nature genetics

Interpreting principal component analyses of spatial population genetic variation

John Novembre[1,3] & Matthew Stephens[1,2]

Nearly 30 years ago, Cavalli-Sforza *et al.* pioneered the use of principal component analysis (PCA) in population genetics and used PCA to produce maps summarizing human genetic variation across continental regions[1]. They interpreted gradient and wave patterns in these maps as signatures of specific migration events[1–3]. These interpretations have been controversial[4–7], but influential[8], and the use of PCA has become widespread in analysis of population genetics data[9–13]. However, the behavior of PCA for genetic data showing continuous spatial variation, such as might exist within human continental groups, has been less well characterized. Here, we find that gradients and waves observed in Cavalli-Sforza *et al.*'s maps resemble sinusoidal mathematical artifacts that arise generally when PCA is applied to spatial data, implying that the patterns do not necessarily reflect specific migration events. Our findings aid interpretation of PCA results and suggest how PCA can help correct for continuous population structure in association studies.

Cavalli-Sforza *et al.*'s classic text "The History and Geography of Human Genes"[3] synthesizes a decades-long survey of human genetic variation. These ground-breaking datasets stimulated development of methods that are now widely used, including application of principal component analysis (PCA) to population genetic variation. In essence, Cavalli-Sforza *et al.* collected count data for many genetic variants ("alleles") 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 summarizing 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 many allele-frequency maps, in that each allele-frequency map can be well approximated by a linear superposition of PC maps.

Figure 1 shows PC maps for Asia, Europe and Africa from refs. 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 of ref. 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 north-western 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).

Because the basis for these 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 equilibrium population genetic models without range expansions, assuming a constant homogeneous short-range migration process across both time and (two-dimensional) space. The results showed highly distinctive structure. For example, the first two PC maps show large-scale orthogonal gradients, and the next two show '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 (**Supplementary Fig. 1** online). Results for the analogous one-dimensional habitat setting are even more structured, resembling sinusoidal functions of increasing frequency (**Fig. 2**, **Supplementary Fig. 2** online). Thus PC maps show local peaks and troughs *even when underlying migration patterns are homogeneous across time and space*. This suggests that the local features of the PC maps do not necessarily indicate specific localized historical migration events. Furthermore, many PC maps obtained by Cavalli-Sforza *et al.* in Asia, Europe and Africa show patterns strikingly similar to those from our simulations (**Fig. 1**, **Supplementary Fig. 1**).

In fact, these highly structured patterns are mathematical artifacts that arise generally when PCA is applied to spatial data in which covariance (similarity) between locations tends to decay with geographic distance. Such data produce highly structured covariance matrices (see, for example, **Supplementary Fig. 3** online), with special mathematical properties. In particular, they have eigenvectors related to sinusoidal waves of increasing frequency (for example, ref. 14). This produces sinusoidal patterns in PC maps because PC maps are visual representations of these eigenvectors (see **Supplementary Methods**

[1]Department of Human Genetics, University of Chicago, 920 E. 58th Street, CLSC 5th floor, Chicago, Illinois 60637, USA. [2]Department of Statistics, University of Chicago, 5734 S. University Ave., Chicago, Illinois 60637, USA. [3]Present address: Department of Ecology and Evolutionary Biology, University of California Los Angeles, 621 Charles E. Young Dr. South, Los Angeles, California 90095, USA. Correspondence should be addressed to M.S. (stephens@galton.uchicago.edu).

Received 1 August 2007; accepted 17 March 2008; published online 20 April 2008; doi:10.1038/ng.139

