Published in final edited form as:
Nat Genet. 2008 May ; 40(5): 646–649. doi:10.1038/ng.139.

Interpreting principal component analyses of spatial population genetic variation

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

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

[2]Department of Statistics, University of Chicago, Chicago, IL 60637

Abstract

Nearly thirty years ago, Cavalli-Sforza et al pioneered the use of principal components analysis (PCA) to summarise data on variation in human gene frequencies across continental regions [1]. Cavalli-Sforza et al produced maps representing each Principal Component (PC), and found these maps exhibited highly distinctive patterns, including gradients and sinusoidal waves. They interpreted these patterns as resulting from specific migration events, such as the migration of agriculturalists out of the Near East [1, 2, 3]. Cavalli-Sforza et al's results have been highly influential (e.g. [4]), and controversial [5, 6, 7, 8, 9], and PCA has become heavily used in population genetics (e.g. [10, 11, 12, 13]). Despite its widespread use, the behavior of PCA with data exhibiting continuous spatial variation, such as might exist within human continental groups, has been little studied. Here, using empirical and theoretical approaches, we find that the distinctive patterns observed by Cavalli-Sforza et al resemble sinusoidal mathematical artifacts that arise generally when PCA is applied to spatial data, implying that the patterns are not necessarily due to population movements. Our results aid the interpretation of PCA results from large-scale analyses of human genetic variation, and suggest that PCA will be helpful in correcting for continuous population structure in association studies.

Cavalli-Sforza et al's 1994 "The History and Geography of Human Genes" [3] stands as a classic text in human population genetics, synthesizing a decades-long survey of human genetic variation. These ground-breaking datasets stimulated development of methods that are now widely used, such as the application of Principal Components Analysis (PCA) to population genetic variation. In essence, Cavalli-Sforza et al collected count data for each genetic variant ("allele") at numerous genetic loci from population samples at many geographic locations, and produced for each allele an *allele-frequency map*, a spatially-interpolated map representing variation in allele frequency across space. They then used PCA, a general method for obtaining low-dimensional summaries of high-dimensional data, to distill the many allele-frequency maps into a smaller number of "synthetic maps", which for brevity we refer to as *PC-maps*. Intuitively, the first few PC-maps summarize the allele-frequency maps, in that each allele-frequency map can be well approximated by a linear

Author contributions
J.N. and M.S. jointly designed the analyses and interpreted results. J.N. performed the analyses and researched the applications of PCA in other fields. J.N. and M.S. wrote the paper.

superposition of the PC-maps (indeed one view of PCA is that it aims to produce PC-maps with this property).

Figure 1 shows PC-maps for Asia, Europe and Africa from [2, 3]. In interpreting these maps, Cavalli-Sforza and colleagues suggest that “if there is a radiation of circular or elliptic lines from a specific area, a [population] expansion is a possible explanation; and its place of origin must be the center of the radiation.” (p. 295 [3]). They also suggest centripetal population movements as an alternative explanation. Examples of their explanations for the European PC-maps in Figure 1 include: expansion of agriculturalists out of the Near East (Europe PC1); migrations of Mongoloid Uralic speakers from northwestern Asia (Europe PC2); migration of the carriers of the proto-Indo-European Kurgan culture in Europe (Europe PC3); and an expansion from Greece (Europe PC4).

Since the basis for Cavalli-Sforza et al’s interpretive guidelines is unclear, we performed simulations to investigate whether such specific migration events are necessary to explain the observed patterns. Specifically, we performed PCA on data simulated under population genetics models without range expansions, assuming a constant homogeneous short-range migration process across both time and (2-dimensional) space. The results showed highly distinctive structure. For example, the first two PC-maps exhibit large-scale orthogonal gradients, and the next two exhibit “saddle” and “mound” patterns (fig. 1). The same four basic patterns occurred consistently in the first few PC-maps across multiple simulations, although not always in the same order (fig. S1). Results for the analogous 1-dimensional habitat setting are even more structured, resembling sinusoidal functions of increasing frequency (fig. 2B, fig. S2). Thus PC-maps exhibit local peaks and troughs *even when underlying migration patterns are homogeneous across time and space*. This implies that interpretation of local features of the PC-maps as necessarily reflecting specific localized historical migration events is inappropriate. Furthermore, the first few PC-maps obtained by Cavalli-Sforza et al in Asia, Europe, and Africa show, with minor exceptions, highly structured patterns strikingly similar to those from our simulations (fig. 1,fig. S1).

