Supplementary Note

Visualizing spatial population structure with estimated effective migration surfaces

D Petkova[1,2], J Novembre[3] & M Stephens[1,3]

[1] *Department of Statistics, University of Chicago*

[2] *Wellcome Trust Centre for Human Genetics*

[3] *Department of Human Genetics, University of Chicago*

Contents

S1 Review of methods that relate genetic data to a geographic map

To clarify how EEMS relates to existing methods, we discuss various other approaches that make use of sampling locations in their analysis or interpretation of genetic data and that have as primary output a visual display of spatial patterns. Our goal is to provide a brief summary but also to emphasize the unique aspects of the EEMS approach to modeling geo-referenced genetic data.

One class of methods first analyze the genetic data to discover patterns of genetic similarities and then utilize the sampling information post hoc to display and interpret the results. For example, principal component analysis (PCA) of genetic data does not incorporate sampling locations but PCA results are often presented by coloring samples from the same sampling unit (e.g. country or region) in the same color. Similarly, Structure, in its original formulation (Pritchard et al., 2000), and FineStructure (Lawson et al., 2012) do not use the sampling locations to specify their cluster-based models of the genetic data, yet the outputs are often organized geographically. Structure infers individual ancestry proportions and these vectors are often arranged into a stacked bar chart so that the individuals from the same geographical unit are plotted together along the x-axis of the chart. FineStructure groups individuals into a hierarchy of genetic clusters. If precise geographic information is available, individuals are plotted on a map and colored according to their cluster membership; if the genetic clusters are geographically localized, spatial patterns of shared ancestry are often displayed in pie charts pinned to the map (Leslie et al., 2015). Spatial ancestry analysis (SPA) (Yang et al., 2012) can produce PCA-like scatter plots of inferred locations on a two-dimensional plane when sampling locations are unknown (or treated as unknown) and individuals are typically colored in a manner similar to PCA plots. Notably, though these approaches produce varied visual summaries, the geographic information is used only to guide interpretation and not taken into account quantitatively to fit the underlying model. (If sampling locations are known, SPA can also model geographically indexed genetic data to estimate spatial trends in allele frequency, in a framework very different from EEMS, as we discuss below).

A second class of methods, and the one that EEMS falls into, use an explicitly spatial model and analyze the genetic and the geographic data jointly. The methods in this class address very different inferential tasks, including spatially aware genetic clustering, e.g. (Guillot et al., 2005); interpolation of allele frequency surfaces and sample localization, e.g. (Yang et al., 2012; Wasser et al., 2004; Baran et al., 2013; Rañola et al., 2014); estimation of mean dispersal distances using spatial autocorrelation or allele distributions, e.g. (Epperson and Li, 1996; Novembre and Slatkin, 2009); and detection of non-homogeneous gene flow in continuous space, e.g. (Duforet-Frebourg and Blum, 2014). This list of tasks is extremely diverse and EEMS addresses only the last one. Other methods for analyzing local variation in dispersal include Monmonier's maximum difference algorithm (Manni et al., 2004), Wombling methods (Manel et al., 2007) and LocalDiff (Duforet-Frebourg and Blum, 2014). Monmonier's algorithm analyzes observed genetic differences but it requires that discrete groups be specified *a priori* to detect strongly differentiated pairs. Wombling methods can detect strong barriers without pre-specified clustering, by estimating spatial gradients in allele frequencies and identifying localized sharp discontinuities. LocalDiff (Duforet-Frebourg and Blum, 2014) uses a spatial Gaussian process to interpolate allele frequencies before assessing local patterns of differentiation among neighboring populations. The algorithmic flavor of these methods contrasts with EEMS, which directly models landscape inhomogeneity

through its effects on expected coalescent times.

We want to particularly emphasize the differences between EEMS, Geneland (Guillot et al., 2005) and SPA (Yang et al., 2012), since these methods all produce visual summaries of spatial patterns from geo-referenced data, and so may appear superficially similar. In fact the three methods have different inferential goals, and so their visualizations are not directly comparable. SPA (Yang et al., 2012) infers spatial variation in local allele frequencies. Such frequency surfaces may, of course, partly reflect complex non-homogeneous patterns of gene flow, but graphical representations of such an allele surface (one for each SNP) cannot be directly compared with an EEMS map, which attempts to highlight variations in rates of gene flow. Put another way, while allele frequency maps visualize spatial variation in allele frequency, EEMS maps attempt to visualize the *effective migration processes* that could have given rise to this spatial variation. Geneland (Guillot et al., 2005) tackles a different problem again: clustering genetic samples. Geneland uses the geographical location of each sample to help inform the clustering, and produces a map in which each location is colored according to its (most likely) cluster membership. This map illustrates the spatial distribution of genetically distinct groups but does not attempt to represent relationships between clusters. For example, there is no indication whether some clusters are more similar than others. And displaying genetic cluster membership on a map is, of course, completely different from displaying effective migration rates. We also note that while both Geneland and EEMS make use of Voronoi tessellations, they do so in fundamentally different ways. In Geneland, each Voronoi tile belongs to one of K genetic clusters that comprise the population, and the tessellation serves as a modeling device to favor cluster assignments where geographically close individuals are assigned to the same genetic cluster. In EEMS, each Voronoi tile has a characteristic migration rate, and the tessellation serves to favor inferences where geographically close areas have the same migration rate. The fact that both methods make use of Voronoi tessellations is, therefore, merely a superficial similarity. Of course, for any given geo-referenced dataset, all three of these tools – and more – may provide useful insights.

S2 Mathematical and computational details of the EEMS method

Here we provide technical details about the EEMS method for estimating effective migration and diversity rates from geo-referenced genetic data. In Section S2.1 we derive expressions for the expected genetic dissimilarity $\mathrm{E}\{D\}$ at a random SNP or microsatellite, as a function of expected coalescence time. In Section S2.2 we derive an approximation for the expected genetic dissimilarities in a spatial (stepping stone) model, in terms of effective resistances in an undirected weighted graph. In Section S2.3 we explain how to compute efficiently the Wishart likelihood $\ell(k, m, q, \sigma^2)$ for the degrees of freedom k, the effective migration rates m, the effective diversity rates q and the variance scale σ^2. In Section S2.4 we explain how the effective resistances R in an undirected weighted graph can be computed efficiently, for the purpose of evaluating the likelihood $\ell(k, m, q, \sigma^2)$. In Section S2.5 we describe how to use birth-death Markov Chain Monte Carlo to sample from the posterior distribution $\pi(k, m, q, \sigma^2 \,|\, D)$.

S2.1 Expected genetic dissimilarities in population genetics

We derive the expected genetic dissimilarity between two distinct individuals i and j as a function of their expected coalescence time T_{ij}. In population genetics, expected coalescence time can be considered a distance metric: the larger T_{ij} is, the more differentiated i and j are, as they would not share mutations either lineage accumulates after the split from their most recent common ancestor (MRCA). We consider a randomly selected genetic marker l, either a SNP or a microsatellite, in a haploid or a diploid species. Let Z_{il} denote the genotype of individual i at locus l. Then the squared difference $(Z_{il} - Z_{jl})^2$ is a measure of the genetic dissimilarity between i and j at the locus l.

Expected dissimilarities at a random SNP in a haploid species

Suppose that SNP l is genotyped in a sample of size n. Following (McVean, 2009), we condition on the event S that the SNP segregates in the sample and we take the limit $\theta \to 0$, where θ is the mutation rate. In a haploid species, $Z_{il} \in \{0, 1\}$ is the allele carried by individual i at SNP l and the event S is equivalent to observing exactly one mutation. The expected genetic difference $(Z_{il} - Z_{jl})^2$ is the probability that $Z_{il} \neq Z_{jl}$ and

$$\mathrm{E}\Big\{(Z_{il} - Z_{jl})^2 \,|\, S\Big\} = \Pr\Big\{Z_{il} \neq Z_{jl} \,|\, S\Big\} = \Pr\{Z_{il} \neq Z_{jl}\} / \Pr\{S\} \tag{S1a}$$

$$= \lim_{\theta \to 0} \frac{\mathrm{E}\Big\{(\theta t_{ij}) \exp(-\theta t_{ij})\Big\}}{\mathrm{E}\Big\{(\theta t_{\mathrm{tot}}/2) \exp(-\theta t_{\mathrm{tot}}/2)\Big\}} = \frac{2T_{ij}}{T_{\mathrm{tot}}}, \tag{S1b}$$

where the t's are random coalescence times and the T's are their expectations, under the coalescent process. Two haploid individuals i and j carry a different allele, $\{Z_{il} \neq Z_{jl}\}$, if and only if a mutation occurs on the path from i to j through the pair's MRCA, which is of expected length $2T_{ij}$. The denominator T_{tot} is the expected total length (sum of all branches) of a random genealogy that describes the history of all n individuals back to the sample's MRCA.

Expected dissimilarities at a random SNP in a diploid species

We model the genotype of a diploid individual i as the sum of two haplotypes i_1 and i_2. Thus $Z_{il} = Z_{i_1 l} + Z_{i_2 l} \in \{0, 1, 2\}$ where the subscript indicates one of two alleles. To derive the expected genetic dissimilarity between two distinct individuals i and j, we condition on the event S and take the limit $\theta \to 0$ as in the haploid case:

$$\mathrm{E}\Big\{Z_{i_1 l}^2 \,|\, S\Big\} = \mathrm{E}\Big\{Z_{i_1 l} \,|\, S\Big\} = T_{\mathrm{mrca}}/T_{\mathrm{tot}}, \tag{S2a}$$

$$\mathrm{E}\Big\{Z_{i_1 l} Z_{j_1 l} \,|\, S\Big\} = \Big(T_{\mathrm{mrca}} - T_{ij}\Big)/T_{\mathrm{tot}}, \tag{S2b}$$

where T_{mrca} denotes the expected coalescence time to the MRCA of all sampled individuals, i.e., the height of the average genealogy of the sample. By applying equations (S2a) and (S2b) repeatedly, we obtain:

$$\mathrm{E}\Big\{(Z_{il} - Z_{jl})^2 \,|\, S\Big\} = \mathrm{E}\Big\{\Big(Z_{i_1 l}^2 + Z_{i_2 l}^2 + Z_{j_1 l}^2 + Z_{j_2 l}^2\Big) + 2Z_{i_1 l} Z_{i_2 l} + 2Z_{j_1 l} Z_{j_2 l} - 2\Big(Z_{i_1 l} Z_{j_1 l} + Z_{i_1 l} Z_{j_2 l} + Z_{i_2 l} Z_{j_1 l} + Z_{i_2 l} Z_{j_2 l}\Big) \,|\, S\Big\}$$

$$= 4\frac{T_{\mathrm{mrca}}}{T_{\mathrm{tot}}} + 2\Big(\frac{T_{\mathrm{mrca}} - T_i}{T_{\mathrm{tot}}}\Big) + 2\Big(\frac{T_{\mathrm{mrca}} - T_j}{T_{\mathrm{tot}}}\Big) - 8\Big(\frac{T_{\mathrm{mrca}} - T_{ij}}{T_{\mathrm{tot}}}\Big) = 2\Big(\frac{4T_{ij} - T_i - T_j}{T_{\mathrm{tot}}}\Big), \tag{S3}$$

where $T_i = \mathrm{E}\{t_{i_1 i_2}\}$ and $T_j = \mathrm{E}\{t_{j_1 j_2}\}$ at a random SNP l.

