Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture

Olivier François,*[,1] Mathias Currat,[2] Nicolas Ray,[3,4,5] Eunjung Han,[5] Laurent Excoffier,[3,4] and John Novembre[6,7]

[1]Laboratoire Techniques de l'Ingénierie Médicale et de la Complexité, Faculty of Medicine, University Joseph Fourier, Grenoble Institute of Technology, Centre National de la Recherche Scientifique UMR5525, La Tronche, France

[2]Laboratory of Anthropology, Genetics and Peopling history, Department of Anthropology and Ecology, University of Geneva, Geneva, Switzerland

[3]Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution, University of Berne, Berne, Switzerland

[4]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[5]EnviroSPACE laboratory, Climate Change and Climate Impacts, Institute for Environmental Sciences, University of Geneva, Carouge, Switzerland

[6]Department of Ecology and Evolutionary Biology, University of California

[7]Interdepartmental Program in Bioinformatics, University of California-Los Angeles

***Corresponding author:** E-mail: olivier.francois@imag.fr.
Associate editor: Jonathan Pritchard

Research article

Abstract

In a series of highly influential publications, Cavalli-Sforza and colleagues used principal component (PC) analysis to produce maps depicting how human genetic diversity varies across geographic space. Within Europe, the first axis of variation (PC1) was interpreted as evidence for the demic diffusion model of agriculture, in which farmers expanded from the Near East ~10,000 years ago and replaced the resident hunter-gatherer populations with little or no interbreeding. These interpretations of the PC maps have been recently questioned as the original results can be reproduced under models of spatially covarying allele frequencies without any expansion. Here, we study PC maps for data simulated under models of range expansion and admixture. Our simulations include a spatially realistic model of Neolithic farmer expansion and assume various levels of interbreeding between farmer and resident hunter-gatherer populations. An important result is that under a broad range of conditions, the gradients in PC1 maps are oriented along a direction perpendicular to the axis of the expansion, rather than along the same axis as the expansion. We propose that this surprising pattern is an outcome of the "allele surfing" phenomenon, which creates sectors of high allele-frequency differentiation that align perpendicular to the direction of the expansion.

Key words: population structure, range expansion, admixture, demic diffusion model, principal component analysis.

Introduction

Since its earliest uses (Cavalli-Sforza and Edwards 1963; Harpending and Jenkins 1973; Menozzi et al. 1978), principal component analysis (PCA) has become a popular tool for exploring multilocus population genetic data (Menozzi et al. 1978; Rendine et al. 1986; Cavalli-Sforza et al. 1993; Cavalli-Sforza et al. 1994; Patterson et al. 2006; Novembre and Stephens 2008). PCA is a general method for representing high-dimensional data, for example, individuals or populations, in a smaller number of dimensions. It has recently regained popularity as a tool to summarize large-scale genomic surveys, by providing covariates that might correct for population structure in genomewide association studies (Patterson et al. 2006; Price et al. 2006) and by unveiling the main factors explaining the structure of genetic variation in large samples (Jakobsson et al. 2008; Li et al. 2008; Novembre et al. 2008).

One way to explain PCA is as an algorithm that iteratively searches for orthogonal axes, described as linear combinations of multivariate observations, along which projected objects show the highest variance, and then returns the positions of objects along those axes (the principal components [PCs]). For many data sets, the relative position of these objects (e.g., individuals) along the first few PCs provides a reasonable approximation of the covariance pattern among individuals in the larger data set. As a result, the first few PC values are often used to explore the structure of variation in the sample.

In one of the largest applications of PCA prior to the advent of large-scale single-nucleotide polymorphism (SNP) data, PCA was used to summarize allele-frequency data collected from worldwide populations of humans (Cavalli-Sforza et al. 1994). The results of the PCA were visualized using "synthetic maps" or "PC maps" depicting how the PC values for each sampled population vary across geographic space (with each PC being displayed on a separate map). Notably, in many of the maps generated from their data, gradients and wave-like patterns were observed.

The interpretation of these gradient and wave-like patterns has been somewhat controversial (Sokal et al. 1999;

Novembre and Stephens 2008). In their original formulation, Cavalli-Sforza et al. (1994) favored explanations in which the gradients and wave-like shapes were signatures of past expansion events. For example, Menozzi et al. (1978) observed a large southeast (SE) to northwest (NW) gradient for PC1 across Europe and concluded that this gradient was the outcome of a SE-to-NW expansion of agriculturalist populations during the Neolithic era. In this "demic diffusion" model, farmers expanded into Europe from the Near East ~10,000 years ago, replacing Paleolithic populations of hunter-gatherers with little or no admixture (Ammerman and Cavalli-Sforza 1984; see also Davies 1998; Diamond and Bellwood 2003). The model implies that agriculture spreads more by the migration of farming populations than by the cultural diffusion of the agricultural technologies.

One complication of this interpretation is that gradient and wave-like shapes arise quite generally in synthetic maps for data that are spatially structured (Novembre and Stephens 2008). In simple scenarios where samples are spaced evenly and covariance decays exponentially with distance, PC maps are expected to show regular patterns where typically the first map is of a gradient, the second is a gradient perpendicular to the first, and the third and fourth PC maps are "saddle"- and "mound"-like wave shapes. Novembre and Stephens (2008) review mathematical arguments that explain these patterns and demonstrate their presence using simulations from simple population genetic models (symmetric migration between populations arranged on a square lattice and mutation–migration–drift equilibrium). Simulations in more complicated scenarios of spatial structure (unequal migration, irregular habitat shape, irregular sampling) evidenced distortions of these basic patterns, but the patterns generally included gradients and wave-like shapes. Thus, the observation of sinusoidal functions in PC maps, such as gradients or waves, is not strong evidence for specific past expansion events because a large range of models inducing spatial structure will give rise to similar patterns.

An unanswered question is to "what extent do sinusoidal patterns and gradients arise if a spatial expansion has occurred?." Novembre and Stephens (2008) do not present simulations of range expansions, nor show, for instance, that observing a SE-to-NW gradient in PC1 is consistent or inconsistent with the Neolithic expansion. One might expect that recent expansions will result in spatially structured data, and thus based on the results of Novembre and Stephens (2008) that some sinusoidal patterns should appear. However, if the patterns appear, is there a systematic distortion of the sinusoidal shapes as a signature of an expansion? For example, all else being equal, one might expect that the largest axis of genetic differentiation would be along the direction of the expansion, and thus if there is a gradient in PC1, its direction would be indicative of the direction of the historical expansion, as supposed in the classic interpretations of PC gradients.

Addressing these questions is particularly relevant to the contentious debate over the Neolithic expansion in Europe. Although it is unlikely that lower PC maps are indicative of unique historical expansions, the direction of the gradient in PC1 in the Menozzi et al. (1978) analysis and in more recent analyses (Lao et al. 2008; Novembre et al. 2008; see also Heath et al. 2008) might be consistent with a recent Neolithic expansion from the southeast toward the northwest.

To address the issue of how PCA behaves on samples of genetic variation obtained after range expansions, we explored a variety of spatial expansion scenarios using computer simulations to mimic massive migration from one or two sources. Previous simulations have been conducted by Rendine et al. (1986), but due to computational advances, we are able to explore a wider range of scenarios. To specifically address the Neolithic expansion in Europe, we modeled an expansion using a spatial model of Europe, parameterized in such a way that migration rates vary according to topography, and incorporating archaeological information about the timing of the arrival of modern humans in Europe as well as start of the Neolithic expansion. In order to get a broader perspective on the problem, we also explored a wide spectrum of other scenarios including more ancient expansions, multiple sources, and expansion on simple regular lattices.