In fact, these highly structured patterns are mathematical artifacts that arise generally when PCA is applied to data exhibiting a “spatial” covariance structure (i.e. where covariances between locations tend to decay with distance, fig. S3A). In population genetic data, a spatial covariance structure will arise when genetic similarity tends to decrease with distance, which would be expected under a wide range of demographic scenarios, including both equilibrium isolation-by-distance models [14] and non-equilibrium models involving population expansions [15, 16]. For intuition into why sinusoidal patterns emerge in PC-maps, we note that the common description of PCA, as searching for directions that explain the most variance in the data, is perhaps not especially helpful here, as these directions are in a very high dimensional mathematical space and not geographic space. Instead recall the property of PC-maps mentioned above: it should be possible to accurately approximate any of the allele-frequency maps using a linear superposition of the first few PC-maps. PC-maps that contain sinusoidal functions of increasing frequency accomplish this in a sensible way: the low-frequency patterns in the first few PC-maps allow for a coarse approximation by reflecting changes in allele frequencies across large spatial scales, while higher-frequency patterns in subsequent PC-maps allow for refinement of the coarse approximations by capturing finer-scale changes.

Mathematically, the explanation for the highly structured patterns is that covariance matrices from spatial data have eigenvectors related to sine waves of increasing frequency [17, 18, 19], and PC-maps are direct visual representations of such eigenvectors (see Supplemental Material). To give three specific examples, 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[1](e.g. fig. S3B), and the eigenvectors of any (large) Toeplitz matrix are known to be closely approximated by sinusoidal functions [19]. A well-studied special case occurs when the covariance between populations decays exponentially with distance[2], where the eigenvectors are approximately the columns of the discrete cosine transform (DCT) matrix [17, 18].
2. If populations are regularly spaced around a circle, then their covariance matrix has a "circulant" structure[3] (e.g. fig. S3B). The eigenvectors of any circulant matrix are the columns of the discrete Fourier transform matrix [19], 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[4] [18]. 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).

We note also that 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 [20].

Although Cavalli-Sforza et al performed PCA on population allele frequency estimates, PCA can also be applied to individual genotype data [12]. The results above apply equally to this context: just as spatial covariance among populations will produce sinusoidal-like PC-maps, so will spatial covariance among individuals. However, in this setting geographical information may not be available for each individual, making PC-maps difficult to produce. Instead PCA results are commonly visualized by producing biplots of one PC against

[1] A Toeplitz matrix is a matrix whose (i, j)th element X_{ij} depends only on $(j - i)$.

[2] Specifically, each element X_{ij} of the matrix equals $\rho^{|i-j|}$ for some constant ρ.

[3] 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.

[4] Indeed, the 2-D DCT is central to the popular JPEG image 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 basis vectors that are specific to each dataset (as in PCA).

another. Under uniform sampling from a 1-dimensional habitat with homogeneous migration, this results in biplots of sinusoidal functions of differing frequencies, producing characteristic patterns known as Lissajous curves [21, 22] (fig. 2C). In particular, the first of these biplots shows a pattern known as the "horseshoe effect" (e.g. [23]). For the analogous 2-dimensional setting, because PC1 and PC2 are typically orthogonal gradients, a biplot of the first 2 PCs essentially reproduces the geographic arrangement of sampled individuals (explaining PCA results on genetic variation in *Arabidopsis* [24] for example), and biplots involving later PCs have intricate patterns analogous to Lissajous curves (fig. S5).