Expected dissimilarities at a random microsatellite in a diploid species

At a microsatellite locus, an allele is coded as the "number of repeats" of a short DNA motif (two to six base pairs). The mutation process at this genetic variant can be modeled by a symmetric stepwise mechanism where the number of repeats increases or decreases by 1, with equal probability (Ohta and Kimura, 1973).

We model the genotype Z_{il} of a diploid individual at microsatellite l as the average of the two alleles, $Z_{i_1 l}$ and $Z_{i_2 l}$. Under the symmetric mutation process,

$$Z_{i_1 l} = a_l + \sum_{k=1}^{K_{i_1 l}} S_k, \tag{S4}$$

where a_l is the ancestral allele at microsatellite l (the allele carried by the MRCA of all lineages in the sample), $K_{i_1 l}$ denotes the number of mutations that occur on the lineage from haplotype i_1 to the MRCA, and the S_ks are independent binary random variables with $\Pr\{S_k = 1\} = \Pr\{S_k = -1\} = \frac{1}{2}$.

Let θ_l denote the mutation rate at microsatellite l. If we assume that the mutations S_k at marker l occur as a Poisson process with mutation rate θ_l, then $K_{i_1 l} \,|\, \theta_l, t_{\text{mrca}} \sim \text{Po}(\theta_l t_{\text{mrca}})$ (Hudson, 1990). Thus, we have:

$$\begin{aligned} \text{E}\Big\{Z_{i_1 l}^2 \,|\, \theta_l\Big\} &= a_l^2 + \text{E}\Big\{\text{E}\Big\{\Big(\sum_{k=1}^{K_{i_1 l}} S_k\Big)^2 \,|\, t_{\text{mrca}}\Big\} \,|\, \theta_l\Big\} = a_l^2 + \text{E}\Big\{\text{E}\Big\{\sum_{k=1}^{K_{i_1 l}} S_k^2 \,|\, t_{\text{mrca}}\Big\} \,|\, \theta_l\Big\} \\ &= a_l^2 + \text{E}\Big\{\text{E}\Big\{K_{i_1 l} \,|\, t_{\text{mrca}}\Big\} \,|\, \theta_l\Big\} = a_l^2 + \text{E}\Big\{\theta_l t_{\text{mrca}} \,|\, \theta_l\Big\} = a_l^2 + \theta_l T_{\text{mrca}}. \end{aligned} \tag{S5}$$

(We have used: $\text{E}\{Z_{i_1 l}\} = a_l + \sum_k S_k = a_l$ and $\text{E}\{S_k S_{k'}\} = 0$ since the S_ks have mean 0, variance 1 and are independent.) Similarly,

$$\text{E}\Big\{Z_{i_1 l} Z_{j_1 l} \,|\, \theta_l\Big\} = a_l^2 + \theta_l\Big(T_{\text{mrca}} - T_{ij}\Big). \tag{S6}$$

Now we can combine equations (S5) and (S6) to obtain:

$$\text{E}\Big\{(Z_{il} - Z_{jl})^2 \,|\, \theta_l\Big\} = \frac{\theta_l}{2}\Big(4T_{ij} - T_i - T_j\Big). \tag{S7}$$

Compared to equation (S3), there is a factor of 1/4 because we model a diploid genotype at a microsatellite locus as the *average* of two alleles rather than the *sum* of two alleles, as in the case of SNPs.

S2.2 Expected genetic dissimilarities in EEMS

EEMS is based on the stepping stone model (Kimura and Weiss, 1964), which specifies that the expected coalescence times between two distinct individuals depends only on their locations:

$$T_{ij} = T_{\delta(i)\delta(j)}, \tag{S8}$$

for individuals i and j drawn randomly from demes $\delta(i)$ and $\delta(j)$, respectively.

Let D be the matrix of observed genetic differences: $D = (D_{ij}) = \Big((Z_{il} - Z_{jl})^2\Big)$ at a randomly selected polymorphic genetic marker l. If $i = j$, both the observed and the expected dissimilarity with self is 0. If $i \neq j$, the expected genetic dissimilarity is given by equations (S1), (S3) and (S7):

$$\text{E}\{D_{ij} \,|\, *\} = \begin{cases} \sigma^2 T_{\delta(i)\delta(j)} \text{ at haploid SNP}; & \sigma^2\Big(4T_{\delta(i)\delta(j)} - T_{\delta(i)} - T_{\delta(j)}\Big) \text{ at diploid SNP}; \\ \sigma_l^2 T_{\delta(i)\delta(j)} \text{ at haploid sat. } l; & \sigma_l^2\Big(4T_{\delta(i)\delta(j)} - T_{\delta(i)} - T_{\delta(j)}\Big) \text{ at diploid sat. } l; \end{cases} \tag{S9}$$

where the symbol $*$ indicates the event that the site segregates in the sample (if the marker is a SNP) and the mutation rate θ_l (if the locus is a microsatellite). The constant of proportionality is $2/T_{\text{tot}}$ for SNPs and $\theta_l/2$ for microsatellites, which we write σ^2 and σ_l^2, respectively. The size of the average sample genealogy T_{tot} and the mutation rate θ_l are not of interest in EEMS.

As a direct consequence of equation (S8), the expected genetic dissimilarity between two distinct individuals depends only on their locations. That is, EEMS assumes that individuals in the same deme are exchangeable.

EEMS decomposes genetic dissimilarities into between-demes and within-demes components

Consider two different demes α and β. Suppose that individual i is assigned to deme α, which we denote by $\delta(i) = \alpha$. Similarly, suppose that $\delta(i^*) = \alpha$ but i and i^* are distinct individuals, that $\delta(j) = \beta$ and $\delta(j^*) = \beta$ but j and j^* are distinct individuals. [Of course, i and j are distinct because they come from different demes.]

EEMS is a spatially explicit model and it is consistent with the following idea: We can expect that (i, j) are more dissimilar than either (i, i^*) or (j, j^*) because i and j come from different locations. However, in general, we can't expect (i, i^*) to be as dissimilar as (j, j^*) as there might be some differences in local genetic diversity. And to model the genetic dissimilarity due to migration in space we should take into account any local variation in genetic diversity.

To capture this idea, EEMS approximates the expected coalescence time between two distinct demes α and β by splitting $T_{\alpha\beta}$ into two components:

$$T_{\alpha\beta} \;=\; \underbrace{T_{\alpha\beta} - \big(T_\alpha + T_\beta\big)/2}_{\text{between demes}} + \underbrace{\big(T_\alpha + T_\beta\big)/2}_{\text{within demes}} \;\approx\; R_{\alpha\beta}/4 + \big(q_\alpha + q_\beta\big)/2, \tag{S10}$$

where $R_{\alpha\beta}$ is the resistance distance between demes α and β in the undirected, connected population grid, and q_α, q_β are the effective diversity rates at α and β, respectively. In this approximation, resistance distances specify the expected genetic differentiation between distinct demes in the habitat (the between-demes component), while the effective diversity rates specify the expected genetic differentiation between distinct individuals from the same deme (the within-demes component).

EEMS uses the approximation given by equation (S10) to specify a model for the expected genetic dissimilarities:

$$\mathrm{E}\{D_{ij} \,|\, *\} \approx \begin{cases} \sigma^2\Big(R_{\delta(i)\delta(j)}/4 + (q_{\delta(i)} + q_{\delta(j)})/2\Big), \text{ hap/SNP}; & \sigma^2\Big(R_{\delta(i)\delta(j)} + (q_{\delta(i)} + q_{\delta(j)})\Big), \text{ dip/SNP}; \\ \sigma_l^2\Big(R_{\delta(i)\delta(j)}/4 + (q_{\delta(i)} + q_{\delta(j)})/2\Big), \text{ hap/sat. } l; & \sigma_l^2\Big(R_{\delta(i)\delta(j)} + (q_{\delta(i)} + q_{\delta(j)})\Big), \text{ dip/sat. } l; \end{cases} \tag{S11a}$$

$$\propto \quad B_{\delta(i)\delta(j)} + (w_{\delta(i)} + w_{\delta(j)})/2 \quad \text{in all four cases.} \tag{S11b}$$

In matrix notation $\mathrm{E}\{D \,|\, *\} \propto \Delta$ and the constant of proportionality is σ^2 for SNPs and σ_l^2 for microsatellite l. Furthermore, the expected dissimilarity matrix Δ has the same form in all four cases, and for the purpose of generality, we can write

$$\Delta = JBJ' + \tfrac{1}{2}Jw1_n' + \tfrac{1}{2}1_n w'J' - W_n, \tag{S12}$$

where $B = (B_{\alpha\beta})$ is the matrix of between-deme dissimilarities, $w = (w_\alpha)$ is the vector of within-demes dissimilarities, $J = (J_{i\alpha})$ is an indicator matrix such that $J_{i\alpha} = 1$ if individual i comes from deme α, and

$W_n = \text{diag}\{Jw\}$. We subtract the diagonal matrix $W_n = \text{diag}\{\frac{1}{2}Jw1_n' + \frac{1}{2}1_n w'J'\}$ because the expected dissimilarity matrix Δ has a main diagonal of 0s. If there are n individuals sampled from o demes, then $J \in \mathbb{Z}^{n\times o}$, $B \in \mathbb{R}^{o\times o}$, $w \in \mathbb{R}^o$. To simplify the notation, we drop the subscripts and write plainly 1 for the vector of 1s. The dimension will be clear from the context because B is an $o \times o$ matrix and Δ is an $n \times n$ matrix.

EEMS models dissimilarities between demes as a function of migration and dissimilarities within demes as a function of diversity

The within-demes component w characterizes the expected genetic dissimilarity between two distinct individuals from the same deme and is a function of the effective diversity rates:

$$w_\alpha = g(q_\alpha). \tag{S13}$$

Here q is a vector of effective diversity rates and q_α is the element that corresponds to deme α. The function g is the identity.

The between-demes component B characterizes the expected genetic dissimilarity between two individuals from distinct demes, *after* correcting for the local differences in genetic diversity, and is a function of the effective migration rates:

$$B_{\alpha\beta} = f(m)_{\alpha\beta}. \tag{S14}$$

Here m is a sparse matrix that represents an undirected, connected, weighted grid, with weights equal to the effective migration rates between adjacent demes. The function f returns the effective resistance distances between vertices in the grid, as a dense matrix, and $f(m)_{\alpha\beta}$ is the element which corresponds to the pair of demes α and β.

In EEMS model the parameters of the model are of greater interest than the expected dissimilarities, i.e., m and q are more interesting than the deterministic functions $f(m)$ and $g(q)$.

S2.3 Computing the Wishart likelihood in EEMS

EEMS represents the population as a connected undirected graph (V, E), with effective migration rates $m = \{(\alpha, \beta) \in E : m_{\alpha\beta}\}$ and effective diversity rates $q = \{\alpha \in V : q_\alpha\}$. Furthermore, EEMS models the observed genetic differences D between n individuals, averaged across p SNPs, through a positive definite transformation:

$$-LDL' \,|\, k, m, q, \sigma^2 \sim \text{W}_{n-1}\Big(k, -\frac{\sigma^2}{k} L\Delta(m, q)L'\Big), \tag{S15}$$

where $\Delta(m, q)$ is the matrix of expected genetic dissimilarities as function of the between-demes component $B(m)$ and the within-demes component $w(q)$. [In Section S2.2, Δ is given by equation (S12), w by Equation (S13) and B by Equation (S14).] In addition, L is a $(n-1) \times n$ basis for contrasts on n elements, k is the degrees of freedom, constrained to lie in the range $[n, p]$, and σ^2 is a scale parameter. See Section S2.2 for a demographic interpretation of σ^2.