A surprising result of our simulation study is that the gradients observed in the first PC map often are found to be, contrary to most often formulated expectations, perpendicular to the main direction of expansion. We found this to be true for parameters representative of hypothesized Neolithic demic expansions into Europe from the Near East. To explore the robustness of this result, we considered various introgression rates in our model of a European Neolithic expansion. We confirmed that the direction of greatest differentiation is perpendicular to the expansion by plotting how genetic differentiation increases with geographic distance along both geographic axes and by applying assignment methods (AM). For example, when $K = 2$, we observed a gradient of assignment probabilities running perpendicular to the expansion. One possible mechanistic explanation for these results is that it is an outcome of the genetic surfing phenomenon (Edmonds et al. 2004; Klopfstein et al. 2006; Currat et al. 2008). We discuss the implications of these findings for the analysis of population structure with PCA and assignment algorithms.

Material and Methods

Spatial Simulations

Spatial simulations of sampled molecular diversity were performed with a modified version of the computer program SPLATCHE, which uses a two-stage coalescent model of migration incorporating topographic information (Currat et al. 2004). Forward in time, the demographic history of a population is simulated in a nonequilibrium stepping-stone model defined on a lattice of regularly spaced subpopulations or demes (fig. 1*A*). In this simulation, spatial information is encoded into a friction value for each deme (fig. 1*B*), and each deme sends migrants to its nearest neighbors at rate m with directional probabilities inversely proportional to the neighbors' friction values. Once a deme is colonized, its population size starts growing according to a standard

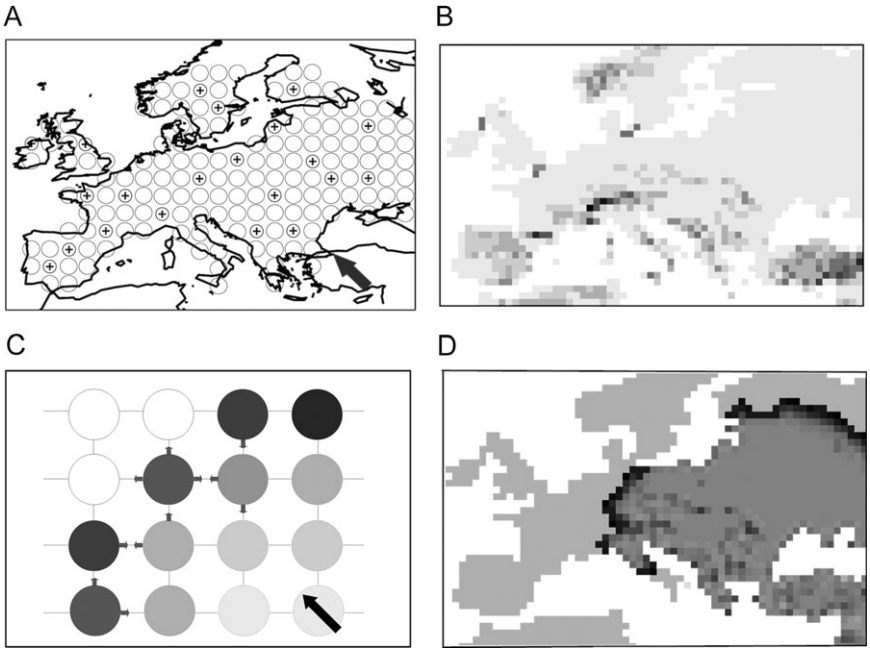

FIG. 1. Illustration of the simulated demographic processes. (*A*) A schematic representation of how Europe is modeled as an irregular array of demes. To simulate genetic data, multilocus genotypes are sampled at uniformly distributed locations, taking 20 individuals at each sampling site (crosses). (*B*) The friction map that encodes the inverse migration rates used in the demographic simulations. Dark values indicate low migration rates. (*C–D*) Picture of the wave-of-advance model at a fixed simulation time. Range expansion starts from the bottom-right corner of the area. Demes with the light gray colors are saturated at their carrying capacities (white demes are empty), whereas the dark gray colors indicate lower densities in particular at the front of the expansion.

logistic model with rate r and carrying capacity C. The model results in a wave-of-advance of the population, as shown in figure 1*C–D*. The shape and speed of the wave-of-advance depend on the parameters of the model, r, C, and m. Backward in time, the demographic parameters are used to generate gene genealogies for samples taken at different geographic locations under a coalescent framework. The population size Ct of a given deme at any time t is used to compute the probability of coalescence for a pair of genes from that deme; backward migration probabilities are calculated using the number of migrants arriving from neighboring demes in the forward step. We used SPLATCHE to simulate various types of genetic markers including short tandem repeats (microsatellite data) and DNA sequence data.

Simulating Neolithic Expansion in Europe

Range expansion occurred in 64 × 42 lattices covering Europe from latitude 38°N to 65°N and from longitude 10°W to 40°E (2,688 cells; fig. 1*B*). In order to enable migration to and from the British Isles and Scandinavia, these regions were connected to the mainland by two narrow bridges associated with friction values 10-fold higher than in plains. The settlement of Europe was fixed at 1,600 generations before the present (Mellars 2006). Regarding this Paleolithic expansion, we used a simplified single-origin model, assuming that modern humans replaced archaic populations without genetic introgression as they arrived in Europe (Currat and Excoffier 2004). Technically, this expansion occurred on a first layer of demes representing hunter-gatherers. The carrying capacity of each deme in this first layer was set to $C = 50$, corresponding to a density of ~0.05 individual per km^2 (Steele et al. 1998). The population size at the onset of the expansion was of 100 individuals (The "density overflow" option was used to spread the ancestral population over patches of up to ~20 demes). Four hundred generations before the present, a second range expansion started from the southeast (Anatolia). This occurred in a second layer of demes representing Neolithic farmer populations who could potentially interbreed with the resident populations. The carrying capacity of Neolithic demes and the size of the ancestral population were set to values 10-fold larger than for hunter-gatherers (Ammerman and Cavalli-Sforza 1984). Hunter-gatherers ultimately disappeared due to density-dependent competition with the farmers (for further details about the competition model used, see Currat and Excoffier 2005). Migration and growth rates have been calibrated to obtain a maximum of 500 generations for the duration of the Paleolithic settlement (Mellars 2004) and around 300 generations for the Neolithic transition (Pinhasi et al. 2005). These scenarios correspond to the following values: migration rates $m = 0.4$ growth rates $r = 0.5$ (Paleolithic) or $r = 0.4$ (Neolithic). Two distinct sources for the Paleolithic expansion were considered: One in the Near-East, representing a starting point for the arrival of modern humans in Europe about 40,000 years ago (Mellars 2004) and one in the center of the Iberian peninsula representing an hypothetical expansion from a glacial refugium 20,000 years ago. Four different values for the rate of interbreeding, γ, have been chosen in order to reproduce extreme as well as intermediate scenarios: i) $\gamma = 0$ is

a pure Neolithic demic diffusion or replacement scenario (100% of Neolithic ancestry in the final European genetic pool); ii) $\gamma = 0.0075$ corresponds to about 80% of Neolithic ancestry in the final genetic pool; iii) $\gamma = 0.04$ corresponds to about 20% of Neolithic ancestry in the final genetic pool; iv) $\gamma = 0.068$ corresponds to less than 10% of Neolithic ancestry in the final genetic pool. These values are similar to the rates of acculturation considered by Cavalli-Sforza and Ammerman (1984) and Barbujani et al. (1995). Allelic states were simulated under a strict stepwise mutation model using $L = 100$ unlinked microsatellite loci, and a mutation rate of 5×10^{-4} per generation per locus. 200 bp DNA sequences were also generated at 2,000 unlinked loci, with a mutation rate of 10^{-7} per bp per generation. To minimize the potentially confounding effect of using an irregular sampling design (McVean 2009), samples of 20 (haploid) individuals were simulated in 60 randomly selected cells (note: the same sampling is used in all simulations of Europe, shown in fig. 2A).