Since the above empirical and theoretical results involve unrealistically simplistic scenarios, we assessed robustness by examining PC-maps for more complex scenarios involving heterogeneous migration processes and irregular sampling of populations across space. Detailed features of the PC-maps were influenced by both factors. Changing the sampling scheme or details of migration can produce a range of continuous distortions of the idealized sinusoidal shapes. Since quantifying this effect is difficult, we instead provide several examples for illustration. Anisotropic migration (i.e. migration is not equal in all directions, fig. S6) and irregularly spaced populations (fig. S7) both distort the PC-maps, and change their order. The direction of the gradient in the first PC-map is influenced by habitat shape (e.g. in fig. 2, PC1 in Africa and Asia are both along the longer axis of the continent), as has also been noted in climatological data [25], and by migration patterns (e.g. under anisotropic migration in a square habitat the gradient in PC1 aligns with the axis of least migration, fig. S6). However, sinusoidal-like patterns consistently emerge. Even when sampling locations are highly clustered within the continuous habitat (a common sampling design in practice, because of logistical challenges to obtaining spatially uniform samples in many species, [26]), the first PCs separate out the clusters as if the sample were obtained from discrete sub-populations, and subsequent PCs show sinusoidal patterns within clusters (fig. S8).

We also examined how quantity of data affects PCA results. With limited data sinusoidal patterns can still emerge. For example, such patterns occur in PC1 and PC2 (figs. S9, S10) from only 62 amplified fragment length polymorphism (AFLP) markers typed in 105 individuals from the ring species complex of greenish warblers (*Phylloscopus trochiloides*, see supplementary text, [27]). However, limited data can lead to less well-defined (or entirely absent) sinusoidal patterns, particularly in higher PC-maps (e.g. PC3 for the same dataset; fig. S9). In general, amounts of data needed to recover sinusoidal patterns will depend on the strength of the population structure (e.g. the amount of differentiation among sampled populations). Thus, for a fixed number of loci, higher effective migration rates, which tend to reduce population structure, lead to less well-defined sinusoidal patterns (fig. S11).

In summary: i) when analyzing data with a spatial covariance structure, PCA produces highly structured results relating to sinusoidal functions of increasing frequency; ii) in as far as PCA results depend on the details of a particular data set, they are affected by factors in addition to the actual underlying spatial population structure, such as the distribution of sampling locations and the amount of data available. These conclusions are supported not only by the results we present here, but by extensive empirical results from other fields where PCA has been applied to spatial data, including pattern analysis of natural images [28], ecology [23] and climatology [25, 29, 30, 31].

Both features i) and ii) above limit the utility of PCA to draw inferences about underlying processes, a fact previously noted in climatology ([25, 29, 30, 31]). In particular, interpreting gradient and wave-like patterns in PC-maps as signatures of historical migration events is problematic because such patterns arise quite generally under a simple condition: that genetic similarity decays with distance. This finding has important implications for interpretation of PCA in population genetics generally, and for Cavalli-Sforza et al’s PCA analyses in particular. Not only is this simple condition likely to be satisfied under many possible models, but, since Cavalli-Sforza et al used spatial interpolation to estimate allele frequencies, their data could satisfy this condition even if absent in the underlying allele frequencies [6, 8]. (Indeed use of interpolation may partly explain the striking similarity between Cavalli-Sforza et al’s PC-maps and those predicted by theory, particularly in Asia where their analysis was based on fewer samples. That said, recent analyses of European data without interpolation also show perpendicular gradients in PC1 and PC2 [32].)

Regarding the specific question of a Neolithic expansion in Europe, we emphasize this paper is not about whether or not such an expansion occurred; a full consideration of this would require a synthesis of multiple types of evidence from many diverse sources (see review in [33] and e.g. [34, 35, 9, 36, 37, 38, 39]). It is true that the NW-SE slope of the PC1 gradient in Europe suggests that this may be the direction of greatest genetic variation in Europe (although a careful analysis would have to take account of the fact that other factors, including the shape of the continent, could also influence the slope direction). However, if a Neolithic expansion could explain this, it is but one of many possible explanations.