By definition, a contrast is a linear combination with coefficients that add to zero, so $L1 = 0$ and

$$L\Delta L' = L\Big(JBJ' + \tfrac{1}{2}Jw1' + \tfrac{1}{2}1w'J' - W_n\Big)L' = L\Big(JBJ' - W_n\Big)L'. \tag{S16}$$

Equation (S16) implies that Δ and $JBJ' - W_n$ are equivalent under the Wishart likelihood (S15) because they give the same likelihood. Therefore, without loss of generality, we can assume that the expected dissimilarity matrix has the form:

$$\Delta = JBJ' - W_n, \tag{S17}$$

where JBJ' is a block matrix and W_n is a diagonal matrix. We can exploit this structure to compute the Wishart log likelihood efficiently, without explicitly constructing the $n \times n$ matrix Δ. As a result, the computational cost scales with the grid size, not with the number of samples. The hard-to-compute terms of the Wishart likelihood (S15) are the determinant and the trace:

$$\operatorname{tr}\left\{(L\Delta L')^{-1}LDL'\right\} = \operatorname{tr}\left\{\Delta^{-1}\Big(\Delta L'(L\Delta L')^{-1}L\Big)D\right\} = \operatorname{tr}\left\{\Delta^{-1}QD\right\}, \tag{S18a}$$

$$\det\left\{-(L\Delta L')^{-1}\right\} = \det\left\{-L'(L\Delta L')^{-1}L\right\} / \det\left\{LL'\right\} = \operatorname{Det}\left\{-\Delta^{-1}Q\right\} / \det\left\{LL'\right\}, \tag{S18b}$$

where det denotes the standard determinant (the product of all eigenvalues), Det denotes the pseudo determinant (the product of the nonzero eigenvalues) and

$$Q = \Delta L'(L\Delta L')^{-1}L = I - 1(1'\Delta^{-1}1)^{-1}1'\Delta^{-1} \tag{S19}$$

is an orthogonal projection matrix with kernel $\{1\}$, the space of constant functions.

The distance matrix $\Delta = JBJ' - W_n$ is the sum of a block matrix and a diagonal matrix, and its inverse Δ^{-1} has similar "almost-block" structure:

$$\Delta^{-1} = JXJ' - W_n^{-1}, \tag{S20}$$

where X is an unknown $o \times o$ matrix. Since $\Delta\Delta^{-1} = I$, the solution X must satisfy:

$$JBCXJ' - W_nJXJ' - JBJ'W_n^{-1} + W_nW_n^{-1} = I \qquad \Leftrightarrow \qquad J(BC - W_o)XJ' = JBW_o^{-1}J', \tag{S21}$$

where $C = J'J = \operatorname{diag}\{n_\alpha\}$ is the diagonal matrix of deme sizes and n_α is the number of geo-referenced individuals assigned to deme α in the population graph. [Demes with $n_\alpha > 0$ are the observed demes.] Since every term in equation (S21) has exact block structure which depends on the sample configuration through J, it is sufficient to solve the lower-dimensional problem:

$$\Big(BC - W_o\Big)X = BW_o^{-1}. \tag{S22}$$

This is a system of linear equations for the unknown matrix X as a function of the between-demes dissimilarities B, the within-demes dissimilarities $W_o = \operatorname{diag}\{w\}$ and the sample counts C. Equation (S22) can be solved efficiently without a matrix inversion, by performing the LU factorization of $Y = BC - W_o$.

We can express the pseudo-determinant $\operatorname{Det}\{-\Delta^{-1}Q\}$ and the trace $\operatorname{tr}\{\Delta^{-1}QD\}$ in terms of the auxiliary matrix X. Using the definition of the orthogonal projection Q in equation (S19) and the properties of the trace,

$$\operatorname{tr}\left\{\Delta^{-1}QD\right\} = \operatorname{tr}\left\{\Delta^{-1}D\right\} - \frac{1}{1'\Delta^{-1}1}\operatorname{tr}\left\{11'\Delta^{-1}D\Delta^{-1}\right\}. \tag{S23}$$

For simplicity of notation, let $c = \mathrm{diag}\{C\}$ and $v = (w_\alpha^{-1})$. We consider each term in Equation (S23):

$$1'\Delta^{-1}1 = 1'\Big(JXJ' - W_n^{-1}\Big)1 = c'\Big(Xc - v\Big), \tag{S24a}$$

$$\mathrm{tr}\Big\{\Delta^{-1}D\Big\} = \mathrm{tr}\Big\{\Big(JXJ' - W_n^{-1}\Big)D\Big\} = \mathrm{tr}\Big\{XJ'DJ\Big\}, \tag{S24b}$$

$$\mathrm{tr}\Big\{11'\Delta^{-1}D\Delta^{-1}\Big\} = \Big(c'X'J' - v'J'\Big)D\Big(JXc - Jv\Big) = \Big(Xc - v\Big)'J'DJ\Big(Xc - v\Big). \tag{S24c}$$

The matrix product in red, $J'DJ$, is a known matrix of order o, where o is the number of observed demes, and it can be precomputed and stored for easy access. Thus we do not need to construct the $n \times n$ matrix Δ^{-1} in order to compute $\mathrm{tr}\{\Delta^{-1}QD\}$; we can work with the $o \times o$ matrix X instead. [It follows that the computational cost scales with the grid size, not with the sample size.]

Next we show how to compute the pseudo determinant $\mathrm{Det}\{-\Delta^{-1}Q\}$. Following (Verbyla, 1990), we can show that

$$\mathrm{Det}\Big\{-\Delta^{-1}Q\Big\} = \frac{\det\{LL'\}}{\det\{-L\Delta L'\}} = \frac{(1'1)/(1'\Delta^{-1}1)}{-\det\{-\Delta\}}. \tag{S25}$$

A distance matrix is conditionally negative definite, and so Δ has one positive eigenvalue and $n-1$ negative eigenvalues (Bapat and Raghavan, 1997). This guarantees that $-\det\{-\Delta\}$ is positive and so it is sufficient to compute $|\det\{\Delta\}| = |JBJ' - W_n|$, which can be obtained from the LU decomposition of $Y = BC - W_o$:

$$\Big|\det\{\Delta\}\Big| = \Big|\det\{W_n\}(-1)^{n-o}\det\{W_o^{-1}BC - I\}\Big| = \Big|\det\{W_n\}\det\{W_o^{-1}\}\det\{BC - W_o\}\Big|. \tag{S26}$$

We can use Equations (S24), (S25) and (S26) to evaluate the the Wishart likelihood (S15) for the parameters k, m, q and σ^2.

S2.4 Computing resistance distances in an undirected graph

In Section S2.3 the between-demes component $B(m)$ is the matrix of pairwise distances between demes. Various distance metrics can be considered but, following (McRae, 2006), EEMS uses the metric "effective resistance" or "resistance distance" $R(m)$.

Let $\mathcal{L}$ be the graph Laplacian of the population graph (V, E) with effective migration rates m. (Resistance distances do not depend on the diversity rates q, so those parameters are not relevant for the following computation.) The graph Laplacian is given by:

$$\mathcal{L} = \mathcal{D} - M, \tag{S27}$$

where $M = (m_{\alpha\beta})$ is the (sparse) matrix of effective migration rates between connected demes and $\mathcal{D} = (\mathcal{D}_{\alpha\alpha})$ is the diagonal matrix with $\mathcal{D}_{\alpha\alpha} = \sum_{\beta:\beta\neq\alpha} m_{\alpha\beta}$.

Following (Babić et al., 2002), we can use $\mathcal{L}$ to compute the effective resistances $R = (R_{\alpha\beta})$ for all pairs of demes in the population graph, by inverting the sum matrix $\Gamma = \mathcal{L} + 11'/c$ where $c > 0$ is a constant. Let $H = \Gamma^{-1}$ and $h = \mathrm{diag}\{H\}$. Then

$$R = 1h' + h1' - 2H. \tag{S28}$$

Importantly, equation (S16) suggests that we can take $B = -2H$ instead of $B = R$ because the matrices $-2H$ and R produce the same likelihood for the EEMS parameters and are therefore equivalent. In Section S2.3 we

use a similar argument to show that $\Delta = JBJ' + \frac{1}{2}Jw1' + \frac{1}{2}1w'J' - W_n$ and $JBJ' - W_n$ are equivalent under the Wishart likelihood (S15). In other words, the likelihood is invariant to adding components with the form $1v'$ or $v1'$ for a vector v.

Furthermore, we can avoid inverting the matrix Γ to obtain the auxiliary matrix X in equation (S20). [Γ is an $d \times d$ matrix where d is the number of demes in the graph; X is an $o \times o$ matrix where o is the number of demes assigned at least one individual.] Let $\Gamma_{o\times o}$ be the $o \times o$ block that corresponds to the observed demes; similarly, let $\Gamma_{(d-o)\times(d-o)}$ be the $(d-o) \times (d-o)$ block that corresponds to the unobserved demes. Then

$$H^{-1}_{o\times o} = \Gamma_{o\times o} - \Gamma_{o\times(d-o)}\Gamma^{-1}_{(d-o)\times(d-o)}\Gamma_{(d-o)\times o}, \tag{S29}$$

which can be computed efficiently by solving a linear system. Finally, the dissimilarities $B_{o\times o}$ between observed demes can be computed from $B^{-1}_{o\times o} = -H^{-1}_{o\times o}/2$. If the population graph is sparsely sampled, as is often the case, it is more efficient to compute the Schur complement of $\Gamma_{o\times o}$ in equation (S29), rather than invert the full-size matrix Γ. This idea is also used in (Hanks and Hooten, 2013).

Computational complexity

The auxiliary matrix Γ is dense, diagonally dominant, positive definite and of order d, where d is the number of observed demes. First we compute the Schur complement $H^{-11}_{o\times o}$ according to equation (S29), and then the between-demes dissimilarities $B_{o\times o}$.

1. Cholesky decomposition $\Gamma_{(d-o)\times(d-o)} = U'U$: $O((d-o)^3)$
2. Forward substitution $U'Y = \Gamma_{(d-o)\times o}$: $O(o(d-o)^2)$
3. Backward substitution $UX = Y$: $O(o(d-o)^2)$
4. Matrix inversion $B_{o\times o} = -H^{-1}_{o\times o}/2$: $O(o^3)$

This procedure has complexity $O((d+o)(d-o)^2 + o^3)$ and, except for very small graphs, it is more efficient than inverting the sum matrix Γ which has complexity $O(d^3)$.

S2.5 Birth-death Markov Chain Monte Carlo estimation

In Section S2.3 we assume, without loss of generality, that the expected dissimilarity matrix has the form:

$$\Delta(m, q) = JB(m)J' - W(q), \tag{S30}$$

where $B(m)$ is the between-demes component, which is a function of the migration rates m, and $W(q)$ is the within-demes component, which is a function the diversity rates q.

