. (e.g. fig. S3b), and the eigenvectors of any (large) Toeplitz matrix are known to be closely approximated by sinusoidal functions [1]. A well-studied special case occurs when the covariance between populations decays exponentially with distance (specifically, each element X_{ij} of the matrix equals $\rho^{|i-j|}$ for some constant ρ), where the eigenvectors are approximately the columns of the discrete cosine transform (DCT) matrix [2, 3].

2. If populations are regularly spaced around a circle, then their covariance matrix has a "circulant" structure. A circulant matrix is a Toeplitz matrix in which each row is obtained from the row above it by a right cyclic shift; that is, by moving the last element of the row to the start of the row (e.g. fig. S3b). The eigenvectors of any circulant matrix are the columns of the discrete Fourier transform matrix [1], which are sinusoidal functions of increasing frequency.

3. If populations are located on a 2-dimensional regular grid, as in Cavalli-Sforza et al's analyses, the covariance matrix has a "block Toeplitz with Toeplitz blocks" form (e.g. fig. S3b), with eigenvectors that are approximated by the two-dimensional DCT commonly used in image compression[3]. The first two eigenvectors are commonly two orthogonal gradients, and the next two have a "saddle" and a "mound" shape (fig. 1). Higher order eigenvectors relate to 2-dimensional sinusoidal functions of increasing frequency (fig. S4).

Further examples of these and related results can be found in various sources. For instance, in time-series analysis, where problems often arise that are analogous to analysis of one-dimensional spatial data taken at regular intervals, it has long been recognized that PCs are closely approximated by the columns of the discrete Fourier transform matrix [4]. In climatology, where PCs are known as empirical orthogonal functions (EOFs), the interpretation of PC-maps has recieved extensive critical attention [5, 6, 7, 8]. In image compression, the 2-D DCT is central to the popular JPEG compression algorithm, and much of its efficacy is due to the fact that the 2-D DCT basis functions so closely approximate the PCA basis (also known as the Karhunen-Loeve basis) without having the burden to compression of having basis vectors that are specific to each dataset (as in PCA)[3].

Selection of PCs in controlling for population structure

One practical issue regarding PCA-based approaches to controlling for population structure in association studies is deciding which PCs to use. Although in simulations for a discrete 2-population

model Price et al [9] found results to be relatively robust to which PCs are used, in general omitting relevant PCs may fail to fully control for structure (e.g. produce an elevated type I error), whereas including irrelevant PCs would be expected to reduce power. One suggestion in [9] is to select PCs based on the"significance" of their eigenvalue [10]. In spatially continuous populations, given enough data, we expect the number of significant eigenvalues to be large. This is because individuals sampled from a continuous population can be thought of as being drawn from a large number of discrete subpopulations exchanging migrants, and for discrete population models, given enough data, the number of "significant" eigenvalues/PCs is one less than the number of subpopulations [10]. (In practice, limits on available amounts of data would be expected to yield fewer significant eigenvalues.) For the example shown in fig. 2 (a sample of 1000 SNPs from 50 individuals from a linear set of 100 demes with effective migration rate $4Nm = 1$), we found using the method of [10] that 12 eigenvalues (of a possible maximum of 49) are "significant" at $p < 0.05$ (fig. S2a). If some of these significant PCs are not correlated with phenotype (once other PCs have been controlled for) then controlling for them is unnecessary, and may reduce power. This suggests that the problem of appropriately choosing the number of PCs may warrant further consideration, and we suggest that an attractive solution to this problem should involve considering which PCs are correlated with phenotype. (Forwards or backwards stepwise regression would be one conventional approach that might achieve this, although preferable approaches may exist.) Further, as mentioned in the main paper, it may be helpful to consider including non-linear functions of early PCs (e.g. quadratic or higher-order polynomial terms, possibly with interactions) as covariates in the regression, as well as, or instead of, later PCs (again, with choice of which non-linear functions and interactions to include being made with reference to the phenotype).

References

[1] Gray, R. M. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory* **2**, 155–239 (2006).

[2] Ahmed, N., Natarajan, T., and Rao, K. R. Discrete cosine transform. *IEEE Transactions on Computers* **C-23**, 90–93 (1974).

[3] Strang, G. The discrete cosine transform. *SIAM Review* **41**, 135–147 (1999).

[4] Brillinger, D. R. *Time Series: Data Analysis and Theory.* Holt, Rinehart, and Winston, (1975).

[5] Richman, M. B. Rotation of principal components. *Journal of Climatology* **6**, 293–335 (1986).