For another example of how our results aid interpretation of PCA, consider the data from Linz et al [13] who found that PC-maps from *Heliobactor pylori* show similar patterns to those in Cavalli-Sforza et al’s human data and who use this as part of an argument that genetic patterns of *H. pylori* reflect a shared migrational history with humans. There are good reasons to suspect genetic variation in *H. pylori* will have been influenced by human migrations. However our results show that similar patterns in PC-maps of two groups does not imply a shared migrationary history; indeed, if each group shows an underlying spatial covariance structure, then similar patterns will often occur in the top few PC-maps even if their histories are independent (e.g. fig. S1).

Despite its limitations for inferring underlying processes that have produced population structure, PCA is undoubtedly an extremely useful tool for investigating and summarising population structure, and we anticipate that it will play a prominent role in the analysis of ongoing genome-wide studies of human genetic variation. Results presented here provide a helpful context for evaluating PCA results, essentially providing a “null” expectation against which observed PCs may be compared and contrasted. On the one hand, a close correspondence between observed and expected PCs may suggest an underlying continuous spatial covariance structure. On the other hand, departures from this null may also be useful, perhaps pointing towards a more discrete ”cluster-like” population structure [12], or to other important structure in the data, such as genotyping error or regions of high linkage disequilibrium [40].

Finally, our results provide some intuitive support for the use of PCA to address the problem of spurious associations produced by population structure in genome-wide association studies [41, 11]. In essence, the problem is that if phenotype mean (or risk) varies among subpopulations, then alleles that have no mechanistic connection to phenotype, but differ in frequency among subpopulations, will be “spuriously” associated with phenotype [42]. Although this problem has been studied mostly in the context of discrete subpopulations, it applies also to continuous (e.g. spatial) variation [43]. One commonly used PCA-based solution [11] controls for stratification by including the first few PCs as covariates in a regression. In populations showing a discrete, cluster-like, structure the first few PCs typically separate out the clusters [12], and so this solution corresponds to allowing for phenotype mean to vary among subpopulations. Our work shows that, for spatially-continuous populations, the PCA-based approach is conceptually similar to modeling smooth geographical trends in phenotype mean, a recognized technique in spatial epidemiology [44]. For example, if PC1 and PC2 are orthogonal gradients in space, including them in a regression essentially controls for latitude and longitude, and allows for linear trends in phenotype across space; including higher-order PCs allows for more flexible spatial trends. While some practical issues remain, including how many PCs should be used (supplementary text) and how best to employ the PCs [24], this analogy between the PCA-based and spatial-statistics approaches gives an intuitively appealing justification for using PCA to control for spurious associations in spatially-structured populations. It also has two important practical advantages over simply using geographic information on each individual directly. First, PCA can be used even when geographic information is not available. Second, PCA will control for continuous population structure even when geographic position does not correlate well with genetic background (as is typical in the United States, for example).

Methods

Simulations

For our population-genetics simulations we assumed a model of D demes that are arranged in a regular square lattice (for two-dimensional habitat) or a line (one-dimensional habitat). Each deme has effective population size $2N$ gametes, and, backwards in time, in each generation, a proportion, m, of gametes swap places with an equal number of gametes in each neighboring deme (e.g. for the 2-d simulations, demes internal on the lattice have four neighbors; demes on the edge have three neighbors; demes in the corners have two neighbors). We assume the population has reached equilibrium (i.e, the population has been evolving in this way for a long time).

We applied PCA to both “population-based” data (as in Cavalli-Sforza et al [1, 2, 3]) and “individual-based” data (as in [12]). For generating population-based data sets, we sampled n individuals from D_S of the D demes, and simulate for each individual data at L independent, bi-allelic polymorphic loci. Assuming independence of loci corresponds to migration of alleles rather than of whole gametes. We experimented with different spatial arrangements of the D_S demes, but for the results shown here (figs. 1, S1,S3,S5,S6), we use a regular square lattice of $D_S = 15 \times 15$ demes embedded in a larger $D = 31 \times 31$ lattice of demes. Allele frequencies in each deme are stimated from the n sampled individuals in that deme, to create a $D_S \times L$ data matrix of allele frequency estimates. For the one-dimensional simulations we report individual-based, rather than population-based PCA. We sampled n diploid individuals randomly from the D demes, and the data matrix consists of an $n \times L$ genotype matrix. See Supporting Online Material for additional details.