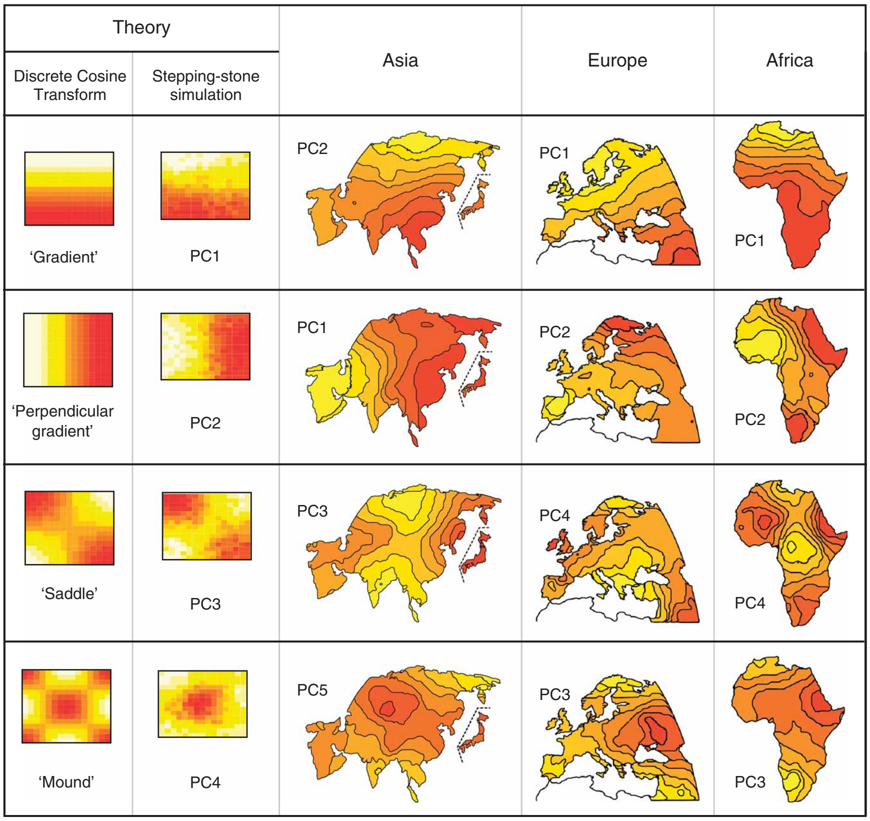

Figure 1 Comparison of PC maps of ref. 3 with theoretical and empirical predictions. The first column shows the theoretical expected PC maps for a class of models in which genetic similarity decays with geographic distance (see text for details). The second column shows PC maps for population genetic data simulated with no range expansions, but constant homogeneous migration rate, in a two-dimensional habitat. The columns marked Asia, Europe and Africa are redrawn from the originals of ref. 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.

online). To give some concrete population genetics examples, if the similarity between two populations depends only on the geographic distance between them, and PCA is applied to populations that are regularly spaced within a linear, circular or two-dimensional habitat, then the resulting covariance matrices have very particular structures (known as Toeplitz, Circulant and Block Toeplitz with Toeplitz Blocks, respectively; **Supplementary Fig. 3b**), with eigenvectors that are sinusoidal functions (columns of the Discrete Cosine, Discrete Fourier and Two-Dimensional Discrete Cosine Transform matrices, respectively; see **Supplementary Note** and **Supplementary Fig. 4** online). The results apply equally when PCA is applied to individual genotype data[11] rather than population allele frequencies. Indeed, they apply quite generally, and have been previously recognized in other fields, including time-series[15], ecology[16] and climatology[17]. Further, although the formal mathematical results inevitably involve idealized scenarios, extensive empirical data in multiple fields[16–18] show sinusoidal patterns emerging from PCA of spatial data.

For insight into why sinusoidal patterns emerge in PC maps, it is perhaps not directly helpful to look to the common description of PCA, as identifying directions of maximum variance, because these directions are in a high-dimensional mathematical space, not geographic space. Instead, consider the property of PC maps mentioned above: it should be possible to accurately approximate each allele-frequency map using a linear superposition of the first few PC maps. PC maps that contain sinusoidal functions of increasing frequency accomplish this sensibly: low-frequency patterns in early PC maps allow a coarse approximation that reflects allele frequency changes across large spatial scales, whereas higher-frequency patterns in subsequent PC maps allow refinement of this approximation to capture finer-scale changes.