Simulation on a Regular Lattice

Additional simulations of demic expansions without admixture were performed on the same lattice as for prehistoric scenarios, using a uniform friction map and sampling 10 individuals in every deme (26,880 individuals simulated for $L = 100$ unlinked loci). For these simpler simulations, we explored a wide range of demographic parameters. Expansions started from the southeast $T = 500$, $T = 1{,}000$, or $T = 2{,}000$ generations ago; migration rates took three distinct values $m = 0.2$, $m = 0.5$ and $m = 0.8$; growth rates took two values $r = 0.5$ and $r = 1.0$; and carrying capacities were set to either $C = 500$ or $C = 1{,}000$, that were equal to the ancestral population size.

PCA and Assignment Algorithms

PCA was performed on a data set of multilocus genotypes (individuals) to mimic the approaches used in the latest analyses of population genetic variation. The genotype matrix was normalized by subtracting the mean and dividing the resulting quantity by the standard deviation of the jth column (as in Patterson et al. 2006; Novembre and Stephens 2008). Given the renormalized matrix, M, we computed the eigenvalues and eigenvectors of the sample covariance matrix, $X = MM'/n$, by applying the "prcomp" function of the R statistical package. Note that the original analyses of Menozzi et al. (1978) applied PCA on a population level. For a fraction of the simulations performed here, we used the population-based approach and we replicated our main results (results not shown). In addition to exploring the behavior of PCA on expansion simulations, we also applied AM to each of the simulated scenarios. These methods are commonly used computational tools for inferring population genetic structure, and the connection between PCA and admixture estimation methods (which are closely related to AM) has been recently investigated by Patterson et al. (2006). In contrast to PCA, AM are model-based methods, which means that they use explicit model definitions for their likelihood function (Beaumont and Rannala 2004). AM programs use assignment of individuals to K putative populations also termed "genetic clusters." The assignment of each individual genotype into each genetic cluster is carried out probabilistically by using Markov chain Monte Carlo methods. AM analyses were carried out by using the computer programs STRUCTURE (Pritchard et al. 2000) and TESS (Chen et al. 2007; Durand et al. 2009) under their default options. Although they used distinct prior distributions, these programs were grouped under a common terminology because their outputs displayed only minor differences for the data sets in our study.

Both the kth PC and membership probabilities in cluster k are vectors of length n with one entry for each individual. Each vector entry is associated with two geographic coordinates. To visualize how these vector values vary across geographical space, we performed spatial interpolation at a set of locations on a regular grid using the kriging method (exponential covariance model; Cressie 1993) and we displayed heat maps for the interpolated values of the PCs and assignment probabilities.

Results

We applied PCA and AM to simulated data sets generated under several demographic models of expansion of the Neolithic farmers in Europe. In these simulations, we modeled demic or cultural diffusion of agriculture with and without admixture between early farmers and resident hunter-gatherers.

Demic Diffusion: Models without Interbreeding

We began our study with spatial scenarios of Neolithic demic expansion in Europe in which there was no admixture between the expanding population and the resident population. Under these conditions, visual inspection of the results reveals that the PC1 maps exhibit continuous gradients for a large majority of the simulated data sets. Remarkably, in 19 of the 20 simulations that ended with 100% of Neolithic ancestry in the European gene pool (full replacement), the gradients are oriented along an axis that starts from the southwest and ends in the northeast of Europe (SW–NE axis, fig. 2A and pattern 1 in table 1). This axis is perpendicular to the direction of expansion that runs along a southeast-to-northwest axis. In order to see if this unexpected result was due to the contours of the European continent, we simulated expansions from the southwest of Europe (source in the center of Spain). We chose southwest Europe not because it is a likely origin for the settlement of Europe but to see how in simulations, the origin of an expansion affects resultant PCA patterns. In this case, we find NW-to-SE gradients in the PC1 map, which are again perpendicular to the main direction of the expansion (SW to NE, 10 of 10 simulations; fig. 2C). For both sources of expansion, PC2 maps generally highlight the regions of Scandinavia (figs. 1C and 2A) and PC3 the British Isles, which presumably reflects their geographic isolation in our simulated habitat (see below for further discussion).

When we ran the AM for $K = 2$ clusters, the resulting assignment probability maps showed patterns that are strongly

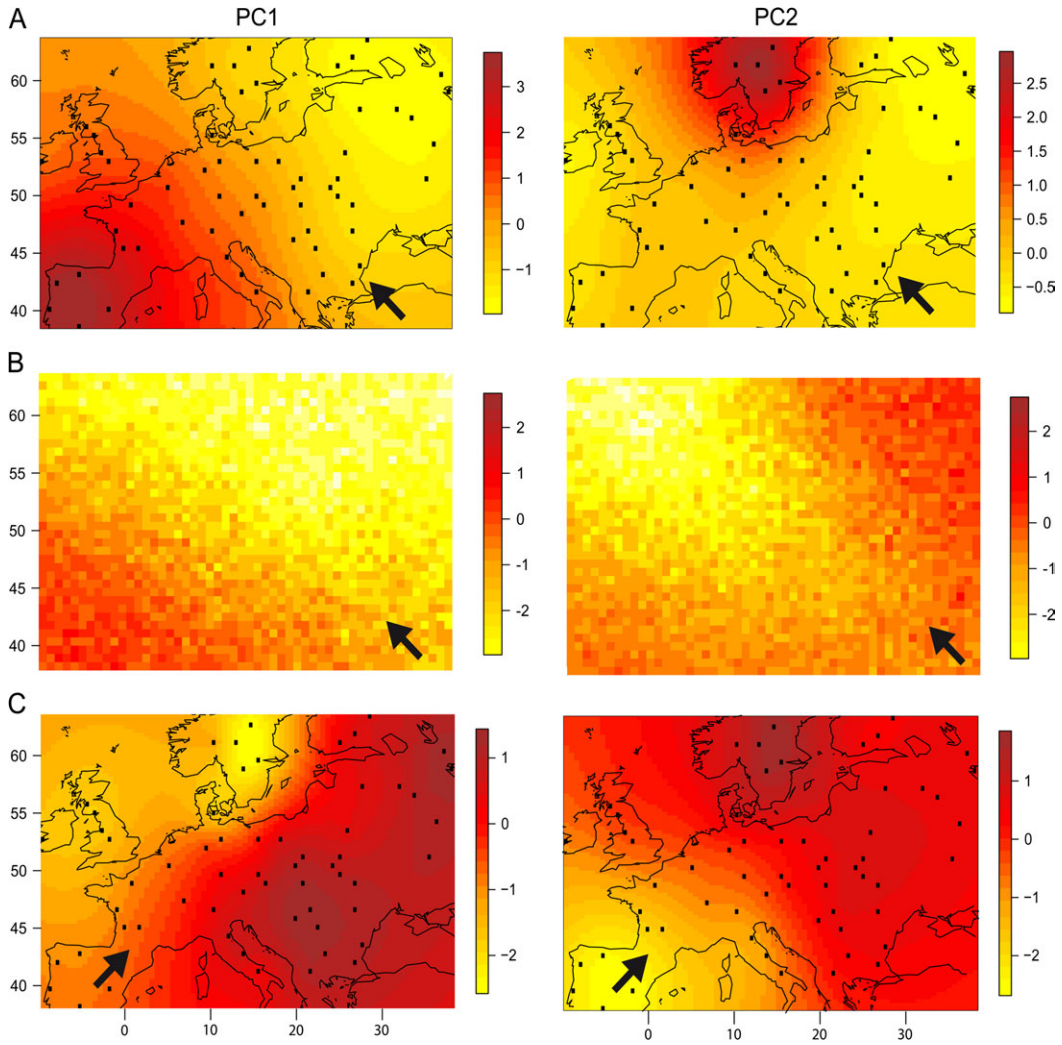

Fig. 2. PC1 and PC2 maps. (*A*) Data set simulated under a spatially realistic scenario of demic diffusion of Neolithic farmers in Europe without interbreeding with Paleolithic residents (100% of Neolithic ancestry in current genomes). (*B*) Range expansion on a regular lattice starting from the bottom-right corner. Ten individuals are sampled in each of the 64 × 42 demes. The average values of the first two eigenvectors are displayed for each deme. (*C*) Data set simulated under a scenario of an hypothetic demic expansion originating in the center of the Iberian Peninsula. Time of origin $T = 400$ generations ago, migration rate $m = 0.5$, growth rate $r = 0.5$, carrying capacity $N = 500$, no admixture with resident populations. The arrows indicate the origin of the expansion.

similar to those observed in PC1 maps, with membership probability in one of the two clusters decreasing along a SW–NE axis (supplementary fig. S1*A* and *B*, Supplementary Material online). AM maps for $K = 3$ clusters exhibit features similar to the PC1 and PC2 maps, showing one cluster either in Scandinavia or in the British Isles and two other partitioning the European mainland along the SW–NE axis (supplementary fig. S1*C* and *D*, Supplementary Material online).