Principal components analysis

To calculate principal components on our simulated data, we use bi-allelic loci and include only the frequency of one of the two alleles. To compute the PCs, we apply the prcomp function of the R statistical package [45] to a matrix based on allele counts. In accord with Cavalli-Sforza's method for creating PCA maps, we do not scale the allele frequencies when conducting population-based PCA; however our methods differ from Cavalli-Sforza et al's in that we applied PCA directly to the observed allele-frequency matrix rather than using allele frequencies spatialy interpolated on a dense grid. This avoids problems with interpolation altering underlying spatial covariance patterns [6]. For individual-based PCA we use an approach similar to that of Patterson et al [12], in that we scale the genotype values across individuals at each locus to have unit variance. For the analysis of AFLP data from greenish warblers, we coded each typed marker by using an indicator variable with 0 or 1 indicating the absence or presence of an AFLP band, respectively. We then normalized each indicator variable to have mean zero and unit variance before applying PCA (again similar to [12], see Supporting Online Material for more detail).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge G. Coop, M. Przeworski, Y. Guan, M. Barber, and C. Becquet for helpful comments on a draft of this manuscript and D. Witonsky for pointing out the connection to Lissajous figures. Funding for this research was provided by an NSF postdoctoral research fellowship in bioinformatics (JN) and NIH grant RO1 HG02585-01 (MS).

References

1. Menozzi P, Piazza A, Cavalli-Sforza L. Science. 1978; 201:786–792. [PubMed: 356262]
2. Cavalli-Sforza L, Menozzi P, Piazza A. Science. 1993; 259(5095):639–646. [PubMed: 8430313]
3. Cavalli-Sforza, LL.; Menozzi, P.; Piazza, A. The History and Geography of Human Genes. Princeton University Press; 1994.
4. Jobling, M.; Hurles, M.; Tyler-Smith, C. Human evolutionary genetics. Garland Science; 2004.
5. Rendine S, Piazza A, Cavalli-Sforza LL. The American Naturalist. 1986; 128:681–706.
6. Sokal RR, Oden NL, Thomson BA. Hum Biol. 1999; 71(1):1–13. [PubMed: 9972095]
7. Rendine S, Piazza A, Cavalli-Sforza LL. Hum Biol. 1999; 71:15–25.
8. Sokal RR, Oden NL, Thomson BA. Hum Biol. 1999; 71(3):447–453. [PubMed: 10380379]
9. Currat M, Excoffier L. Proc Biol Sci. 2005; 272(1564):679–688. [PubMed: 15870030]
10. Hanotte O, Bradley DG, Ochieng JW, Verjee Y, Hill EW, Rege JEO. Science. 2002; 296(5566): 336–339. [PubMed: 11951043]

11. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Nat Genet. 2006; 38(8): 904–909. [PubMed: 16862161]

12. Patterson N, Price A, Reich D. PLoS Genet. 2006; 2(12):e190. [PubMed: 17194218]

13. Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, Yamaoka Y, Graham DY, Perez-Trallero E, Wadstrom T, Suerbaum S, Achtman M. Nature. 2007; 445(7130):915–918. [PubMed: 17287725]

14. Rousset, F. Genetic Structure and Selection in Subdivided Populations (MPB-40) (Monographs in Population Biology). Princeton University Press; 2004.

15. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Proc Natl Acad Sci U S A. 2005; 102:15942–15947. [PubMed: 16243969]

16. Prugnolle F, Manica A, Balloux F. Curr Biol. 2005; 15(5):R159–R160. [PubMed: 15753023]

17. Ahmed N, Natarajan T, Rao KR. IEEE Transactions on Computers. 1974; C-23:90–93.

18. Strang G. SIAM Review. 1999; 41:135–147.

19. Gray RM. Foundations and Trends in Communications and Information Theory. 2006; 2:155–239.

20. Brillinger, DR. Time Series: Data Analysis and Theory. Holt, Rinehart, and Winston; 1975.

21. Lissajous J. Annales de Chimie et de Physique. 1857; 51

22. Freiberger, W., editor. The International Dictionary of Applied Mathematics. Princeton, NJ: D. Van Norstrand Company, Inc.; 1960.