[6] Jolliffe, I. *Principal Component Analysis.* Springer Series in Statistics. Springer, (1986).

[7] Preisendorfer, R. *Principal Component Analysis in Meteorology and Oceanography.* Number 17 in Developments in Atmospheric Science. Elsevier, Amsterdam, (1988).

[8] Richman, M. Comments on: 'the effect of domain shape on principal components analyses'. *International Journal of Climatology* **13**, 203–218 (1993).

[9] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–9 (2006).

[10] Patterson, N., Price, A., and Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

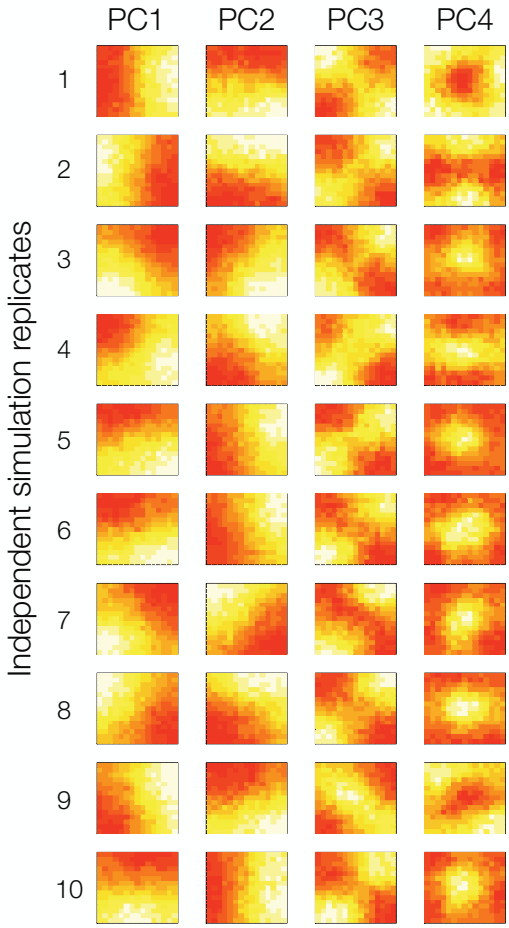
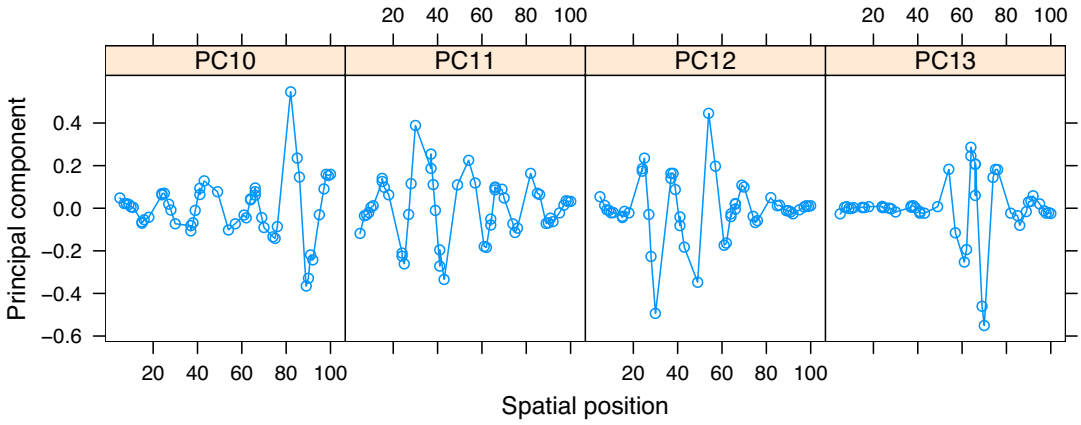


Figure S1: **Plots of the first four PC-maps for 10 independent simulation replicates where there is constant homogeneous migration in a 2-dimensional habitat.** Parameters are the same as in fig. 1 of the main text, i.e. $4Nm = 0.1$, $D_s = 225$ (i.e., 15×15), $D = 961$ (i.e., 31×31), $n = 200$, and $L = 500$. Noteworthy features include: (1) the exact angle of the gradient in PC1 varies across runs but PC1 is consistently a gradient across the habitat and PC2 is consistently a perpendicular gradient to that of PC1. (2) PC3 is typically a saddle-like shape. (3) PC4 is typically a mound- or bowl-like shape (note: the sign of the PC is arbitrary, so whether one views the shape as a mound or bowl is arbitrary). (4) The order of PC patterns sometimes fluctuates: In replicate 2, PC4 has changed order with the PC-map that is typically expected as PC5, so that the mound-like shape is present in the PC5 map (not shown). This re-ordering of the PCs occurs more frequently when smaller numbers of individuals or loci are used (not shown). (5) In addition to the overall similarity of results across independent replicates, in many cases replicates show similarity in detail (e.g. PC1 gradients that are in the same direction). For instance, replicates 4,5,6, and 8 all show a "north-west / south-east" gradient in PC1 even though the individual histories of migration in each simulation are independent of one another. Amongst all 45 possible pairwise comparisons ~10 show roughly equivalent patterns for PC1 and PC2 [e.g. pairs (1-2),(3-7),(3-9),(4-5),(4-6),(4-8),(5-6),(5-8),(6-8),(7-9), and recalling that the sign of PC values is arbitrary so that two patterns that are opposite in sign are appropriately considered to be equivalent]