One concern might be that the unexpected result is influenced by the specific set of 60 sampling locations or by the habitat shape and friction surfaces used for the simulations. To investigate this possibility, we ran additional simulations on a lattice of the same size as implemented in our spatial simulations for expansions starting from the southeast but with uniform migration rates and regular sampling across space. In addition, we sampled the complete set of 2,688 demes, with 10 individuals per deme. For a majority of the tested combinations of the model parameters, the first PC separates southwestern populations from northeastern ones. Again, this direction is perpendicular to the main axis of expansion. An example of this typical pattern is shown in figure 2*B*, for $m = 0.2$, $r = 0.5$, $C = 100$, and $T = 1{,}000$ (C is the carrying capacity of each deme, T is the number of generations since the onset of the expansion). In all the 36 simulations, PC2 showed a gradient running in the direction orthogonal to that apparent in the PC1 map. The pattern visible in PC1 consistently changed over all replicates from a SW–NE to an EW gradient when T increased, and the gradients in the maps of PC1 and PC2 become weaker and eventually nonexistent as genetic variation homogenizes across the habitat with time (example replicates shown in supplementary fig. S2, Supplementary Material online). For example, this happens when the age of the expansion is set to $T = 1{,}000$ generations, and when the migration rate is simultaneously increased to $m = 0.5$ implying that Cm was greater than 50 (supplementary fig. S2, Supplementary

Table 1. Frequency of observed patterns in PC1 maps for simulations of Neolithic range expansions from the southeast (80 replicates).

Final Levels of Neolithic Ancestry (%)	Pattern 1 (SW–NE gradient)	Pattern 2 (W–E gradient)	Pattern 3 (SE–NW gradient)	Other Patterns
Paleolithic expansions from the southeast				
100	**9/10**	**1/10**		
80	**8/10**			**2/10**
20	**9/10**	**1/10**		
10	**1/10**	**9/10**		
Paleolithic expansions from the southwest				
100	**10/10**			
80	**7/10**	**1/10**		**2/10**
20	**7/10**	**2/10**		**1/10**
10			**10/10**	

NOTE.—The top panel of results is for the case where Paleolithic expansions were modeled from the southeast and the bottom panel for the case with expansions from the southwest. The first row within each panel (100% Neolithic ancestry) corresponds to demic Neolithic expansions without admixture with resident Paleolithic populations. Subsequent rows (<100% final Neolithic ancestry) are for Neolithic expansions in which there was admixture with resident Paleolithic populations. Pattern 1 displays a SW–NE gradient. Pattern 2 displays an east–west gradient. Pattern 3 exhibits a gradient from the SE to the NW. A summary of the three patterns can be found in figure 4. The other patterns observed in PC1 represent clusters in Scandinavia or in the British Isles. The bolded values represent the number of simulation replicates exhibiting one of 3 typical patterns.

Material online). For the lowest values of T, m, and C (T = 500, m = 0.2, and C = 100), we find that the direction of the PC1 gradient is variable from replicate to replicate—aligning with the expansion ~50% of the time. This phenomenon is reminiscent of variation in the direction of PC1 observed amongst replicate simulations from equilibrium stepping-stone models in which there is no directional spatial pattern in the data (see supplementary fig. S1, Supplementary Material online, of Novembre and Stephens 2008) and was not observed after restoring the European habitat shape for the same demographic parameters or after increasing the carrying capacities to C = 500.

For these simulations, we also simulated sequence data sets consisting of 2,000 loci of 200 bp each. The mutation rate, equal to 10^{-7}/bp/generation, is a comparable rate of novel mutant alleles as having a more realistic mutation rate of 10^{-8}/bp distributed in 2,000 nonrecombining sequences of 2 kb. We measured the extent of isolation-by-distance for m = 0.2, r = 0.5, C = 100, and T = 1,000. Isolation-by-distance was assessed by regressing the logarithm of genetic distances [measured as $F_{ST}/(1 - F_{ST})$] between pairs of samples on the logarithm of their geographic distances (Slatkin 1993), where F_{ST} was obtained according to the definition of Hudson et al. (1992). Figure 3 provides evidence that genetic distances increased significantly faster with geographic distances along the transect perpendicular to the expansion than along the direction of the expansion ($P < 10^{-9}$).

Admixture Models: Interbreeding with Paleolithic Residents

We next examined what would happen if there was any interbreeding between an expanding Neolithic population from the southeast and a resident Paleolithic population.

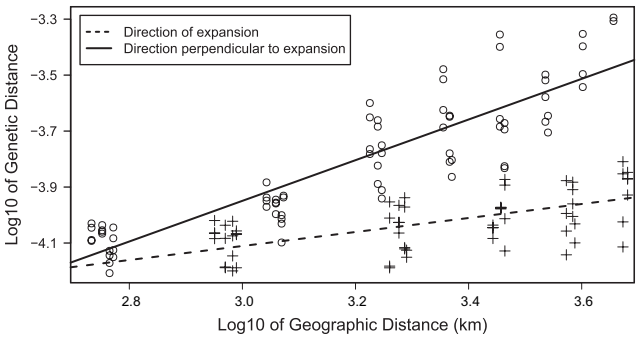

FIG. 3. Isolation by distance. Regression of genetic distance, computed as $F_{ST}/(1-F_{ST})$, on the logarithm of geographic distance for a simulation of range expansion on a regular lattice (start from the bottom-right corner, no admixture). Dashed line: demes in the direction of expansion (main diagonal of the habitat). Solid line: demes in the direction perpendicular to expansion (second diagonal).

To this aim, we reproduced the framework and the choice of parameters of Currat and Excoffier (2005) to simulate the genetic impact of the Neolithic transition. Briefly, the first expansion started around 1,500 generations ago from the Near East on a first layer of demes representing Paleolithic hunter-gatherer populations and covering all Europe. We specified levels of local gene flow between the resident and the invading populations so that the Paleolithic genes represented ~20%, 80% or more than 90% of the current European gene pool. These proportions were computed by the program SPLATCHE at each sampling location and then averaged over the sampling area. In ~77.5% (31 of 40) of the simulations ending with 20% or 80% of Paleolithic ancestry, we observe patterns similar to those obtained under a pure Neolithic demic diffusion model (table 1 and fig. 4*A*). In other words, PC1 exhibits a gradient along the SW–NE axis that runs perpendicular to the Neolithic expansion axis. As previous simulations have revealed the existence of gradients of admixture along the Neolithic expansion axis (Currat and Excoffier 2005), we computed maps of the fraction of Neolithic ancestry in current populations (fig. 5). These maps represent the local proportions of Neolithic genes in the European genetic pool. In the examples with 20% and 80% of final Paleolithic ancestry, we obtain a gradient of introgression along the direction of Neolithic expansion (see fig. 5 for the case of a final Paleolithic contribution equal to 80%). Thus, this pattern can occur at the same time as a PC gradient is running perpendicular to the same axis. We also observed a similar behavior for genetic diversity, computed as the (average) variance in microsatellite allele size, which displays a gradient running along the recent expansion axis (supplementary fig. S3, Supplementary Material online). In conclusion, if the proportion of ancient lineages in the current genetic pool is not very high (<80%), the direction of PC1 gradient is found to be perpendicular to the most recent (Neolithic) expansion.