23. Podani J, Miklos I. Ecology. 2002; 83(12):3331–3343.

24. Zhao K, Aranzana M, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M. PLoS Genet. 2007; 3(1):e4. [PubMed: 17238287]

25. Richman MB. Journal of Climatology. 1986; 6:293–335.

26. Serre D, Paabo SP. Genome Research. 2004; 14(9)

27. Irwin DE, Bensch S, Irwin JH, Price TD. Science. 2005; 307(5708):414–416. [PubMed: 15662011]

28. Heidemann G. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006; 28:822–826. [PubMed: 16640267]

29. Jolliffe, I. Principal Component Analysis. Springer Series in Statistics. Springer; 1986.

30. Preisendorfer, R. Principal Component Analysis in Meteorology and Oceanography. Amsterdam: Number 17 in Developments in Atmospheric Science. Elsevier; 1988.

31. Richman M. International Journal of Climatology. 1993; 13:203–218.

32. Bauchet M, McEvoy B, Pearson LN, Quillen E, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD. Am J Hum Genet. 2007; 80:948–956. [PubMed: 17436249]

33. Barbujani G, Chikhi L. Heredity. 2006; 97(2):84–85. [PubMed: 16721387]

34. Zilhao J. Proc Natl Acad Sci USA. 2001; 98(24):14180–14185. [PubMed: 11707599]

35. Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ, Zhivotovsky LA, King R, Torroni A, Cavalli-Sforza LL, Underhill PA, Santachiara-Benerecetti AS. Am J Hum Genet. 2004; 74(5):1023–1034. [PubMed: 15069642]

36. Haak W, Forster P, Bramanti B, Matsumura S, Brandt G, Tanzer M, Villems R, Renfrew C, Gronenborn D, Alt KW, Burger J. Science. 2005; 310(5750):1016–1018. [PubMed: 16284177]

37. Pinhasi R, Fort J, Ammerman AJ. PLoS Biol. 2005; 3(12):e410. [PubMed: 16292981]

38. Belle EMS, Landry P-A, Barbujani G. Proc Biol Sci. 2006; 273(1594):1595–1602. [PubMed: 16769629]

39. Sampietro ML, Lao O, Caramelli D, Lari M, Pou R, Marti M, Bertranpetit J, Lalueza-Fox C. Proc Biol Sci. 2007; 274(1622):2161–2167. [PubMed: 17609193]