In some settings, particularly those involving individual genotype data, geographic information may not be available for each sample, making PC maps difficult to produce. Instead PCA results are commonly visualized by producing biplots of one PC against another. Under uniform sampling from a one-dimensional habitat with homogeneous migration, this results in biplots of sinusoidal functions of differing frequencies, producing characteristic patterns known as Lissajous curves[19] (**Fig. 2c**). In particular, the biplot of PC1 versus PC2 shows a pattern known as the "horseshoe effect" (for example, see refs. 16,20). For the analogous two-dimensional setting, because PC1 and PC2 are typically orthogonal gradients, plotting PC1 versus PC2 essentially reproduces the geographic arrangement of sampled individuals (explaining PCA results for *Arabidopsis thaliana*[21], for example), and biplots involving later PCs have intricate patterns analogous to Lissajous curves (**Supplementary Fig. 5** online).

To assess the effects of deviations from the simplistic scenarios assumed in our initial simulations, we examined 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. Because quantifying this effect is difficult, we instead provide several examples for illustration. Anisotropic migration (migration that is not equal in all directions, **Supplementary Fig. 6** online) and irregularly spaced populations (**Supplementary Fig. 7** online) both distort the PC maps, and change their order. The direction of the gradient in the first PC map is influenced by habitat shape (for example, 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[17], and by migration patterns (for example, under anisotropic migration in a square habitat, the gradient in PC1 aligns with the axis of least migration; **Supplementary Fig. 6**). However, sinusoidal-like patterns consistently emerge. Even when sampling locations are highly clustered within the continuous habitat (a common sampling design because of logistical challenges to obtaining spatially uniform samples in many species), the first PCs separate out the clusters as if the sample were obtained from discrete subpopulations, and subsequent PCs show sinusoidal patterns within clusters (**Supplementary Fig. 8** online).

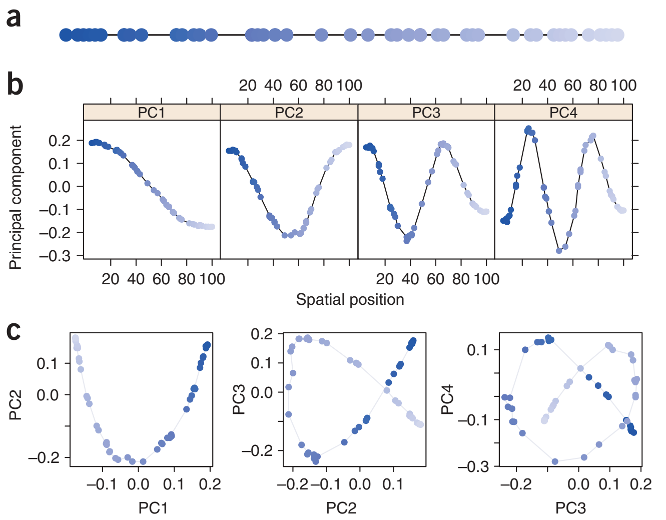

Figure 2 Results of PCA applied to data from a one-dimensional habitat. (**a**) Schematic of the one-dimensional habitat, with circles marking sampling locations and shades of blue marking order along the line. (**b**) One-dimensional PC maps (that is, plots of each PC element against the geographic position of the corresponding sample location). (**c**) Biplots of PC1 versus PC2, PC2 versus PC3, and PC3 versus PC4. Colors correspond to those in **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.

Although the results above all involve large datasets (many loci), sinusoidal patterns can also emerge within smaller datasets. For example, such patterns occur in PC1 and PC2 (**Supplementary Fig. 9** online) 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 Methods**[22] and **Supplementary Fig. 9**). However, limited data can produce less well-defined (or entirely absent) sinusoidal patterns, particularly in higher PC maps (for example, PC3 for the same dataset; **Supplementary Fig. 9**). In general, amounts of data needed to recover sinusoidal patterns will depend on the strength of the population structure (for example, higher migration rates reduce population differentiation, giving less well-defined sinusoidal patterns; **Supplementary Fig. 2b**).

In summary, we have shown that (i) when analyzing spatial data, PCA produces highly structured results relating to sinusoidal functions of increasing frequency; and (ii) insofar as PCA results depend on the details of a particular dataset, they are affected by factors in addition to population structure, including distribution of sampling locations and amounts of data. Both these features limit the utility of PCA for drawing inferences about underlying processes, a fact previously noted in climatology (for example, ref. 17). In particular, interpreting gradient- and wave-like patterns in PC maps as signatures of historical migration events is problematic because such patterns arise generally under a simple condition: that genetic similarity decays with distance. This condition would be expected to be satisfied under a wide range of demographic scenarios, including both equilibrium isolation-by-distance models and nonequilibrium models involving population expansions[23]. Furthermore, because Cavalli-Sforza *et al.* used spatial interpolation to estimate allele frequencies, their data could satisfy this condition even if the condition were absent in the underlying allele frequencies[5]. (Use of interpolation may partly explain the 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[12] show perpendicular gradients in PC1 and PC2.)