When the local levels of interbreeding are higher than γ ~6–7%, we get two categories of patterns that depend on

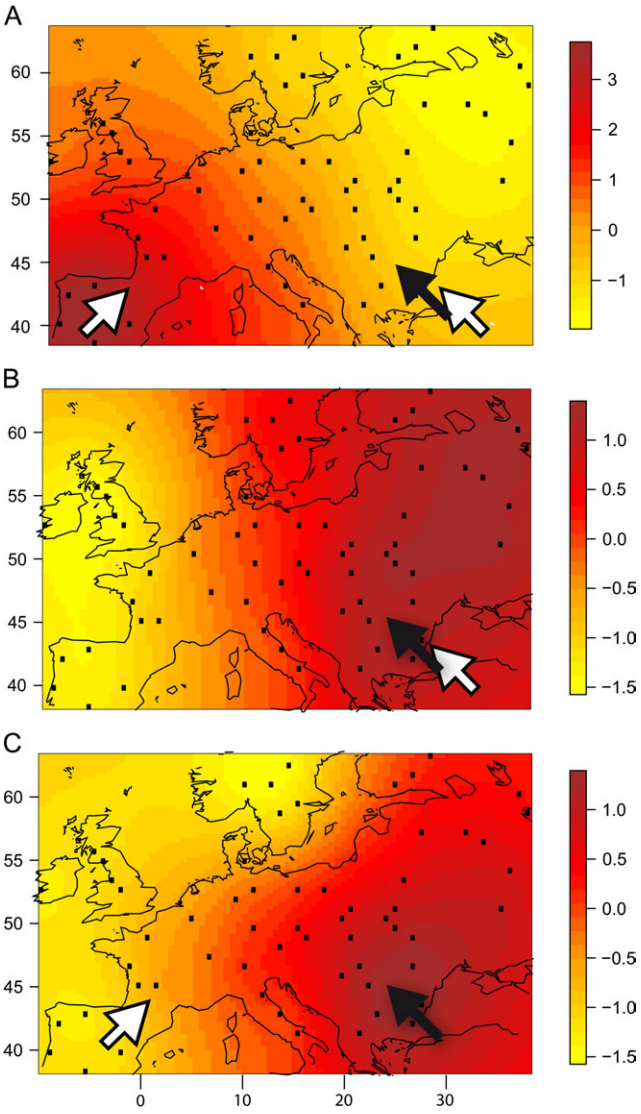

FIG. 4. The three main patterns observed in PC1 maps under spatially realistic models of the demic expansion of Neolithic farmers in Europe with admixture with resident hunter-gatherer populations. (*A*) Simulations with more than 20% of Neolithic ancestry in current genomes (Paleolithic expansions starting either from the SE or from the SW). (*B*) Simulations with less than 20% of Neolithic ancestry in current genomes (Paleolithic expansion from the southeast). (*C*) Simulations with less than 20% of Neolithic ancestry in current genomes (Paleolithic expansion from the southwest). The black arrows indicate the origin of the Neolithic expansion, and the white arrows indicate the origin of the Paleolithic expansion.

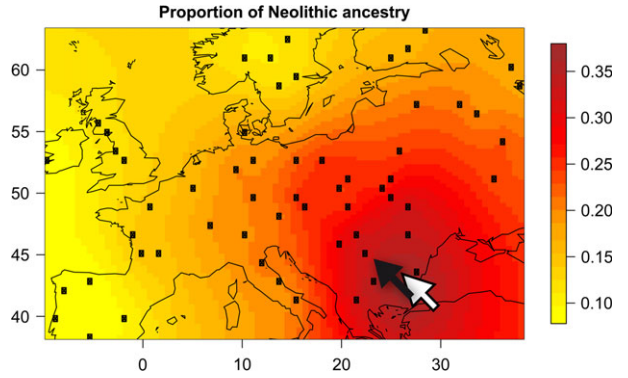

FIG. 5. Proportions of Neolithic ancestry in current genomes. Simulation with 20% of Neolithic average contribution. Similar maps were obtained regardless of the Paleolithic origin of the residents.

where the Paleolithic expansion took place: 1) Assuming a Paleolithic expansion that starts from the SW at the onset of the last glacial maximum (~20,000 years ago) and a Neolithic expansion that starts from the SE, the gradients in the PC1 map align with the main direction of Neolithic expansion (SE–NW axis; table 1 and fig. 4*C*); 2) When both the ancient and the recent expansions start from the SE (arrival of modern humans followed by the Neolithics), then the direction of PC1 gradients is along the east–west axis in most simulations (9 of 10 simulations; table 1 and fig. 4*B*). For these simulations with proportions of Paleolithic ancestry in current genome ancestry reaching values ~90%, the patterns of genetic variation in the current populations are more influenced by the Paleolithic population and where it expanded from than by Neolithic movements. In agreement with this, in cases where the Paleolithic expansion is from the SE, the PC gradients along the EW axis are similar to those obtained under an ancient expansion from the SE (supplementary fig. S4, Supplementary Material online). Likewise, the gradients of genetic diversity do not run parallel to the direction of the most recent expansion, but in the direction of the most ancient one (supplementary fig. S3, Supplementary Material online).

Discussion

Gradients in PC1 Are Often Perpendicular to the Main Direction of Expansion

In our computer simulations of the colonization of Europe by southeastern populations of early farmers, we observe gradients in PC1 maps in agreement with a spatial structuring of genetic variation across the continent. An important and striking result is that when the local rates of admixture between Neolithic colonists and Paleolithic residents are low, these gradients are consistently oriented in a direction perpendicular to the axis of the Neolithic expansion, rather than along the same axis as the expansion. Another important result is that when the final genetic pool is highly introgressed by the ancient (Paleolithic) population (>80% introgression), we found the PC1 gradient to be perpendicular to the direction of the Paleolithic expansion and as a result can in some cases be parallel to the direction of the most recent (Neolithic) expansion. For example, if there has been an ancient expansion from a southwestern refugium and the level of Neolithic ancestry in the current gene pool is less than 20%, our results show that PC1 evidences a SE-to-NW gradient.

To confirm these results, we ran simulations of expansions in a homogeneous environment and found that PC1 maps again showed a gradient running perpendicular to the expansion front, just as in simulations including

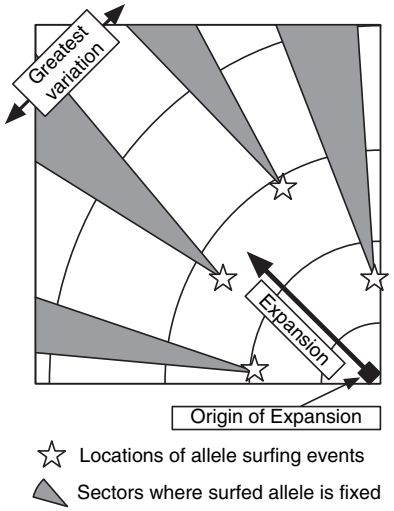

FIG. 6. Recurrent founder effects during range expansions create sectors where one allele is completely fixed, whereas the same allele is absent elsewhere. These regions have approximately conic shapes, and they increase genetic differentiation along the axis perpendicular to the direction of expansion.

realistic environmental features. The results suggest that PC1 map patterns are due to the process of expansion, rather than being an artifact of the geographical constraints we simulated. Assignment programs with $K = 2$ clusters also inferred gradients of probability that run perpendicular to the expansion axis, validating the PC1 gradient as an important axis of differentiation. Finally, by measuring the extent of genetic differentiation as function of distance, we found a stronger extent of genetic differentiation on an axis perpendicular rather than along the expansion axis. It thus seems that a gradient perpendicular to the expansion is not an artifact of a given method but that it rather reflects a true main underlying axis of differentiation among the populations. The question arises as to why differentiation would be perpendicular to the direction of expansion.