40. Wellcome Trust Case Control Consortium. Nature. 2007; 447(7145):661–678. [PubMed: 17554300]

41. Zhu X, Zhang S, Zhao H, Cooper RS. Genet Epidemiol. 2002; 23(2):181–196. [PubMed: 12214310]

42. Pritchard JK, Rosenberg NA. Am J Hum Genet. 1999; 65(1):220–228. [PubMed: 10364535]

43. Rosenberg NA, Nordborg M. Genetics. 2006; 173(3):1665–1678. [PubMed: 16582435]

44. Wakefield J. Biostatistics. 2007; 8(2):158–183. [PubMed: 16809429]

45. R Development Core Team. R: A Language Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2007.

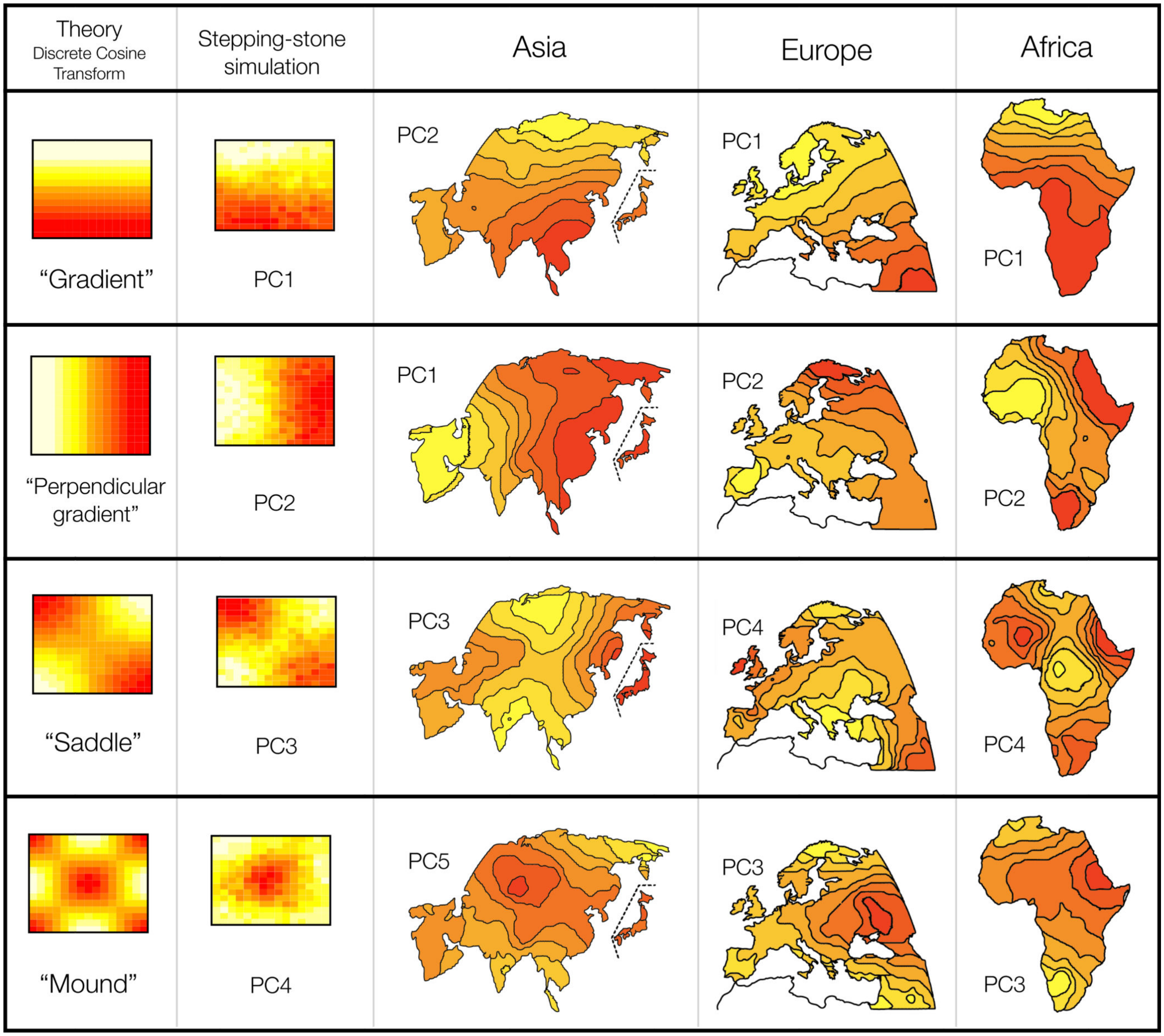

Figure 1.

Comparison of PC-maps of [3] with theoretical and empirical predictions. The first column shows the theoretical expected PC-maps for a class of models in which genetic similarity decays with geographic distance (see text for details). The second column shows PC-maps for population genetic data simulated with no range expansions, but constant homogeneous migration rate in a 2-dimensional habitat. The columns marked Asia, Europe, and Africa are redrawn from the originals of [3]. Each map is marked by which PC it represents. The order of maps in each of the last three columns was chosen to correspond with the shapes in the first two columns.