a



b

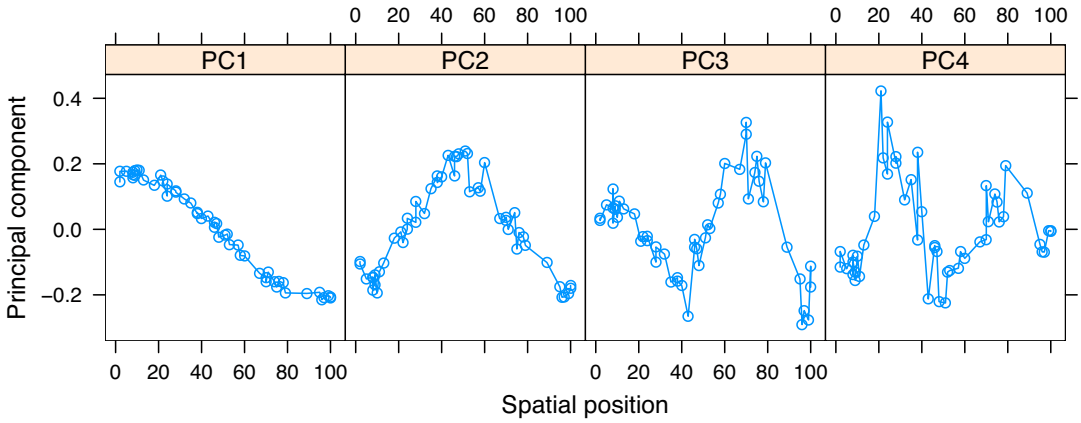


Figure S2: **(a) One-dimensional PC-maps for PCs 10-13 for the same case as in fig. 2 in the main text.** PC12 is the last "significant" axis of variation according to the method of [7]. **(b) One-dimensional PC-maps for the same case as in fig. 2 in the main text, but with the effective migration parameter** $4Nm = 100$ **rather than** $4Nm = 1$**.** With increasing effective migration rates, the sinusoidal patterns become more noisy, particularly for higher PCs (e.g. PC4 here). In both cases, additional parameters for these individual-based simulations are: $n = 50$, $D = 100$, $n = 1$, and $L = 1000$.