The Surfing Phenomenon

One possible explanation for the direction of the gradients we observed is the "allele surfing phenomenon" (Edmonds et al. 2004; Klopfstein et al. 2006; Excoffier and Ray 2008). In the surfing phenomenon, the repeated founder effects that occur at the edge of an expansion wave create conditions for low-frequency alleles to "surf" to higher frequencies and even to fixation at the wave front. As the wave moves forward, large patches of habitat become colonized with the "surfing" allele and form "sectors" of low genetic diversity at a given locus (Hallatschek et al. 2007). These sectors are often fixed for an allele that has low frequency elsewhere in the habitat, leading to strong differentiation between sectors (Hallatschek et al. 2007; Hallatschek and Nelson 2008). Because these sectors are aligned along the direction of expansion, there is actually the potential for substantial differentiation "perpendicular" to the axis of expansion (as illustrated in fig. 6; see also Excoffier and Ray 2008).

Common Allele–Frequency Distributions

To investigate whether common alleles show patterns consistent with surfing, we examined a particular data set from our geographic simulations (1,200 individuals, 60 samples, no interbreeding). In these simulations, we generated sequence data (400 kb per individual distributed evenly over 2,000 independent loci, mutation rate $= 10^{-7}$/bp/generation). For these data, we obtained 10,581 segregating sites with a frequency spectrum highly skewed toward low-frequency alleles (fig. 7A). The high frequency of singletons (ca. 80%) indicates a strong departure from the constant-size neutral frequency spectrum, for which the expected value is around $0.14 : \left(\sum_{1}^{600} \frac{1}{i}\right)^{-1}$. When PC1 was computed from the loci with minimum allele frequency (MAF) >10, the synthetic map was not different from the result obtained with all the data, although these high-frequency mutations occur only at a small fraction of the polymorphic sites (108 sites; fig. 7B–C). In contrast, when PC1 was computed from the sites with MAF less than 10, a strikingly distinct picture emerged, displaying an optimum at the center of the area (fig. 7D). This suggests that the PC1 gradient is driven strongly by the geographic distribution of the common alleles, many of which are likely to have become common due to allele surfing (Currat and Excoffier 2005; Excoffier and Ray 2008). To study the common alleles in more detail, we generated allele-frequency maps for the most common mutations (MAF > 30). We found that their spatial distributions exhibit regions where one allele was nearly absent and others where the same allele was completely fixed (supplementary fig. S5, Supplementary Material online). These regions have approximately conic shapes, and they approximate the sectors described by Hallatschek et al. In geographically explicit simulations, sectors of high frequencies were also observed in areas accessible only through the narrow bridges in Scandinavia and in the British Isles, where spatial bottlenecks might have reinforced genetic drift.

How Likely Is Allele Surfing to Be a Determinant of Genetic Structure?

The question arises of whether allele surfing is an exceptional phenomenon that only occurs due to our specific simulated parameter values or if it is expected to play a role in real populations. The probability of surfing alleles depends on many factors, including the amount of local diversity (Edmonds et al. 2004), the demographic parameters (Klopfstein et al. 2006), potential admixture with resident populations during the expansion phase (Currat et al. 2008), and geographical heterogeneity (Burton and Travis 2008). For conditions approximating a mutation rate of 10^{-8}/bp/generation, we find that about one mutation per 100 kb has a chance to reach a final frequency over 20% (fig. 7A). Although these surfing mutations represent less than 1% of the total number of all mutations, this small fraction of high-frequency mutations seems to dominate the variability represented by PC1. Although a rare phenomenon in our simulations, surfing indeed deeply influences the patterns uncovered in PC or AM maps. In

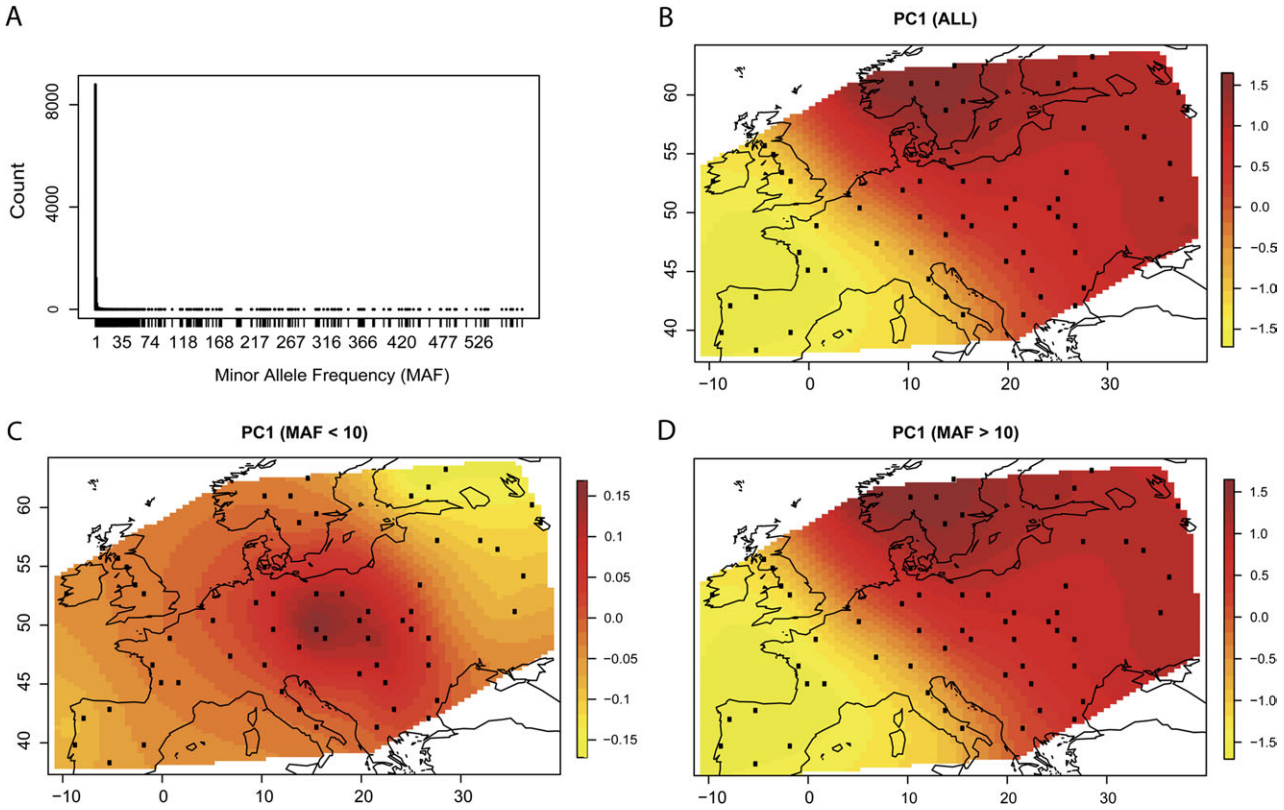

FIG. 7. Common alleles carry out population structure. Sequence data including 2,000 unlinked sequences of length 200 bp simulated under a regular lattice (1,200 individuals, 60 samples, mutation rate $= 10^{-7}$/bp/generation). (*A*) Folded frequency spectrum computed from more than 10,000 polymorphic sites. (*B*) PC1 map for all polymorphic sites. (*C*) PC1 map for sites with MAF < 10. (*D*) PC1 map for sites with MAF > 10.

addition, as surfing is not restricted to mutations arising during the expansion phase, rare alleles present in the gene pool of the expanding population, and those introduced via introgression also have the possibility to produce surfing patterns. Further, as the size of the population at the source of the expansion is rather small ($C = 100$ for the hunter-gatherers and $C = 1{,}000$ for the farmers), most mutations have occurred during or after the expansion in our simulations, whereas a large fraction of the mutations present in current European populations originated in Africa and were therefore already present in the populations having initially colonized Europe. It follows that the surfing of standing variants has probably been underestimated in our simulations.