Regarding the Neolithic expansion into Europe (and other migration events that have been argued for using PCA), we emphasize that this paper is not about whether or not such events have occurred; a full consideration of this would require, in each case, a synthesis of evidence from many diverse sources (for example, refs. 7, 24–26). The northwest-southeast 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 account for the potential influence of other factors, such as the shape of the continent). However, if a Neolithic expansion could explain this, it is but one of various 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 patterns similar 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 that genetic variation in *H. pylori* has been influenced by human migrations. However, our results show that similar patterns in PC maps of two groups does not necessarily 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 (for example, **Supplementary Fig. 1**).

Despite limitations for inferring underlying processes that have produced population structure, PCA is undoubtedly an extremely useful tool for investigating and summarizing population structure, and we anticipate it playing a prominent role in analyses of ongoing studies of population genetic variation. The analyses 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 structure. On the other hand, departures from this null may also be useful, perhaps pointing toward a more discrete 'cluster-like' population structure[11] or to other important structure in the data, such as genotyping error or regions of high linkage disequilibrium[27].

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[10,28]. For simplicity, we focus on association mapping of a quantitative trait using a population sample. In essence, the problem is that if phenotype mean varies among subpopulations, then alleles that have no mechanistic connection to phenotype, but differ in frequency among subpopulations, will be 'spuriously' associated with phenotype[29]. Although this problem has been studied mostly in the context of discrete subpopulations, it applies also to continuous (for example, spatial) variation[21]. One solution recently suggested[10] is to include the first few PCs as covariates in a regression. In populations showing a discrete, cluster-like structure, these PCs typically separate out the clusters[11], and so this solution corresponds to allowing for phenotype mean to vary among subpopulations. Our work now suggests that, for spatially continuous populations, the PCA-based approach is conceptually similar to modeling smooth geographical trends in phenotype mean. 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 mean across space.

If later PCs relate to sinusoidal waves of increasing frequency, then including them in the regression allows for more flexible spatial trends. Controlling for smooth geographic trends in phenotype is a recognized technique in spatial epidemiology[30], so this view gives some intuitive support for the use of PCA to control for spurious associations in spatially structured populations. Further, PCA has important practical advantages over the use of geographic information on each individual directly: it can be used even when geographic information is not available, or when geographic position does not correlate well with genetic background (as is typical in the United States, for example). Nonetheless, practical issues remain. For example, it may be more robust to use nonlinear functions of the first two PCs, rather than higher PCs, to capture nonlinear spatial trends. And, we suggest, some attempt should be made to control for only those PCs that are correlated with phenotype, as controlling for other PCs is unnecessary and may reduce power.

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 of $2N$ gametes, and, backward in time, in each generation, a proportion, m, of gametes swap places with an equal number of gametes in each neighboring deme (for example, for the two-dimensional 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 (that is, 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.*[1–3]) and 'individual-based' data (as in ref. 11). For generating population-based datasets, we sampled n individuals from D_s of the D demes, and simulate, for each individual, data at L independent, biallelic polymorphic loci. Assuming independence of loci corresponds to assuming 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 (**Fig. 1** and **Supplementary Figs. 1**, **3**, **5** and **6**), we used a regular square lattice of $D_s = 15 \times 15$ demes embedded in the center of a larger $D = 31 \times 31$ lattice of demes. Allele frequencies in each deme are estimated 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 **Supplementary Methods** for additional details.