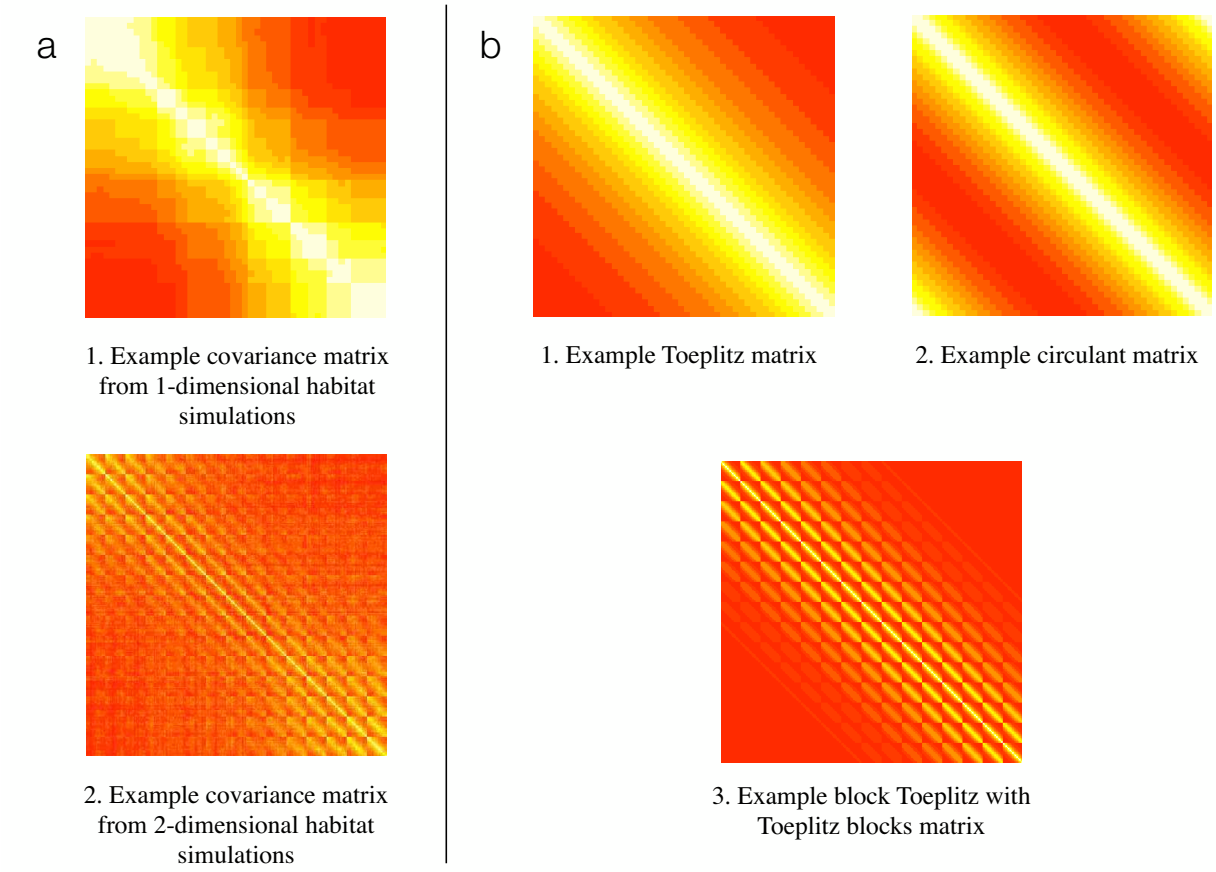


Figure S3: **Examples of the major classes of matrices referred to in the main text.** The matrices are depicted by coloring each element of each matrix in proportion to the magnitude of the value in the element, where whiter colors represent larger values. **Panel a: Examples of sample covariance matrices from simulated data.** For (a1) the rows of the covariance matrix are ordered by the geographic position of each individual and the simulated data are from the simulations shown in fig. 2 of the main text. The decrease in values as one moves away from the matrix diagonal reflects how covariance decays with distance between individuals (note though that the data also show a boundary effect that increases covariance among individuals near either end of the habitat). For (a2) the rows of the covariance matrix are also ordered by the geographic position of the individuals in the 2-d habitat (such that individuals are ordered from "west" to "east" and then from "north" to "south"). Specifically, this covariance matrix corresponds to the simulated data used in fig. 1 of the main text and it also shows a general decay of covariance with distance. **Panel B: Structured matrices that arise from idealized scenarios (see main text).** Theoretical results presented in the main text relate to the example toeplitz (b1), circulant (b2), and block Toeplitz with Toeplitz block matrices (b3). Of particular importance is how (a1) shows a similar structure to a Toeplitz matrix (b1) for which theoretical results are available and likewise (a2) shows a similar structure to a block Toeplitz with Toeplitz blocks matrix (b3).