EEMS uses two Voronoi tessellations, which independently partition the habitat: one parametrizes the effective migration rates m, and the other – the effective diversity rates q. Specifically, the migration rates m are determined by a Voronoi tessellation with C_m cells, seeds $s_1, \ldots, s_{C_m}$, migration effects $e_1, \ldots, e_{C_m}$, and overall migration rate (on the $\log_{10}$ scale) μ, while the diversity rates are determined by another independent Voronoi tessellation with C_q cells, seeds $t_1, \ldots, t_{C_q}$, and diversity effects $f_1, \ldots, f_{C_q}$. The overall diversity rate

is assumed to be 0 on the $\log_{10}$ scale (1 on the original scale). We fix the overall diversity rate because the two components of the expected dissimilarity matrix scale so that $B(m/2) = 2B(m)$, $W(2q) = 2W(q)$. With the current parametrization, fixing the overall diversity rate to 1 makes the scale σ^2 identifiable. Finally, the migration cell effects $e_1, \ldots, e_{C_m}$, have variance ω_m^2, while the diversity cell effects $f_1, \ldots, f_{C_q}$ have variance ω_q^2.

We use birth-death Markov Chain Monte Carlo (MCMC) to estimate the number of cells C in each Voronoi tessellation because the dimension of the seeds and the cell effects changes as C increases or decreases. [The same procedure is used to update the migration and the diversity Voronoi tessellations, so instead of C_m or C_q we write C.] In each step, we propose the birth (addition) or death (removal) of a cell, with equal probability (Stephens, 2000). For a birth proposal, the acceptance probability is

$$\alpha(\Theta, \Theta^*) = \min\left\{1, u\frac{c+r}{c+1}\frac{\ell(\Theta^*)}{\ell(\Theta)}\right\}, \tag{S31}$$

where c is the current number of cells; (r, u) are the parameters of the negative binomial prior on C, Θ is the current parameter state (with c cells) and Θ^* is the proposed parameter state, with one additional cell added at a random location within the habitat and assigned a random effect drawn from a (truncated) normal prior. Small probability of success u means small acceptance probability α unless the likelihood ratio indicates strong evidence in favor of adding the new cell.

For a death proposal, one cell is randomly chosen to be removed. There should be at least one cell in the Voronoi tessellation at each step, so let $c + 1$ be the current number of cells, i.e., there are at least two cells currently. The acceptance probability for a death proposal has the form:

$$\alpha(\Theta, \Theta^*) = \min\left\{1, \frac{1}{u}\frac{c+1}{c+r}\frac{\ell(\Theta^*)}{\ell(\Theta)}\right\}. \tag{S32}$$

For SNP data, the parameter state $\Theta = (k, m, q, \sigma^2)$ consists of the degrees of freedom k, the migration rates m, the diversity rates q and a scale parameter σ^2; the likelihood $\ell(\Theta)$ is given by equation (3). For microsatellite data, the parameter state $\Theta = (m, q, \sigma_1^2, \ldots, \sigma_p^2)$ consists of the migration rates m, the diversity rates q and locus-specific scale parameters $\sigma_1^2, \ldots, \sigma_p^2$; the likelihood $\ell(\Theta)$ is given by equation (7).

For a given number of Voronoi cells C_m and C_q in the two Voronoi tessellations, the cell effects and their locations, the overall migration rate μ and (for SNPs data only) the effective degrees of freedom k are each updated in turn with a random-walk Metropolis-Hastings step:

$$\alpha(\Theta, \Theta^*) = \min\left\{1, \frac{p(\Theta^*)}{p(\Theta)}\frac{\ell(\Theta^*)}{\ell(\Theta)}\right\}, \tag{S33}$$

where $p(\Theta)$ is the prior and $\ell(\Theta)$ is the likelihood.

Finally, the scalar variance parameters are $\omega_m^2, \omega_q^2, \sigma^2$ for SNP data, and $\omega_m^2, \omega_q^2, \sigma_1^2, \ldots, \sigma_p^2$ for microsatellite data. These parameters are updated with a Gibbs step by sampling from the corresponding full conditional distribution, which is inverse gamma. For example, the variance in relative migration among cells, ω_m^2, is drawn from

$$\omega_m^2 \,|\, C, e_1 \ldots, e_C \sim \text{Inv-G}\Big((c_\omega + C)/2, (d_\omega + SS_e)/2\Big), \tag{S34}$$

where the sum of squares is $SS_e = \sum_{c=1}^{C} e_c$.

S3 Visualizing estimated migration rates as a color contour plot

To simplify comparisons, we plot relative migration rates on the same scale throughout the paper; the blue-and-orange color palette is based a collection of divergent color schemes suitable for people with deficient red-green vision. As with any contour plot, the choice of color and scaling can affect – sometimes profoundly – the resulting image and the message it conveys. At the simplest level, a scheme that is too broad will wash out any differences in effective migration rate between regions, while a scheme that is too narrow may over-emphasize trivial differences. The color scheme used here, as well as some alternatives, is described in (Light and Bartlein, 2004) and available at `http://geog.uoregon.edu/datagraphics/color_scales.htm`.

An estimated effective migration surface is a color contour plot. The accompanying R package `rEEMSplots` (`http://www.github.com/dipetkov/eems`) generates such a plot according to the following procedure:

1. Choose a grid of interpolation points (x, y). These do not correspond to the demes in the population graph and a dense interpolation grid is required to generate a smooth migration surface.

2. For each Voronoi tessellation drawn from the posterior distribution on the effective migration rates, compute the rate $\log_{10}(m_{xy})$ at every interpolation point. A Voronoi tessellation partitions the habitat and each interpolation point is assigned the migration rate $\log_{10}(m_c)$ of the cell c that it falls into.

3. Standardize the migration rates so that the mean over the interpolation points is 0.

The Voronoi cells do not necessarily have the same area and so may contain a different number of interpolation points. Therefore, the unweighted average over the interpolation points corresponds to a weighted average across the Voronoi cells, with weights proportional to the area of each cell. With the normalization described above, the "average color" across the color contour plot is white.

S4 Details about four empirical datasets analyzed with EEMS

EEMS is a general method for visualizing spatial population structure and it might be appropriate to apply quality control steps that are customary in population structure analyses, such as pruning SNPs because of long-range LD or high missingness. Measures for SNP and sample quality control have been applied to each of the empirical datasets analyzed here.

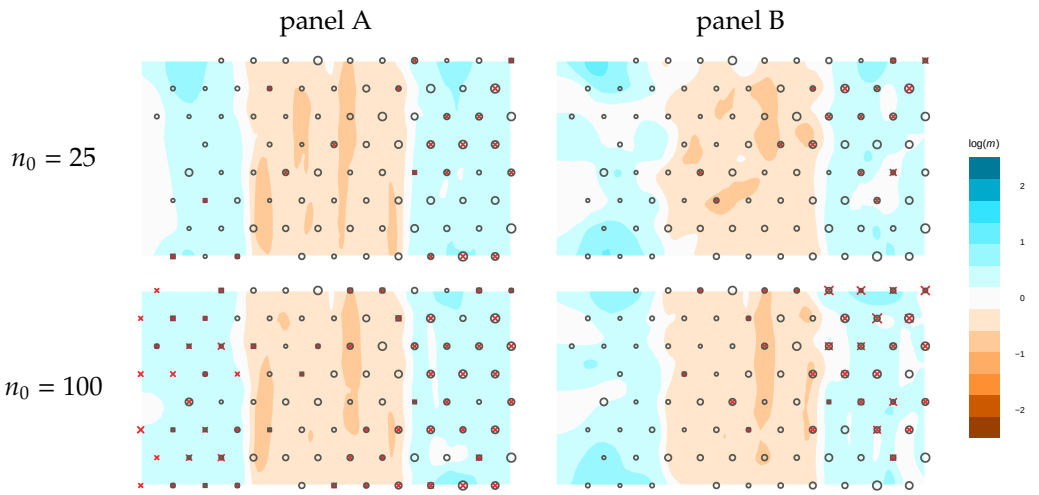
Elephant data. The African elephant dataset is collected and genotyped as part of a large collaborative study to develop assignment methods for determining the geographic origin of elephant samples from across Sub-Saharan Africa (Wasser et al., 2004). Samples from both forest and savanna elephants have been collected and genotyped at 16 microsatellite loci. Although the two subspecies can be accurately discriminated using the 16 microsatellites, there is observational and genetic evidence that forest and savanna elephants can hybridize (Wasser et al., 2004). We analyzed the geo-referenced data from (Wasser et al., 2015), which excludes putative hybrids (samples with posterior probability of being hybrid greater than 0.01) and consists of 211 forest and 913 savanna elephants, from 75 distinct locations in 28 countries in Sub-Saharan Africa. The microsatellite data and sample locations are available on the Dryad Digital Repository (`http://dx.doi.org/10.5061/dryad.435p4`).

Human European and African data. The European dataset was collected and genotyped as part of the POPRES (Population Reference Sample) project (Nelson et al., 2008) and can be accessed at https://www.ebi.ac.uk/ega/studies/phs000145.v2.p2. We used a focal subset of 197,146 autosomal SNPs and 1,379 individuals analyzed in a previous publication (Novembre et al., 2008), with the individual IDs and marker list available from https://github.com/NovembreLab/Novembre_etal_2008_misc. We analyzed a subset of 1,201 individuals from 13 Western European countries: Austria (AT), Belgium (BE), Denmark (DK), France (FR), Germany (DE), Ireland (IE), Italy (IT), Netherlands (NL), Portugal (PT), Scotland (Sct), Spain (ES), Switzerland (CH), United Kingdom (UK). The samples from Switzerland (CH) are split into three subpopulations: French, Italian and German speaking Swiss, coded as CHf, CHi and CHg, respectively. We removed five samples from Italy (7623, 33242, 34049, 38532, 49500) that project outside the main Italian cluster in PC1-PC2 space and therefore are identified as possible outliers in (Novembre et al., 2008). [For example, these samples might have insular Italian ancestry – Sardinian or Sicilian.] The resulting dataset is described in Section S7.