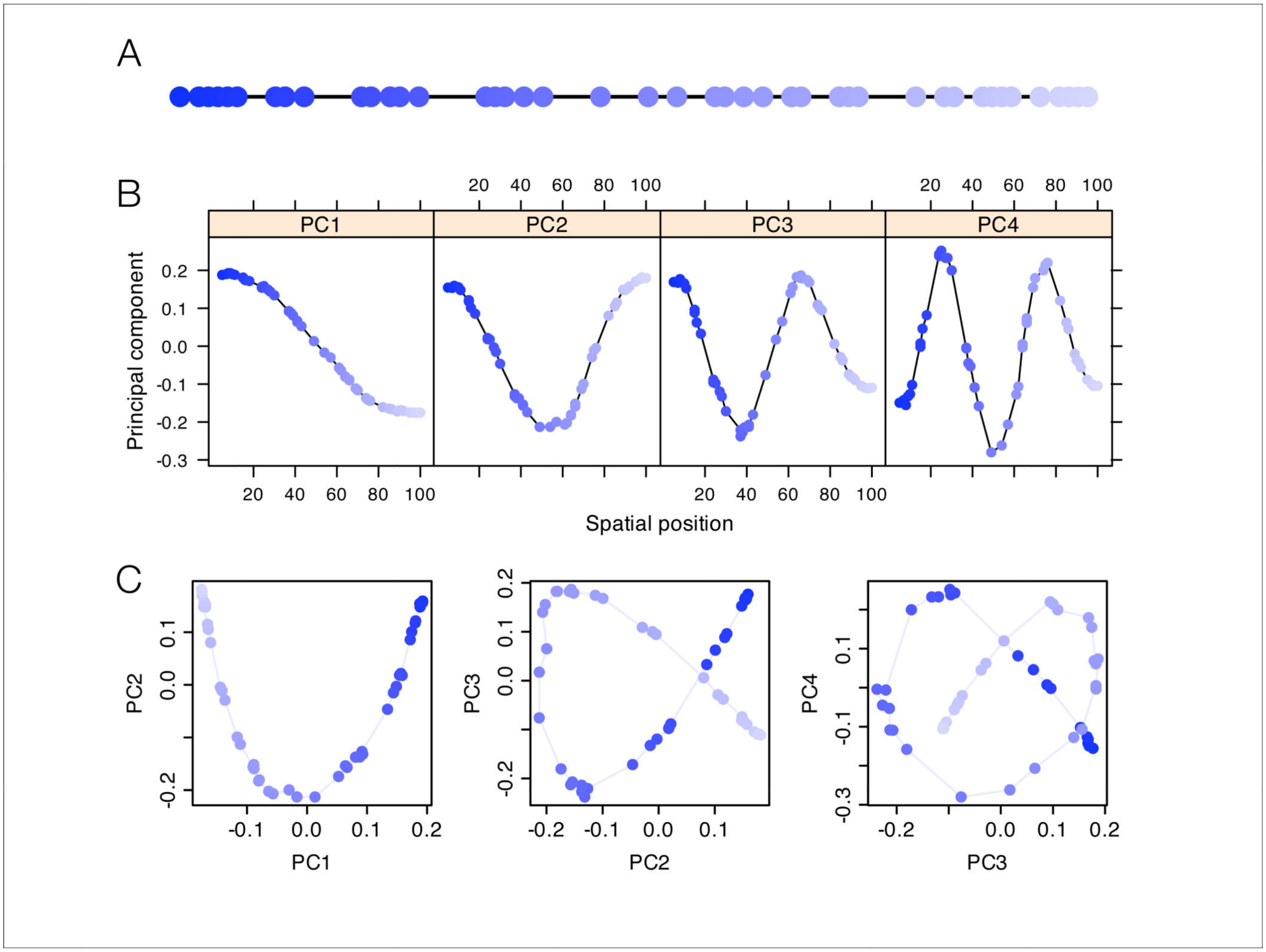

Figure 2.

Results of PCA applied to data from a one-dimensional habitat. (A) Schematic of the one-dimensional habitat, with circles marking sampling locations and shades of blue marking order along the line. (B) One-dimensional PC-maps (i.e. plots of each PC element against the geographic position of the corresponding sample location). (C) Biplots of PC1 vs. PC2, PC2 vs. PC3, and PC3 vs. PC4. Colors correspond to those in Panel A. In many datasets without spatially referenced samples, the colors and the lines connecting neighboring points would not be observed; here they are shown to aid interpretation.