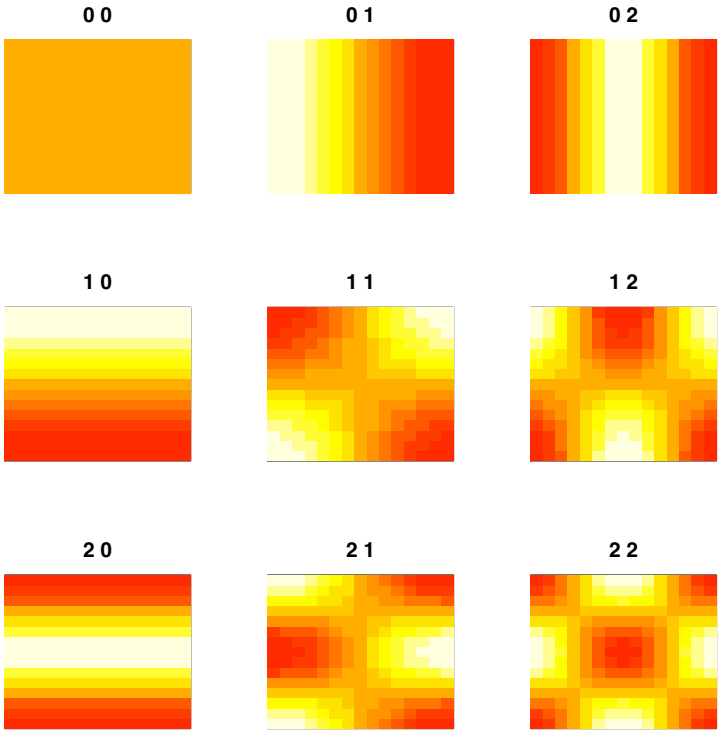


Figure S4: **Plots of 9 basis functions of the two-dimensional Discrete Cosine Transform (DCT) for 15×15 sample points.** The i, jth element of each plot is equal to $\cos(\frac{2\pi(2i+1)u}{2 \cdot 15}) \cos(\frac{2\pi(2j+1)v}{2 \cdot 15})$, where u and v are given as an ordered pair above the image. To obtain the complete set of 15^2 basis functions, one must take the corresponding plots for all possible ordered pairs of $u = 0 \ldots, 14$ and $v = 0 \ldots, 14$.