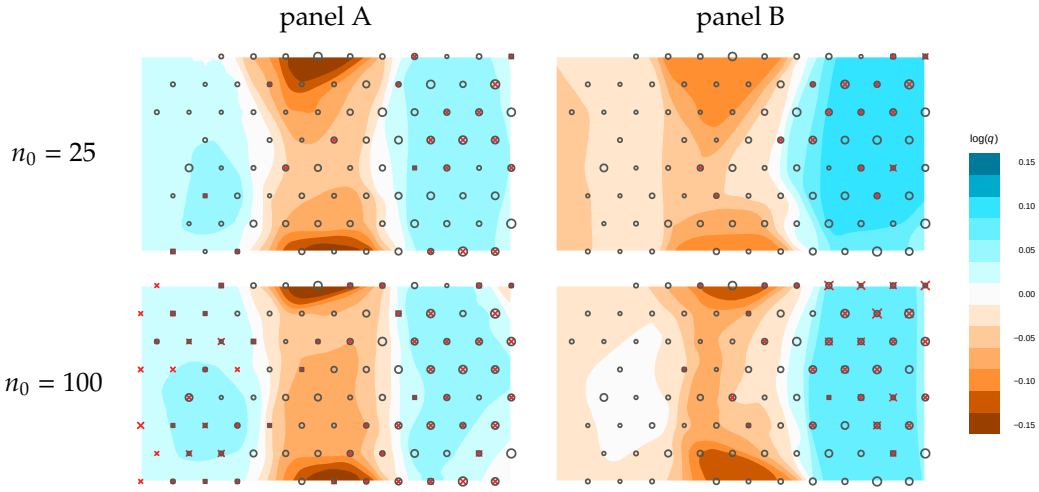
The African dataset was compiled from a subset of two published SNP array datasets: one presented in (Xing et al., 2010) and available from http://jorde-lab.genetics.utah.edu/?page_id=23 and the other presented in (Henn et al., 2011) and available from http://www-evo.stanford.edu/repository/. From the Xing *et al.* dataset we extracted the populations: Alur (Al), Bambaran (Ba1), Dogon (Do), Hema (He), Nguni (Ng), Pedi (Pe) and Sotho/Tswana (ST); from the Henn *et al.* dataset we extracted all samples from the populations: Bamoun (Ba2), Brong (Br), Bulala (Bu), Fang (Fa), Hausa (Ha), Igbo (Ig), Kaba (Ka), Kongo (Ko), Mada (Ma2), Mandenka (Ma3) and Xhosa (Xh) and as well as the Yoruba (Yo) samples in the Human Genetic Diversity Project (HGDP) and the Luhya (Lu) and Maasai (Ma1) samples in the HAP1117 subset of HapMap phase 3 (Pemberton et al., 2010). The two subsets were merged at SNPs that have been genotyped in both datasets. From the merged dataset, we then removed SNPs with more than 5% missingness per marker and samples with more than 5% missingness per individual, as well as two Hema individuals that are classified as likely relatives and outliers in most analyses of Sub-Saharan samples in (Wang et al., 2012). After these exclusions, we analyzed a dataset composed of 314 samples from 21 Sub-Saharan populations genotyped at 27,825 polymorphic SNPs. The resulting dataset in described in Section S7.

***Arabidopsis thaliana* data**. The *Arabidopsis thaliana* dataset was collected and genotyped as part of the RegMap (Regional Mapping) project (Horton et al., 2012) and is available at http://bergelson.uchicago.edu/regmap-data/. We downloaded unimputed SNP genotypes for 1,193 samples with high-quality geographic coordinates (latitude and longitude), categorized into twelve geographic regions. From these we analyzed 1,160 accessions from North America and Europe, genotyped at 214,051 SNPs using the Affymetrix Arabidopsis 250K SNP chip (Horton et al., 2012). These include 180 accessions from the region *Americas* and 979 accessions from the (European) regions *British-Isles*, *Fennoscandia*, *France*, *Iberia*, *North-West Europe*, *South-Central* and *Austria-Hungary*. We excluded three accessions (Yo-0, Van-0, Buckhorn Pass) from the western coast of North America because the rest are collected from the eastern and central United States, as well as one accession (Can-0) because it is collected from Spain's Canary Islands and one accession (Da(1)-12) from the Czech Republic because its exact latitude/longitude coordinates are missing.

S5 Additional simulations under the stepping stone model

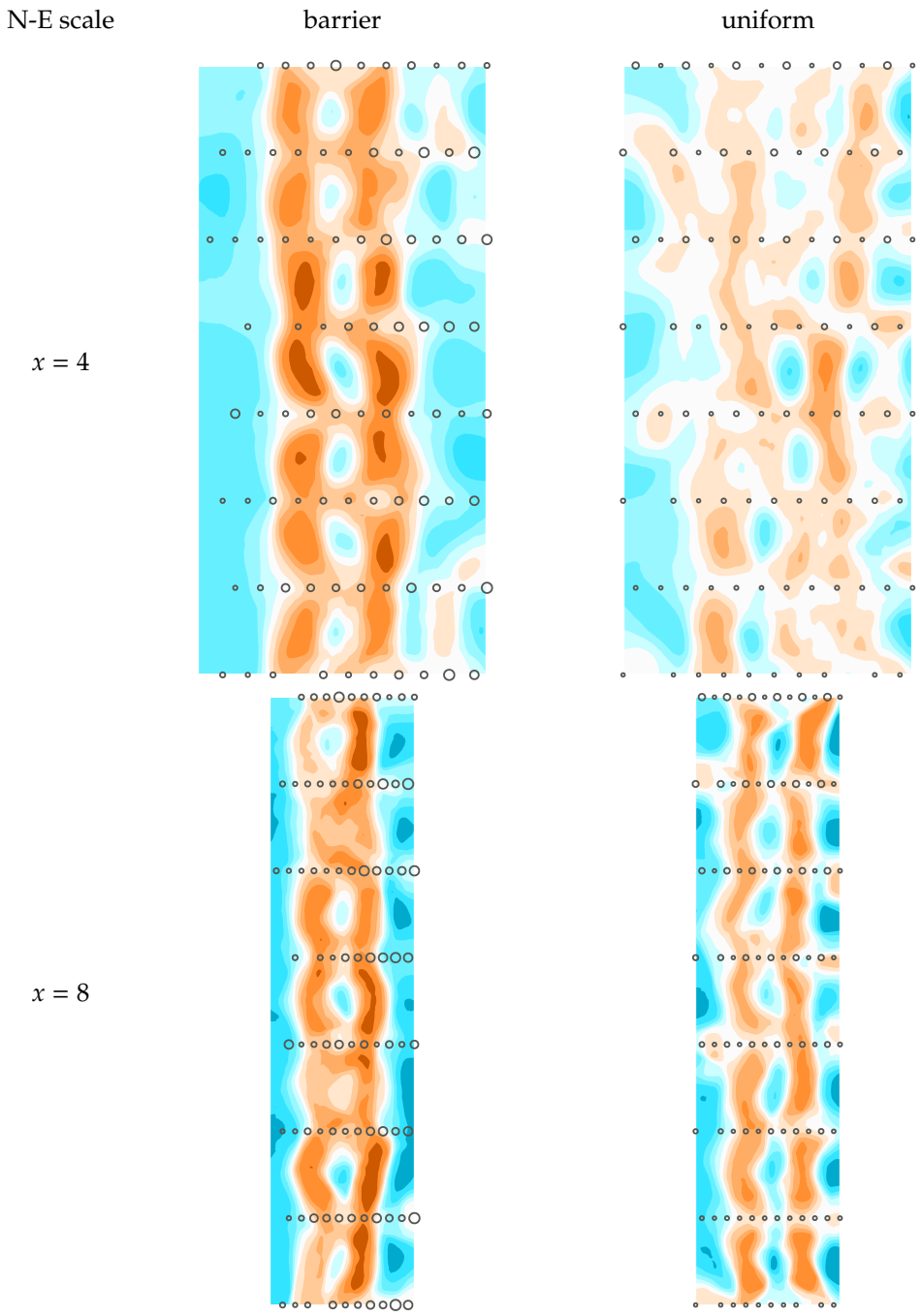


(a)

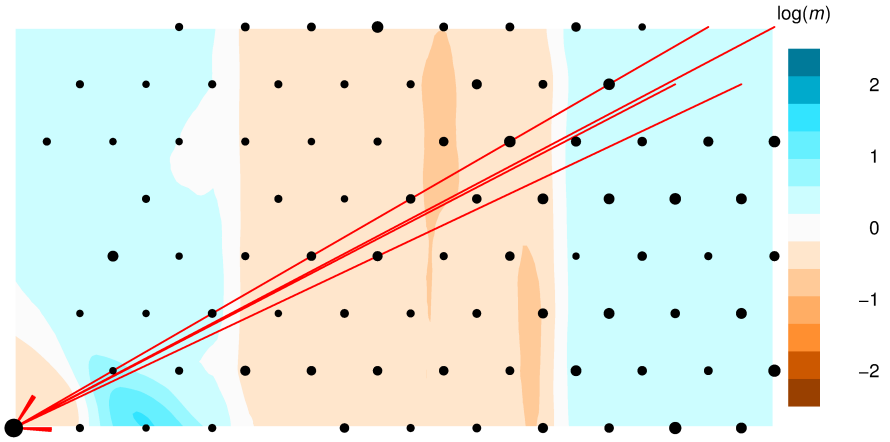


(b)

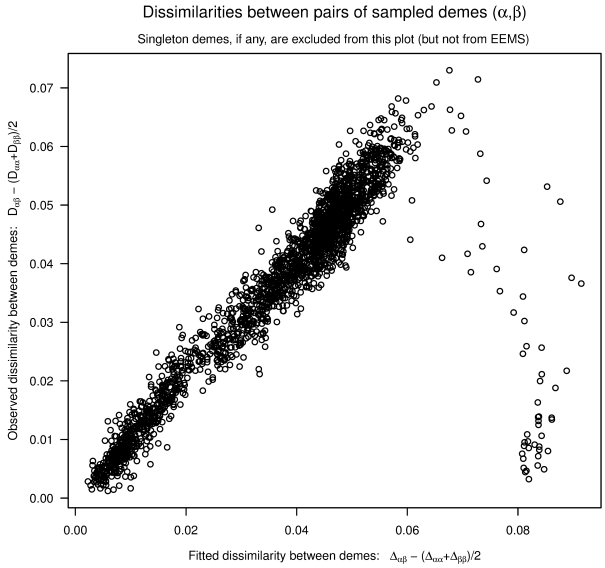
Supplementary Figure 1 Geographic bias in SNP ascertainment can affect EEMS inference about effective migration and diversity. As (McVean, 2009) points out, two samples are involved in SNP ascertainment: first a panel to discover SNPs for genotyping on a microchip and then a sample to genotype and analyze. The ascertainment is biased if the discovery panel is not representative of the genetic variation in the population. Here we illustrate SNP ascertainment with geographic bias under a barrier to migration scenario. We simulated $n = n_0 + n_1$ individuals, of which n_0 are designated discovery panel (red crosses) and $n_1 = 300$ are designated geo-referenced sample (black circles). The SNP ascertainment is geographically biased because the discovery panel is preferentially sampled from the right than from the left of the barrier. **(a)** Estimated effective migration rates on the same $\log_{10}$ scale indicated by the color bar. **(b)** Estimated effective diversity rates, also on a common $\log_{10}$ scale. In these simulations, panel B has a stronger geographic bias than panel A and the ascertainment bias affects the diversity estimates more than the migration estimates. In panel B, the effective diversity is highest in the region where most of the discovery panel is sampled from, on the right of the barrier. Intuitively, after ascertainment, it is more likely to "discover" mutations that have arisen in the discovery region, which has increased effective diversity as a result.



Supplementary Figure 2 Migration directionality can affect EEMS inference about effective migration. EEMS assumes that migration is undirected and that migration rates are locally similar; both assumptions can be violated in practice. For a straightforward simulation of directed migration, we used two datasets analyzed in Figure 2 and we multiplied the latitude (vertical) coordinate of every sampling location by a fixed factor x. Thus N-S migration is x times as fast as E-W migration. (Equivalently, under the coalescent, a lineage can travel x times as far in the N-S direction than in the E-W direction, in the same amount of time.) We vary both the N-S scale factor: $x = 4$ in the top row, $x = 8$ in the bottom row; and the underlying true migration scenario: a barrier to migration in the left column, uniform migration in the right column. EEMS is not well suited to modeling migration directionality; nevertheless, it attempts to explain the spatial patterns in genetic dissimilarities by inferring vertical N-S barriers which effectively "slow down" migration in the E-W direction.