Effects of Geographic Constraints

When a realistic geography of Europe is taken into account in the simulation, PC maps often reveal strong differentiation at the edge of the range of the expansion, typically in Scandinavia, or less frequently in the British Isles and in the Iberian Peninsula. The very common Scandinavian cluster does not persist when geographic constraints are removed and when simulations are performed into a uniform environment. Clusters arising at the edge of the continental area might be interpretable as a combination of the effects of isolation-by-distance and the effects of geographic bottlenecks, like land-bridges across seas or corridors in mountain ranges (Burton and Travis 2008). The narrow land bridge we have introduced to connect the south of Sweden to Denmark is likely to lead to increased genetic drift and founder effects and thus to differentiation between populations on opposite sides of the Baltic Sea. Note that the Scandinavian and the British populations were also identified in separate clusters by the AM programs. A second point is that we do not expect our results to hold in long rectangular (approximately one dimensional) habitats. Indeed, in expansions into linear habitats, the wave front is necessarily very narrow, and it will be difficult for sectors to form, and so alleles that surf will likely not be distributed in patches perpendicular to the axis expansion.

Criticisms of the Simulation Model

Although our simulation model realistically accounts for contours and geographic barriers in Europe, it is not meant to be a very detailed model of European prehistory. First of all, there is unavoidable uncertainty about the parameters used for characterizing population densities, rates of expansion, and rates of migration (see Rowley-Conwy 2009). Events that occurred at small geographic scales, like fluctuations in carrying capacities due to variation in resource availability or due to local changes in the environment, are ignored. Thus, it is possible that the model fails to reproduce every particular aspect of local genetic diversity. However, the simulation model is still useful for giving interesting insights as it captures large-scale temporal and spatial aspects of European prehistory, including, for

example, the timing of the spread of agriculture, the relative densities of hunter-gatherer and farmer populations, and admixture between hunter-gatherer and farmer populations. Previous work has shown how expansions with admixture produce clines in the proportion of Neolithic ancestry that sensibly follow the direction of expansion (Currat and Excoffier 2005), and we show here how diversity decreases as one moves along the direction of the expansion. Both these patterns are expected in expansion models, and they suggest that the simulations, which also reproduce observed patterns (Chiaroni et al. 2009), are meaningful. In this framework, the PC maps described in this study are robustly observed under a wide range of model parameters.

Implications for the Interpretation of Human Genetic Variation in European Populations

For some time, population geneticists have been attempting to reconstruct the ancient demographic history of the Europeans, and it has been the source of considerable debate (e.g., Barbujani and Goldstein 2004; Jobling et al. 2004). Major ancestral processes that have been suggested are an initial Paleolithic colonization, later re-expansions from southern refugia, the Neolithic dispersal of early farmers, or trans-Mediterranean gene flow. The relative importance of these events for explaining standing patterns of genetic variation is however difficult to assess from archaeological data. As a result, a variety of genetic data sets and analysis methods have been used to study this problem. Our goal here has been to clarify how to best interpret results produced by PCA analysis, one particular exploratory tool used in this long debate. We found that at odds with conventional wisdom, the gradient in PC1 can orient perpendicular to the direction of an expansion under a wide range of conditions. It thus appears that NW–SE gradients previously observed in PC1 plots of Europe are inconsistent with many simple models of Paleolithic or Neolithic expansions from the Near East. The simulation results might suggest a role of expansions from southwestern refugia after the last glacial maximum. However, Heath et al. (2008) found PC1 to align with an E–W gradient in Europe, Lao et al. (2008) found a PC1 gradient that ran N–S, NW–SE gradients were observed by Menozzi et al. (1978) and by Cavalli-Sforza et al. (1994), and a NNW–SSE gradient was observed by Novembre et al. (2008). The direction of PC gradients is difficult to interpret due to the influence of the sampling scheme (Novembre and Stephens 2008; McVean 2009). Because of these uncertainties, we withhold making conclusions and suggest that future progress will occur by more directly looking at spatial patterns of variation in Europe (such as potential sector patterns) in place of methods such as PCA.

The simulation study we conducted here gives two general insights about spatial patterns of variation that might be observed under models of population expansions: 1) Spatial patterns of genetic variation (gradients/clines) can arise under a broad range of expansion scenarios, just as they do in equilibrium isolation-by-distance models; 2) There can be substantial differentiation along an axis perpendicular to the direction of an expansion, presumably due to allele surfing. Many studies have shown that gradients in variation exist across Europe (Menozzi et al. 1978; Sokal and Menozzi 1982; Sokal et al. 1989; Barbujani and Pilastro 1993; Chikhi et al. 1998; Rosser et al. 2000; Chikhi et al. 2002; Dupanloup et al. 2004), most recently finding that such gradients even exist at spatial scales on the order of hundreds of kilometers (Bauchet et al. 2007; Heath et al. 2008; Lao et al. 2008; Novembre et al. 2008; Tian et al. 2008; Price et al. 2009; Sabatti et al. 2009).

Evidence for the directionality of spatial patterns is more difficult to summarize as substantial differences exist across studies, and one needs to be specific about exactly what aspect of variation is being observed. Using directional correlograms, Sokal et al. (1989) show many loci consistent with NW–SE clines (particularly human leukocyte antigen loci), but other loci show evidence for other directional clines. Inferences of the proportion of Neolithic ancestry have shown both patterns that decay with distance from the Near East using Y-chromosome markers (Chikhi et al. 2002) as well as East–West patterns using eight loci (Dupanloup et al. 2004). The recent availability of large-scale SNP data helps alleviate concerns about making inferences from a small number of loci and promises to reveal more consistent genomewide patterns. In this vein, two more recent large-scale SNP-based studies (Lao et al. 2008; Auton et al. 2009) have both observed a gradient in levels of haplotype diversity and linkage disequilibrium that are roughly north–south and with high levels of diversity in the Italian and Iberian Peninsulas. Notably, these patterns seem unexpected under a demic diffusion model from the Near East and are more consistent with an impact of trans-Mediterranean gene flow (Auton et al. 2009), larger population sizes in the southwest (Lao et al. 2008) or with hypotheses of southern glacial refugia. However, the analysis of high-throughput SNP data from European populations is still in an exploratory phase. Further work, for instance, looking specifically for patterns of variation consistent with sectors generated by surfing alleles, will likely shed more light on the genetic history of European populations. As always, the results will need to be integrated with other approaches, and an additional promising avenue of work is ancient DNA analyses. Comparing mitochondrial DNA sequences from 20 hunter-gatherer skeletons with those from modern Europeans, Bramanti et al. (2009) found that most of the ancient hunter-gatherers in Central Europe share haplotypes that are rare in Europeans today, perhaps pointing toward a highly dynamic history of human population movements in Europe.

Conclusions

A previous study showed that the original patterns observed in PCA might not reflect any expansion events (Novembre and Stephens 2008). Here, we find that under very general conditions, the pattern of molecular diversity produced by an expansion may be different than what was

expected in the literature. In particular, we find conditions where an expansion of Neolithic farmers from the southeast produces a greatest axis of differentiation running from the southwest to the northeast. This surprising result is seemingly due to allele surfing leading to sectors that create differentiation perpendicular to the expansion axis. Although a lot of our results can be explained by the surfing phenomenon, some interesting questions remain open. For example, the phase transition observed for relatively small admixture rates between Paleolithic resident and Neolithic migrant populations occurs at a value that is dependent on our simulation settings, and further investigations would be needed to better characterize this critical value as a function of all the model parameters. Another unsolved question is to know why the patterns generally observed in PC2 maps for our simulation settings sometimes arise in PC1 maps instead. These unexplained examples remind us that PCA is summarizing patterns of variation in the sample due to multiple factors (ancestral expansions and admixture, ongoing limited migration, habitat boundary effects, and the spatial distribution of samples). In complex models such as our expansion models with admixture in Europe, it may be difficult to tease apart what processes give rise to any particular PCA pattern. Our study emphasizes that PC (and AM) should be viewed as tools for exploring the data but that the reverse process of interpreting PC and AM maps in terms of past routes of migration remains a complicated exercise. Additional analyses—with more explicit demographic models—are more than ever essential to discriminate between multiple explanations available for the patterns observed in PC and AM maps. We speculate that methods exploiting the signature of alleles that have undergone surfing may be a powerful approach to study range expansions.