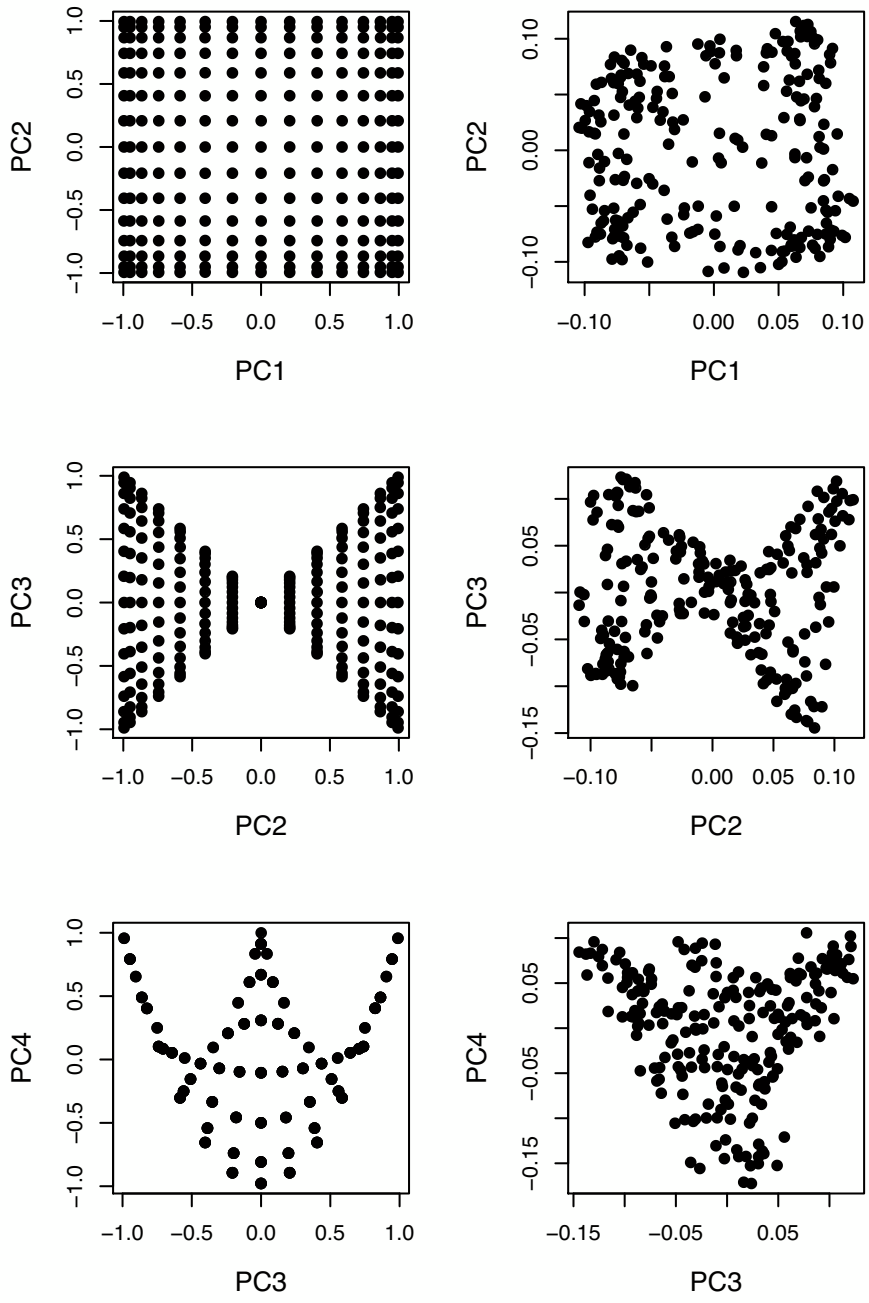


Figure S5: **Example of biplots of PCs from 2-dimensional spatial data.** The left-hand column contains biplots of the 4 idealized PCs expected from the 2-D DCT. The right-hand column contains biplots of the 4 observed PCs from data from a stepping-stone model simulation (same simulated data as in fig. 1 of main text, i.e. $4Nm = 0.1$, $D_s = 225$ (i.e., 15×15), $D = 961$ (i.e., 31×31), $n = 200$, and $L = 500$.)

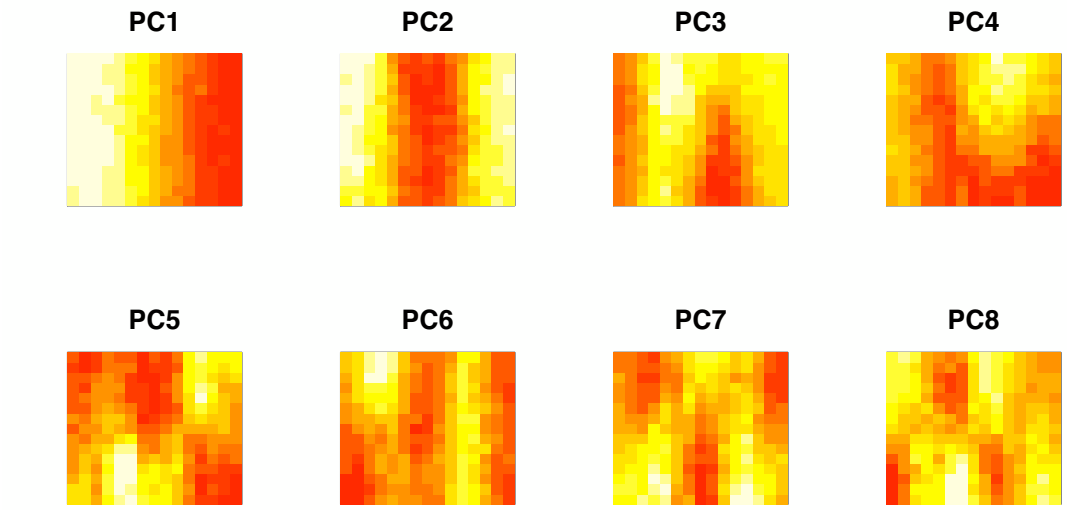
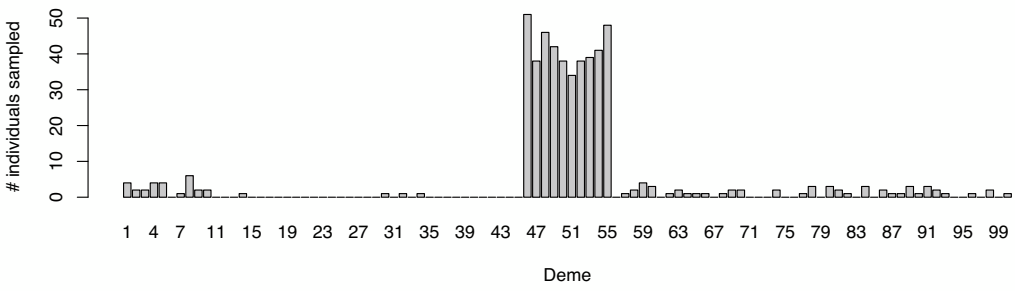
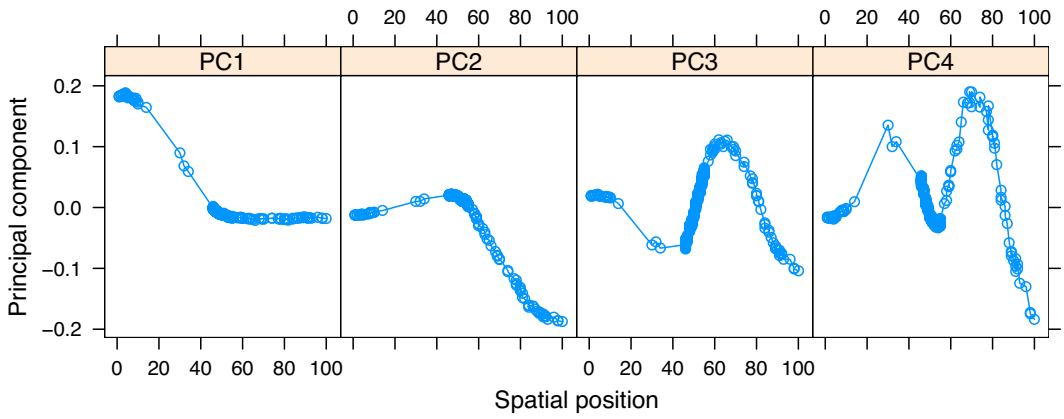