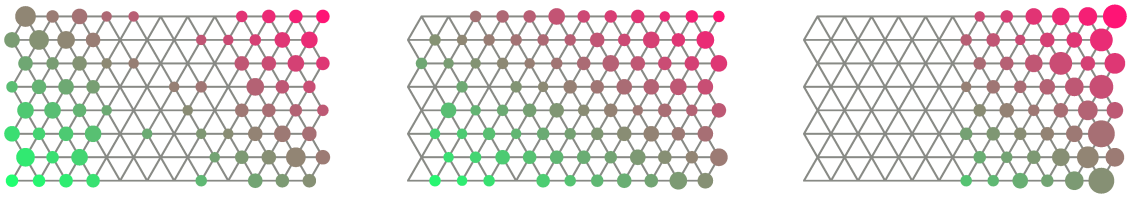


(a)

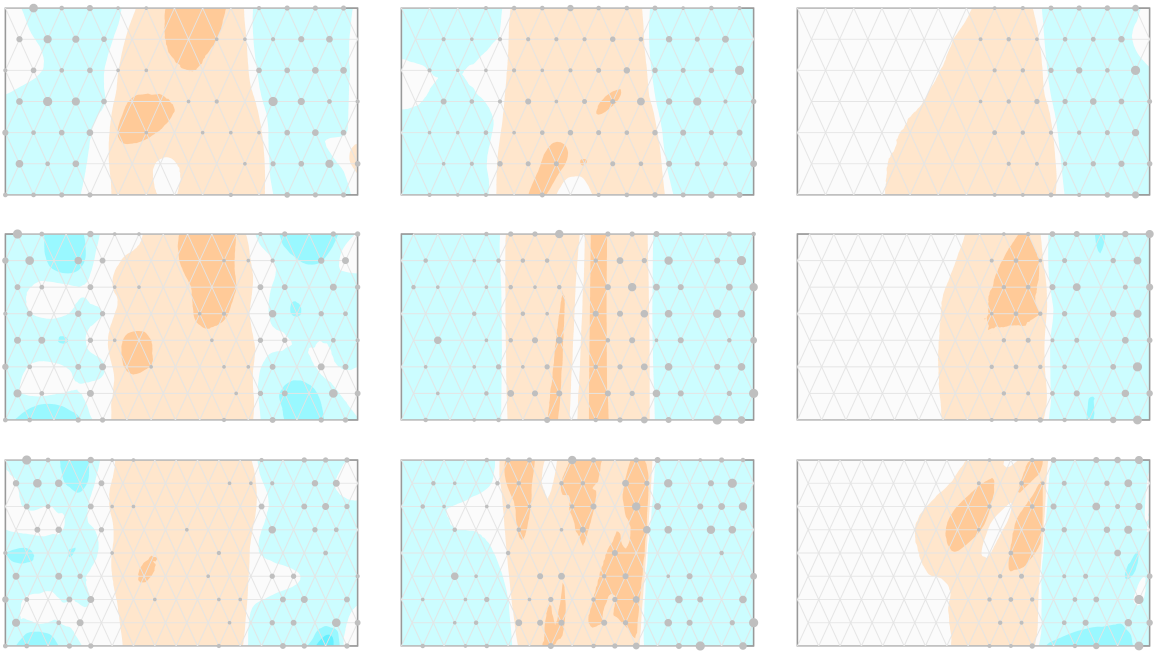


(b)

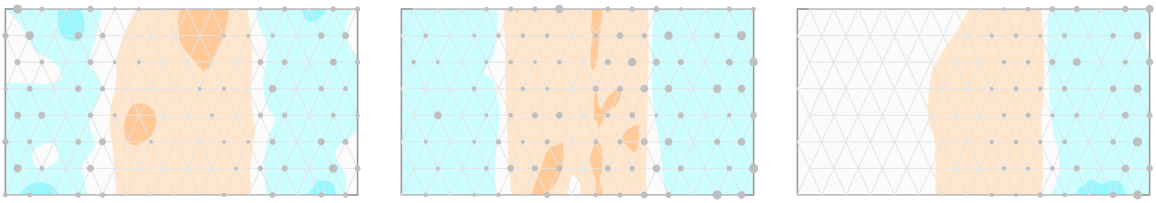
Supplementary Figure 3 Lack of fit due to recent migrants or incorrect geographic labels. In **(a)** 21 individuals from the top right corner are incorrectly assigned to the bottom left corner, as indicated by the the red arrows. The small region of low effective migration in the bottom left creates a barrier around these "migrants", capturing the fact that they are genetically distinct from other nearby individuals. Effectively, the "recipient" deme is isolated from its neighbors. At the same time, the "migrants" are genetically similar to the distant individuals in the top right corner but this spatial pattern is not represented in the estimated effective migration surface. **(b)** Indeed EEMS considerably overestimates the corresponding expected dissimilarities; the lack of fit is highlighted by the outliers in the diagnostic between-demes scatter plot. The diagnostic scatter plot is generated automatically by the EEMS software, by plotting the pairwise genetic differences predicted by the fitted model against the pairwise genetic differences observed in the data.



(a) Sampling configuration on the true 12×8 population grid.



(b) Estimated effective migration rates on 13×7, 15×8 and 17×9 grids.



(c) Estimated effective migration rates, averaged over the 13×7, 15×8 and 17×9 grids.

Supplementary Figure 4 The vertices V and the edges E in the population grid are specified as inputs to the EEMS program and are fixed during estimation; the choice for V and E can affect the inferred migration rates. For example, a very coarse grid might not allow for sufficient variation in resistance distances. Therefore, we recommend averaging the surfaces estimated with grids of different sizes. In practice this will moderate the discretization implicit in assigning the samples collected in a continuous habitat to the vertices of a discrete population graph. Here we use the barrier-to-migration simulation in Figure 2 for illustration. **(a)** The data is simulated on a 12×8 grid. **(b)** We use 13×7, 15×8 and 17×9 grids to estimate effective migration rates. As the grid dimensions change, so do the occupied demes because samples are assigned to the closest deme in the grid; small details in the estimated surfaces change as well. **(c)** After averaging the estimated migration rates over grids of two different sizes, the broad features remain the same.

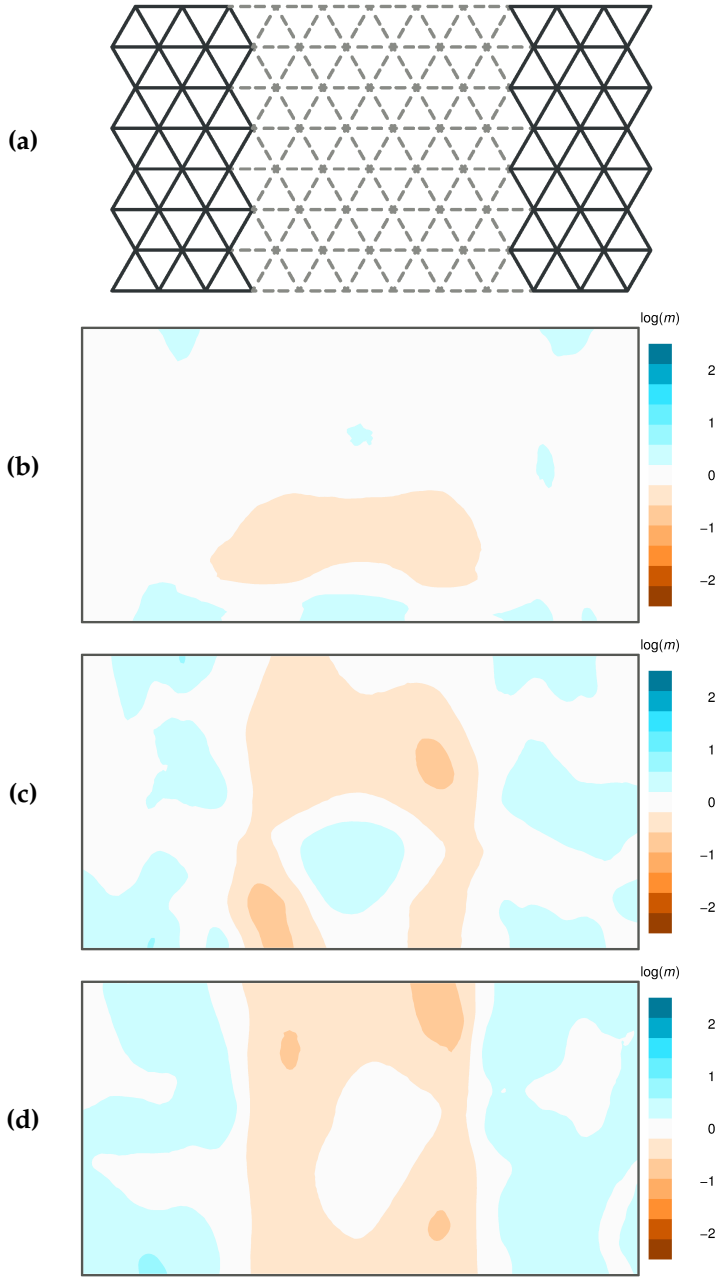
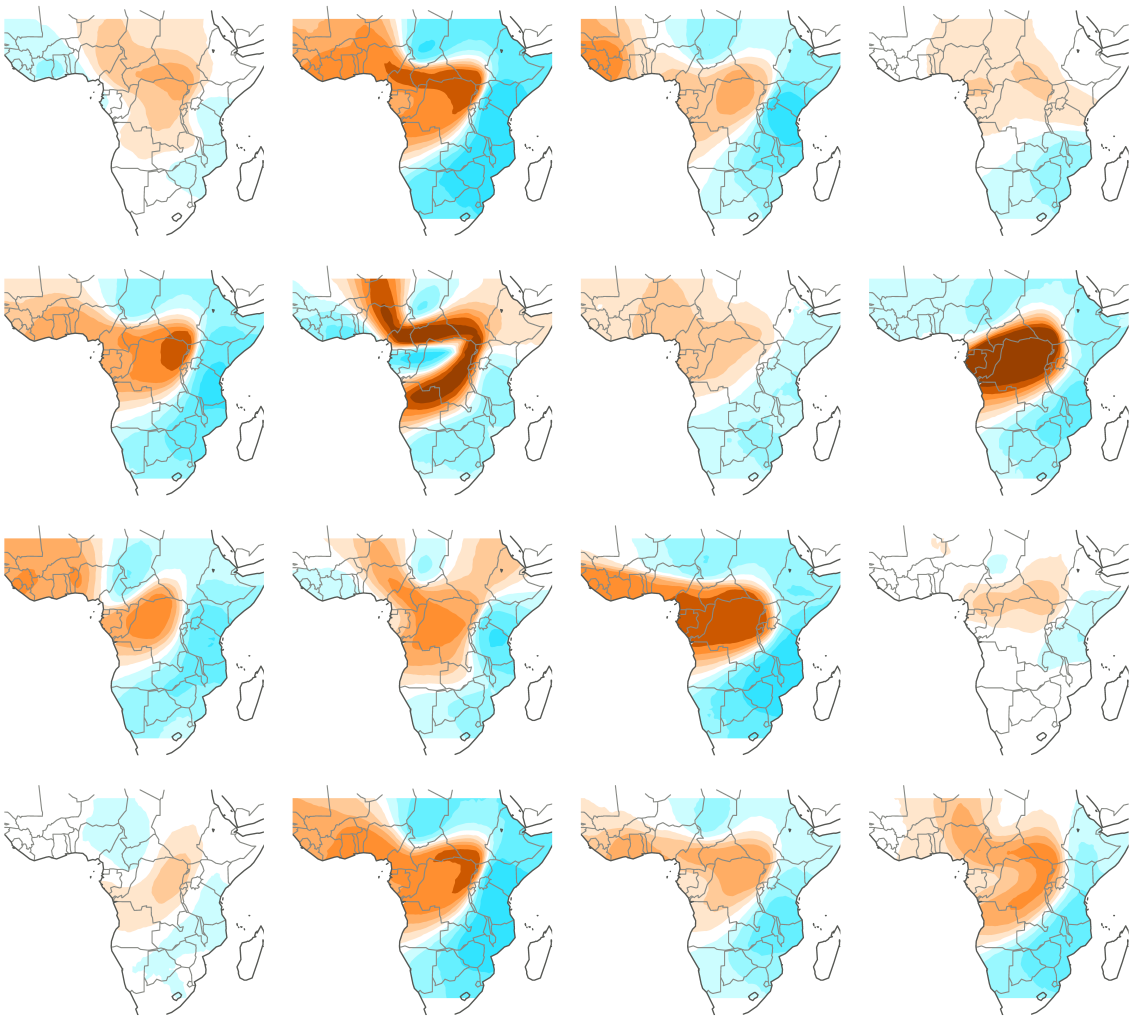
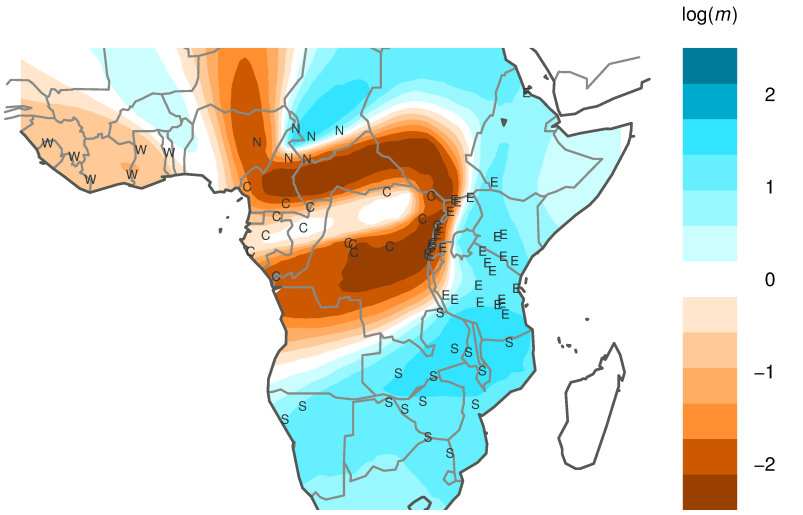