Supplementary Material

Supplementary figures S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

N.R. and L.E. were partially supported by Swiss National Science Foundation (NSF) grants 3100-112072 and 3100-126074 to L.E. M.C. was supported by Swiss NSF grant 3100A0-112651 to Alicia Sanchez-Mazas whom we thank for her support. O.F. was partially supported by a French Agence Nationale de la Recherche grant BLAN06-3146282 MAEV, and he thanks the IXXI Institute of Complex Systems. J.N. was supported by the Searle Scholar Program. J.N. and E.H. were supported by NSF grant 0733033.

References

Ammerman AJ, Cavalli-Sforza LL. 1984. The neolithic transition and the genetics of populations in Europe. Princeton (NJ): Princeton University Press.

Auton A, Bryc K, Lohmueller KE, et al. (13 co-authors). 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res*. 19:795–803.

Barbujani G, Sokal RR, Oden NL. 1995. Indo-European origins: a computer-simulation test of five hypotheses. *Am J Phys Anthropol*. 96:109–132.

Barbujani GG, Goldstein DB. 2004. Africans and Asians abroad: genetic diversity in Europe. *Annu Rev Genomics Hum Genet*. 5:119–150.

Barbujani GG, Pilastro A. 1993. Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. *Proc Natl Acad Sci U S A*. 90:4670–4673.

Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD. 2007. Measuring European population stratification with microarray genotype data. *Am J Hum Genet*. 80:948–956.

Beaumont MA, Rannala B. 2004. The Bayesian revolution in genetics. *Nat Rev Genet*. 5:251–261.

Bramanti B, Thomas MG, Haak W, et al. (16 co-authors). 2009. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* 26(5949):137–140.

Burton OJ, Travis JM. 2008. Landscape structure and boundary effects determine the fate of mutations occurring during range expansions. *Heredity* 101(4):329–340.

Cavalli-Sforza LL, Edwards AWF. 1963. Analysis of human evolution. In: Geerts SJ, editor. Genetics today: Proceedings of the 11th International Congress of Genetics, The Hague, The Netherlands. New York: Pergamon. Vol. 3. p. 923–993

Cavalli-Sforza LL, Menozzi P, Piazza A. 1993. Demic expansions and human evolution. *Science* 259:639–646.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton (NJ): Princeton University Press.

Chen C, Durand E, Forbes F, François O. 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Mol Ecol Notes*. 7:747–756.

Chiaroni J, Underhill PA, Cavalli-Sforza LL. 2009. Y chromosome diversity, human expansion, drift, and cultural evolution. *Proc Natl Acad Sci U S A*. 106:20174–20179.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G. 1998. Clines of nuclear DNA markers suggest a recent Neolithic ancestry of the European gene pool. *Proc Natl Acad Sci U S A*. 95:9053–9058.

Chikhi L, Nichols RA, Barbujani G, Beaumont MA. 2002. Y genetic data support the Neolithic demic diffusion model. *Proc Natl Acad Sci U S A*. 99:11008–11013.

Cressie NAC. 1993. Statistics for spatial data. Wiley, New York.

Currat M, Excoffier L. 2004. Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biol*. 2:2264–2274.

Currat M, Excoffier L. 2005. The effect of the Neolithic expansion on European molecular diversity. *Proc R Soc B*. 272:679–688.

Currat M, Ray N, Excoffier L. 2004. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol Ecol Notes*. 4(1):139–142.

Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. *Evolution* 62:1908–1920.

Davies N. 1998. Europe: a history. Harper Perennial, New York.

Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. *Science* 300:597–603.

Dupanloup I, Bertorelle G, Chikhi L, Barbujani G. 2004. Estimating the impact of prehistoric admixture on the genome of Europeans. *Mol Biol Evol* 21:1361–1372.

Durand E, Jay F, Gaggiotti OE, François O. 2009. Spatial inference of admixture proportions and secondary contact zones. *Mol Biol Evol*. 26:1963–1973.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci U S A*. 101:975–979.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends Ecol Evol*. 23:347–351.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR. 2007. Genetic drift at expanding frontiers promotes gene segregation. *Proc Natl Acad Sci U S A*. 104:19926–19930.

Hallatschek O, Nelson DR. 2008. Gene surfing in expanding populations. *Theor Popul Biol*. 73:158–170.

Harpending HC, Jenkins T. 1973. Genetic distance among southern African populations. In: Crawford M, Workman P, editors. Method and theory in anthropological genetics. Albuquerque (NM): University of New Mexico Press.

Heath SC, Gut IG, Brennan P, et al. (27 co-authors). 2008. Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet* 16:1413–1429.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132(2):583–589.

Jakobsson M, Scholz SW, Scheet P, et al. (24 co-authors). 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003.

Jobling MA, Hurles ME, Tyler-Smith C. 2004. Human evolutionary genetics: origins, peoples and disease. London: Garland Science Publishing.

Klopfstein S, Currat M, Excoffier L. 2006. The fate of mutations surfing on the wave of a range expansion. *Mol Biol Evol*. 23:482–490.

Lao O, Lu TT, Nothnagel M, et al. (33 co-authors). 2008. Correlation between genetic and geographic structure in Europe. *Curr Biol*. 18:1241–1248.

Li JZ, Abscher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

McVean G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet*. 5(10):e1000686.

Mellars P. 2004. Neanderthals and the modern human colonization of Europe. *Nature* 432:461–465.

Mellars P. 2006. Archeology and the dispersal of modern humans in Europe: deconstructing the "Aurignacian". *Evol Anthropol*. 15:167–182.

Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* 201:786–792.

Novembre J, Johnson T, Bryc K, et al. (12 co-authors). 2008. Genes mirror geography within Europe. *Nature* 456:98–101.

Novembre J, Stephens M. 2008. Interpreting principal components analyses of spatial population genetic variation. *Nat Genet*. 40:646–649.

Patterson NJ, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190.

Pinhasi R, Fort J, Ammerman AJ. 2005. Tracing the origin and spread of agriculture in Europe. *PLoS Biol*. 3:e410.

Price AL, Helgason A, Palsson S, Stefansson H, St Clair D, Andreassen OA, Reich D, Kong A, Stefansson K. 2009. The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet*. 5(6):e1000505.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 8:904–909.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Rendine S, Piazza A, Cavalli-Sforza LL. 1986. Simulation and separation by principal components of multiple demic expansions in Europe. *Am Nat*. 128:681–706.

Rosser ZH, Zerjal T, Hurles ME, et al. (63 co-authors). 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet*. 67:1526–1543.

Rowley-Conwy P. 2009. Human prehistory: hunting for the earliest farmers. *Curr Biol*. 19:R948–R949.

Sabatti C, Service SK, Hartikainen AL, et al. (25 co-authors). 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*. 41:35–46.

Slatkin M. 1993. Isolation-by-distance in equilibrium and non-equilibrium populations. *Evolution* 47:264–279.

Sokal RR, Harding RM, Oden NL. 1989. Spatial patterns of human gene frequencies in Europe. *Am J Phys Anthropol*. 80:267–294.

Sokal RR, Menozzi P. 1982. Spatial autocorrelation of HLA frequencies in Europe support demic diffusion of early farmers. *Am Nat*. 119:1–17.

Sokal RR, Oden NL, Thomson BA. 1999. A problem with synthetic maps. *Hum Biol*. 71:1–13.

Steele J, Adams JM, Sluckin T. 1998. Modeling Paleoindian dispersals. *World Archaeol*. 30:286–305.

Tian C, Plenge RM, Ransom M, et al. (11 co-authors). 2008. Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet*. 4:e4.