Figure S6: **Plots of PC-maps for PC1-PC8 from a two-dimensional stepping-stone simulation where $4Nm = 0.1$ in one dimension ("east-west") and $4Nm = 1$ in the other ("north-south").** The PCs no longer show the four canonical shapes, but they still have clear sinusoidal patterns. For example PC2 is analogous to the DCT basis function with $u = 0$,$v = 2$ (fig. S4). Additional parameters for these simulations are: $D_s = 225$ (i.e., 15×15), $D = 961$ (i.e., 31×31), $n = 200$, and $L = 500$.

a



b



c

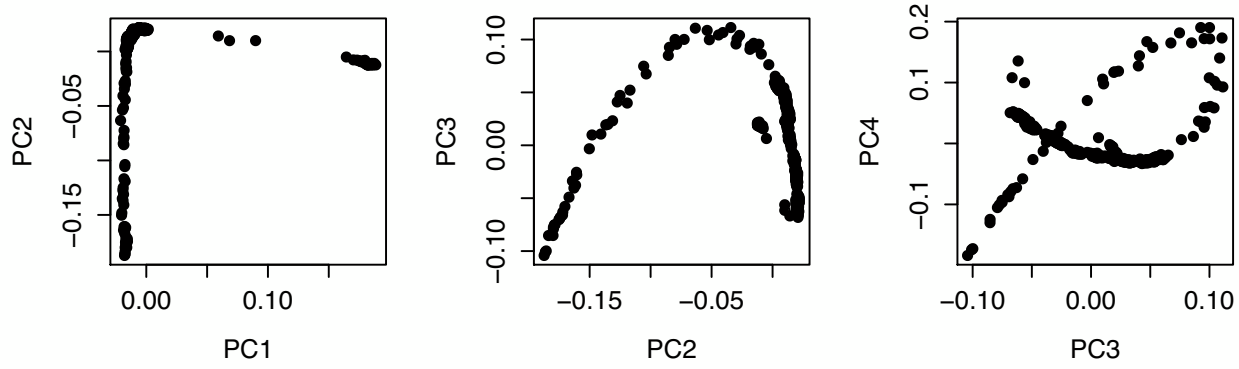


Figure S7: **An example of the distortion of idealized PCs due to biased spatial sampling.** 500 individuals are sampled from a habitat of 100 demes arranged along in a line and genotyped at 500 polymorphic sites. The effective migration parameter is set to $4Nm = 1$. The sampling distribution (a) is biased towards sampling individuals in the center of the habitat and then sampling out to the edges of the habitat but with an added bias towards sampling one end of the habitat slightly more than the other. The resulting PCs [shown as PC-maps in (b) and as bi-plots in (c)] are "distortions" of the PCs found from samples drawn uniformly from the habitat (fig. 2).