Figure A past demographic event can produce a barrier to effective migration. **(a)** An ancestral population splits into two subpopulations, E in the east and W in the west, after the migration rates in the central region drop simultaneously to 0. This event, which occurs x units of time in the past, creates an "effective barrier" to migration. The further back in time the split event occurs, the more differentiated subpopulations E and W are. The split occurs at **(b)** $x = 1$; **(c)** $x = 4$; **(d)** $x = 9$ units of time in the past; x is measured in N_0 generations. Before the split, migration rates are all set to 1 (on the same coalescent scale N_0); after the split, migration rates on either side of the central region remain 1.

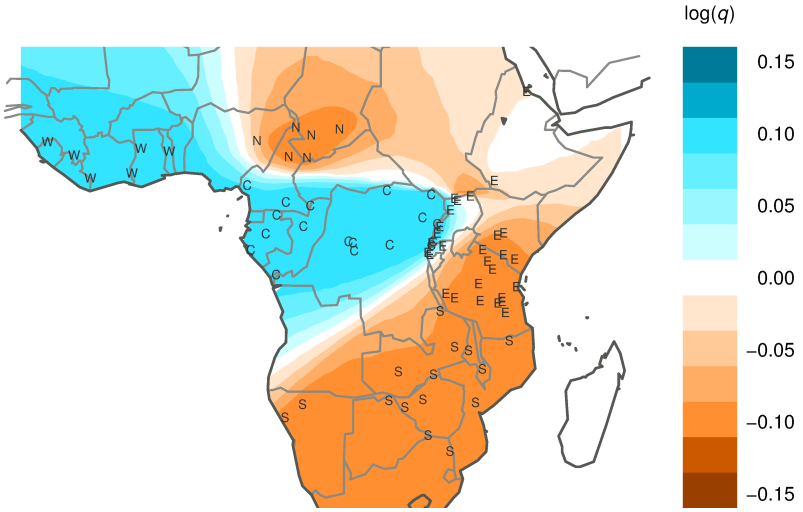
S6 Additional analysis of African elephant population structure



Supplementary Figure 5 Effective migration rates for the African elephant at sixteen microsatellite loci analyzed separately. The loci are highly polymorphic and therefore informative for the sample genealogy at each site as every mutation on the genealogy contains information about the branch lengths in the tree. (Longer branches are more likely to carry a mutation, if the mutation rate is constant in time.) The sixth locus is extremely informative, presumably because it has the highest mutation rate, and it successfully captures the strong effective barrier to migration between the habitat ranges of forest and savanna elephants.

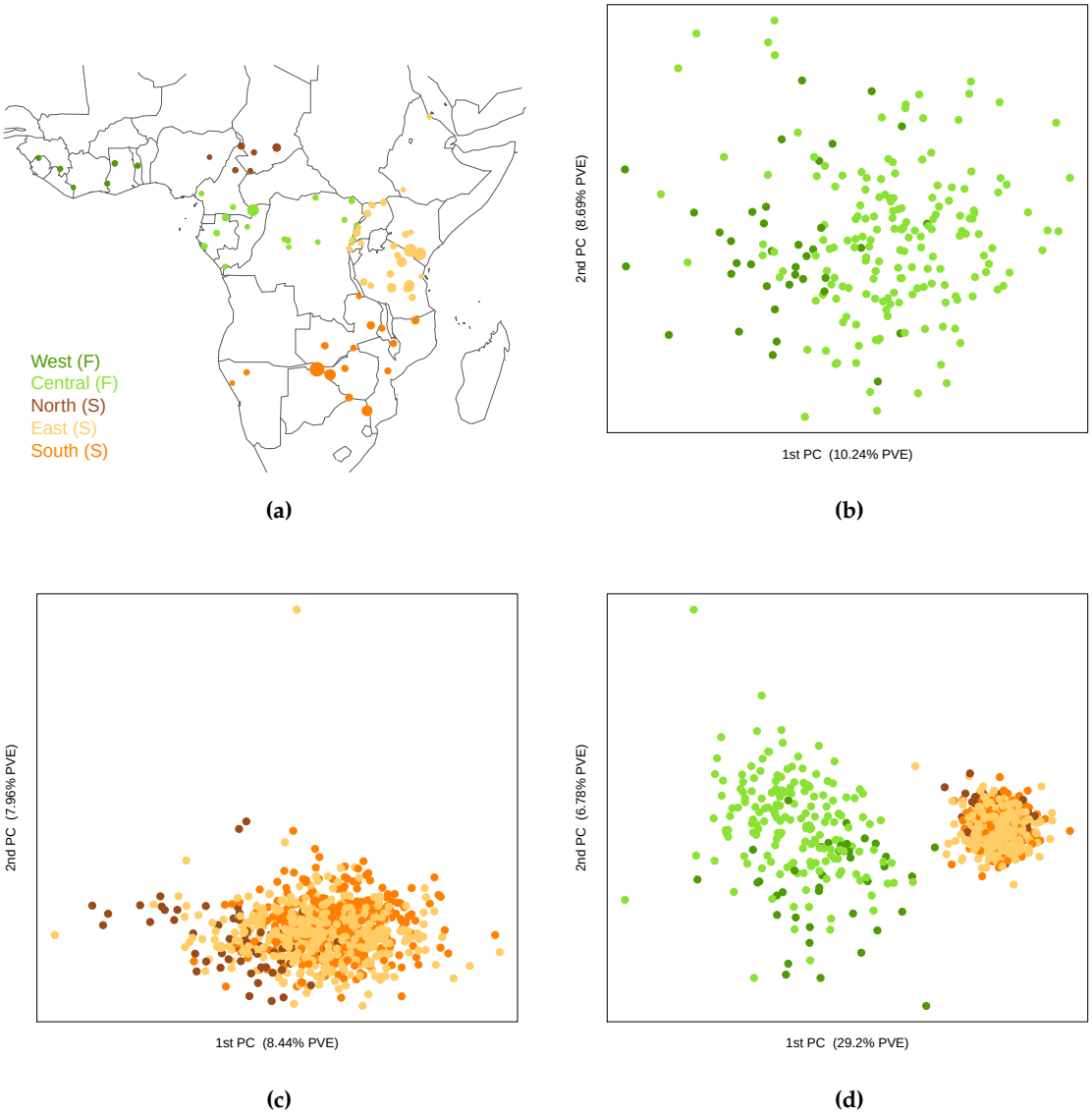


(a)

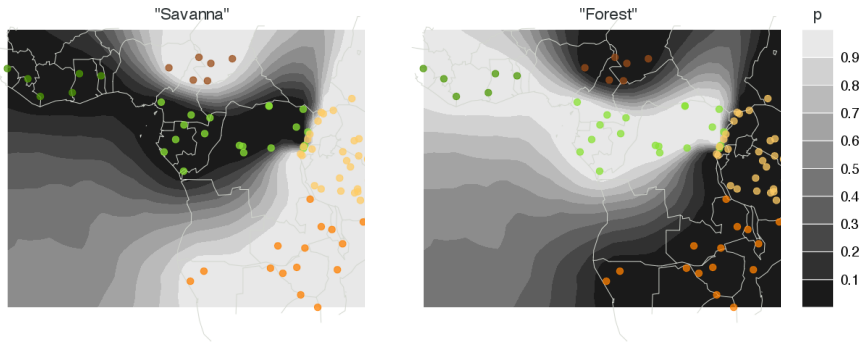


(b)