Principal component analysis. To calculate principal components on our simulated data, we use biallelic loci and include only the frequency of one of the two alleles. 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 spatially interpolated on a dense grid. This avoids problems with interpolation altering underlying spatial covariance patterns[5]. For individual-based PCA, we use an approach similar to that of Patterson *et al.*[11], in that we scale the genotype values across individuals at each locus to have unit variance (see **Supplementary Methods**). 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 the approach of ref. 11; see **Supplementary Methods** for more detail).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank G. Coop, M. Przeworski, D. Reich, N. Patterson, Y. Guan, M. Barber and C. Becquet for helpful comments and D. Witonsky for pointing out the connection to Lissajous figures. Funding for this research was provided by a US National Science Foundation postdoctoral research fellowship in bioinformatics (J.N.) and US National Institutes of Health grant RO1 HG02585-01 (M.S.).

AUTHOR CONTRIBUTIONS

J.N. and M.S. jointly designed the analyses and interpreted results. J.N. performed the analyses. J.N. and M.S. wrote the paper.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792 (1978).
2. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. Demic expansions and human evolution. *Science* **259**, 639–646 (1993).
3. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton University Press, Princeton, New Jersey, USA, 1994).
4. Rendine, S., Piazza, A. & Cavalli-Sforza, L.L. Simulation and separation by principal components of multiple demic expansions in Europe. *Am. Nat.* **128**, 681–706 (1986).
5. Sokal, R.R., Oden, N.L. & Thomson, B.A. A problem with synthetic maps. *Hum. Biol.* **71**, 1–13 (1999).
6. Rendine, S., Piazza, A. & Cavalli-Sforza, L.L. A problem with synthetic maps: Reply to Sokal et al. *Hum. Biol.* **71**, 15–25 (1999).
7. Currat, M. & Excoffier, L. The effect of the Neolithic expansion on European molecular diversity. *Proc. Biol. Sci.* **272**, 679–688 (2005).
8. Jobling, M., Hurles, M. & Tyler-Smith, C. *Human Evolutionary Genetics* (Garland Science, New York, 2004).
9. Hanotte, O. *et al.* African pastoralism: genetic imprints of origins and migrations. *Science* **296**, 336–339 (2002).
10. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
11. Patterson, N., Price, A. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
12. Bauchet, M. *et al.* Measuring European population stratification with microarray genotype data. *Am. J. Hum. Genet.* **80**, 948–956 (2007).
13. Linz, B. *et al.* An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**, 915–918 (2007).
14. Ahmed, N., Natarajan, T. & Rao, K.R. Discrete cosine transform. *IEEE Trans. Comput.* **C-23**, 90–93 (1974).
15. Brillinger, D.R. *Time Series: Data Analysis and Theory* (Holt, Rinehart, and Winston, New York, 1975).
16. Podani, J. & Miklos, I. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology* **83**, 3331–3343 (2002).
17. Richman, M.B. Rotation of principal components. *J. Climatol.* **6**, 293–335 (1986).
18. Heidemann, G. The principal components of natural images revisited. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 822–826 (2006).
19. Freiberger, W., ed. *The International Dictionary of Applied Mathematics* (D. Van Nostrand Co., Princeton, New Jersey, USA, 1960).
20. Diaconis, P., Goel, S. & Holmes, S. Horseshoes in multidimensional scaling and kernel methods. *Ann. Appl. Stat.* (in the press).
21. Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3**, e4 (2007).
22. Irwin, D.E., Bensch, S., Irwin, J.H. & Price, T.D. Speciation by distance in a ring species. *Science* **307**, 414–416 (2005).
23. Handley, L.J.L., Manica, A., Goudet, J. & Balloux, F. Going the distance: human population genetics in a clinal world. *Trends Genet.* **23**, 432–439 (2007).
24. Semino, O. *et al.* Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the Neolithization of Europe and later migratory events in the Mediterranean area. *Am. J. Hum. Genet.* **74**, 1023–1034 (2004).
25. Haak, W. *et al.* Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* **310**, 1016–1018 (2005).
26. Pinhasi, R., Fort, J. & Ammerman, A.J. Tracing the origin and spread of agriculture in Europe. *PLoS Biol.* **3**, e410 (2005).
27. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
28. Zhu, X., Zhang, S., Zhao, H. & Cooper, R.S. Association mapping, using a mixture model for complex traits. *Genet. Epidemiol.* **23**, 181–196 (2002).
29. Pritchard, J.K. & Rosenberg, N.A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228 (1999).
30. Wakefield, J. Disease mapping and spatial regression with count data. *Biostatistics* **8**, 158–183 (2007).