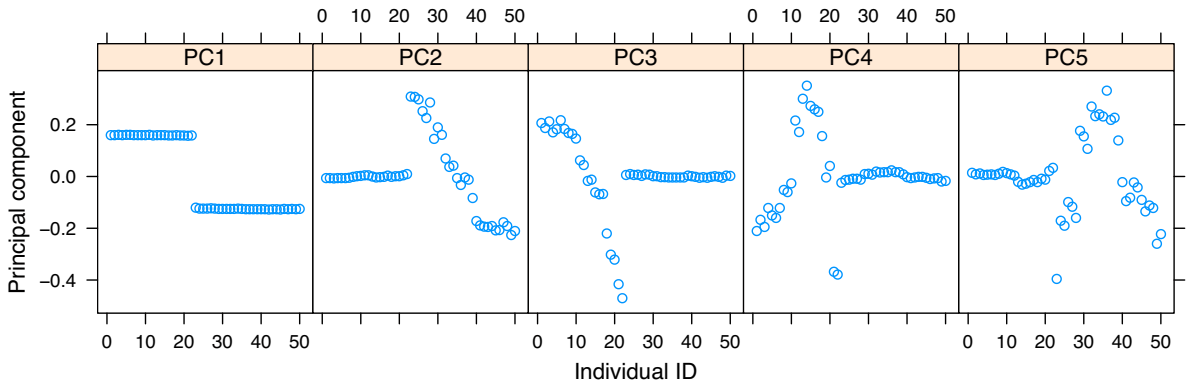
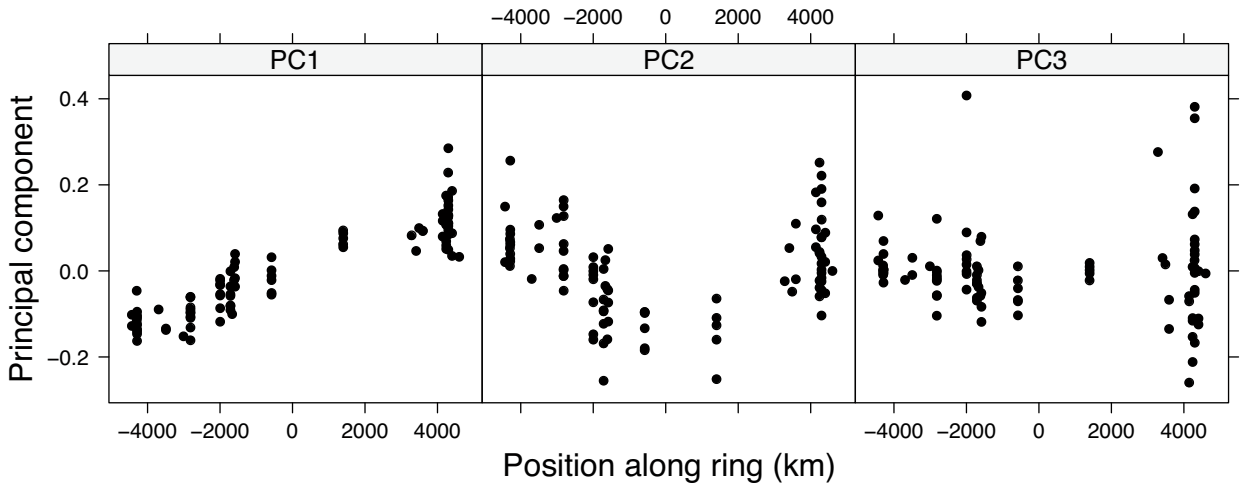


Figure S8: **An example of PCs from a sample with discrete and continuous patterns of variation.** Fifty individuals were drawn at random from one of two sampling areas within a habitat consisting of 100 demes arranged in a line and genotyped at 1000 polymorphic loci. The individual IDs reflect the order of individuals habitat. The first 22 individuals are from area 1 (the first ten demes in the linear array of demes) and the last 28 are from region 2 (the last ten demes in the linear array of demes). PC1 separates out individuals of the 2 sampling areas. PC2 and PC3 reflects the "linear" component within area 2 and area 1, respectively. PC4 and PC5 are the "distance from the center shape" for area 1 and area 2. For the simulations, $4Nm = 1$.

a



b

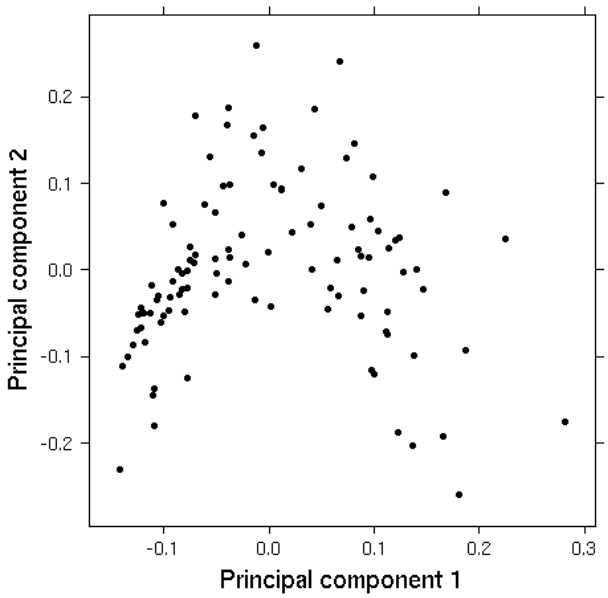


Figure S9: **PCA results for the *P. trochiloides* data.** (a) The first three one-dimensional PC-maps for the *P. trochiloides* data. Geographic position in this case is equivalent to the position along the ring-shaped habitat. (b) PC1 vs PC2.