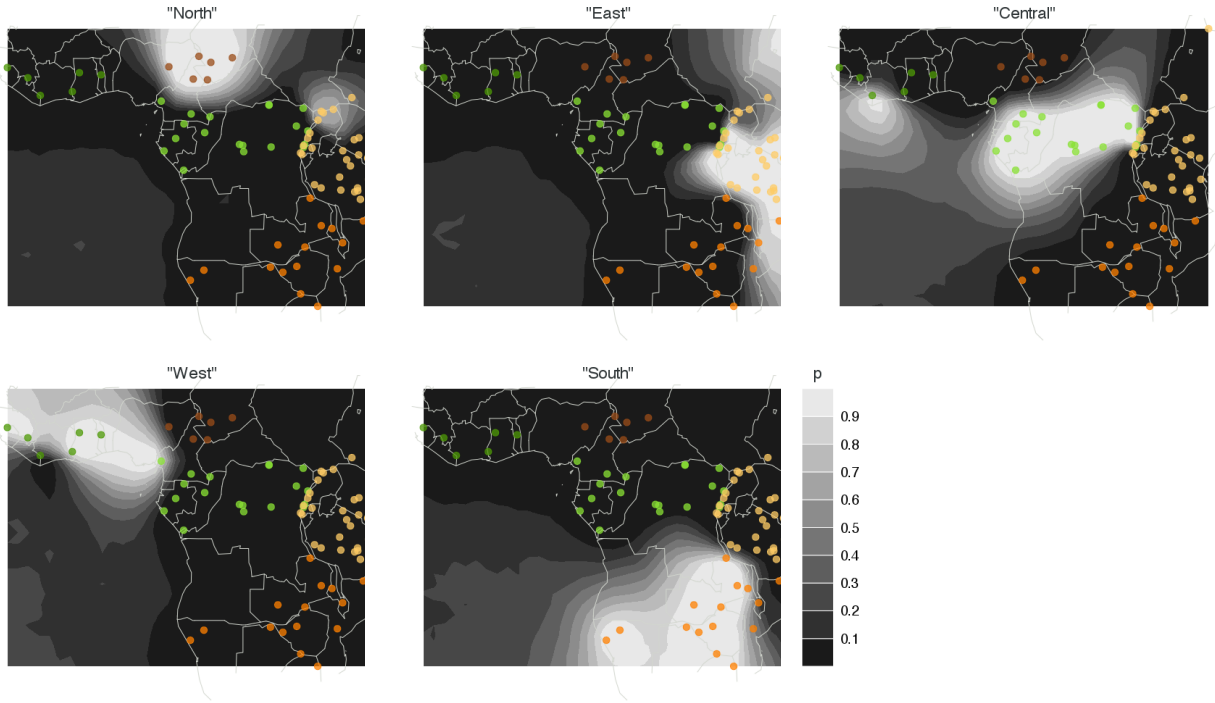
Supplementary Figure 6 Further EEMS analysis of the population structure of the African elephant data from (Wasser et al., 2015). According to the categorization in (Wasser et al., 2004), forest elephants come from two regions: West (W) and Central (C); savanna elephants come from three regions: North (N), East (E) and South (S). **(a)** Estimated effective migration surface, after excluding the most variable locus – the sixth locus in **Supplementary Figure 5**. The inferred surface, which does not change qualitatively after removing that locus, separates the two subspecies of African elephants, forest and savanna. **(b)** Estimated effective diversity rates using all sixteen loci. Forest elephants have higher effective diversity than savanna elephants. This is consistent with previous analysis which indicates that forest elephants have higher average heterozygosity (Comstock et al., 2002).



Supplementary Figure 7 Principal component analysis (PCA) of the African elephant data. **(a)** Original sampling locations in Sub-Saharan Africa. According to the categorization in (Wasser et al., 2004), the forest elephant subspecies inhabits the West and Central regions (in shades of green); the savanna elephant subspecies inhabits the North, East and South regions (in shades of orange). **(b-d)** First and second principal components of genetic variation, and the proportion of variance explained (PVE) as a percentage. **(b)** PCA of 211 forest elephants. **(c)** PCA of 914 savanna elephants. **(d)** PCA of 1124 African elephants, both forest and savanna.

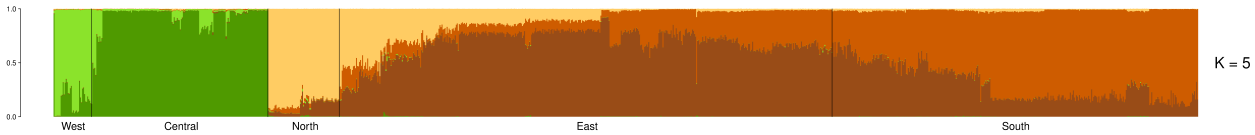


(a)

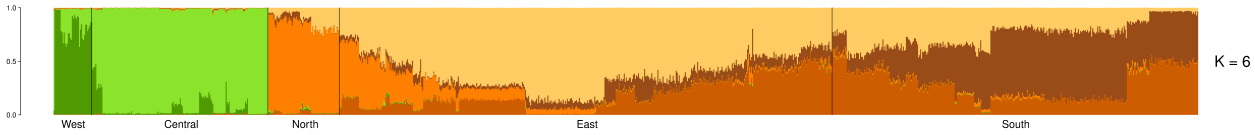


(b)

Supplementary Figure 8 GENELAND analysis of the African elephant data. GENELAND (Guillot et al., 2005) is a cluster-based method which uses Voronoi tessellations to encourage spatially continuous clusters and find sharp boundaries between genetically differentiated groups. **(a)** Posterior probabilities for belonging to each of two inferred clusters, which correspond to the forest and savanna habitats. **(b)** Posterior probabilities for belonging to each of five inferred clusters, which correspond roughly to the five biogeographic regions defined in (Wasser et al., 2004): "North", "South", "East", "West" and "Central". GENELAND successfully detects differences in allele frequencies between the two subspecies and the five biogeographic regions. However, GENELAND does not model the relationships between the regions: the five clusters in (b) are as distinct from each other as the two clusters in (a), even though the "West" and "Central" clusters are inhabited by forest elephants while the "North", "East" and "South" clusters – by savanna elephants.

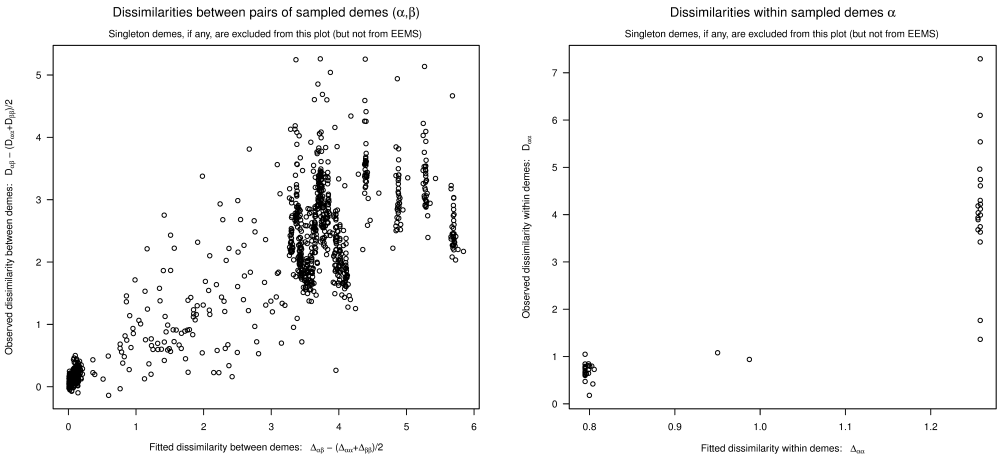


(a)

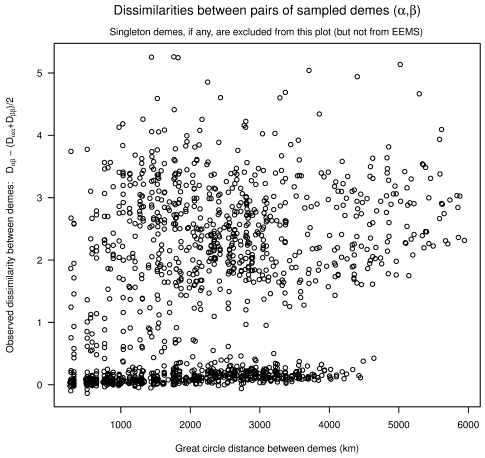


(b)

Supplementary Figure 9 STRUCTURE analysis of the African elephant data. **(a)** Membership proportions for belonging to five inferred clusters. **(b)** Membership proportions for belonging to six inferred clusters. Individuals are ordered by sampling location and the five biogeographic regions and separated by black vertical lines; the ancestral populations of forest and savanna elephants are colored in green and brown hues, respectively. Unlike GENELAND (Guillot et al., 2005), STRUCTURE (Pritchard et al., 2000) with a sampling location prior (Hubisz et al., 2009) provides intuition for the relationship between the five biogeographic regions. STRUCTURE clearly detects the difference between forest elephants (West and Central regions) and savanna elephants (North, East and South regions) as they fall into different clusters. Furthermore, STRUCTURE shows some evidence for isolation by distance, particularly in savanna elephants, as most of these individuals are represented as weighted mixtures of ancestral clusters that do not correspond to distinct geographic areas.



(a) **(b)**



(c)

Supplementary Figure 10 Observed vs fitted genetic dissimilarities, between and within demes, for the African elephants. The observed genetic dissimilarities between individuals, D_{ij}, are summarized according to the deme assignment: $D_{\alpha\beta} = \frac{1}{n_{\alpha\beta}} \sum_{\delta(i)=\alpha,\delta(j)=\beta,i\neq j} D_{ij}$ where $n_{\alpha\beta}$ is the number of pairs (i, j) such that sample i is assigned to deme α, sample j is assigned to deme β and i, j are distinct individuals. Singleton demes (those assigned one sample) are excluded from both scatter plots. **(a)** Dissimilarities between demes. The values $B_{\alpha\beta} = \Delta_{\alpha\beta} - (\Delta_{\alpha\alpha} + \Delta_{\beta\beta})/2$ comprise the between-demes component of genetic dissimilarity, B, which is modeled by the effective migration rates m (equation S14). **(b)** Dissimilarities within demes. The values $W_\alpha = \Delta_{\alpha\alpha}$ comprise the within-demes component of genetic dissimilarity, W, which is modeled by the effective diversity rates q (equation S13). If the EEMS model fits the data well, we expect a strong linear relationship between the observed and fitted values in both scatter plots. **(c)** Genetic dissimilarities against geographic distances between demes. Under isolation by distance (IBD), we expect a strong linear relationship between genetic dissimilarity and geographic distance. The African elephant violates exact IBD because forest and savanna elephants are strongly differentiated even though their habitats meet in the hybrid zone in Central Africa.

S7 Additional analysis of human population structure in Europe and Africa

Population	Symbol	Size	Comment about sample exclusions
Austria	AT	14	
Belgium	BE	43	
Denmark	DK	1	
France	FR	89	
Germany	DE	71	
Ireland	IE	61	
Italy	IT	214(219)	Removed 7623, 33242, 34049, 38532, 49500 as PCA outliers.
Netherlands	NL	17	
Portugal	PT	128	
Scotland	Sct	5	
Spain	ES	136	
Swiss-French	CHf	125	
Swiss-German	CHg	84	
Swiss-Italian	CHi	13	
United Kingdom	UK	200	

Table Description of the Western Europe dataset, which is a subset of POPRES data analyzed in (Novembre et al., 2008). It comprises 15 populations from 13 countries; their names and abbreviations, which generally correspond to ISO country codes, are given in the first and second column. The samples from Switzerland (CH) are split into three subpopulations: French, Italian and German speaking Swiss, coded as CHf, CHi and CHg, respectively. The number of samples from each population are given in the third column. We excluded five individuals as possible outliers based on their position in PC1-PC2 space: they project outside of the main Italian cluster and thus might have insular Italian ancestry – Sardinian or Sicilian (Novembre et al., 2008).

Population	Symbol	Size	Dataset	Comment about sample exclusions
Alur	Al	10	(Xing et al., 2010)	
Bambaran	Ba1	25	(Xing et al., 2010)	
Bamoun	Ba2	18	(Henn et al., 2011)	
Brong	Br	7 (8)	(Henn et al., 2011)	Removed 3572B (missingness > 5%).
Bulala	Bu	15	(Henn et al., 2011)	
Dogon	Do	24	(Xing et al., 2010)	
Fang	Fa	15	(Henn et al., 2011)	
Hausa	Ha	11 (12)	(Henn et al., 2011)	Removed NGHA019 (missingness > 5%).
Hema	He	13 (15)	(Henn et al., 2011)	Removed AFH7, AFH10 as PCA outliers.‡
Igbo	Ig	13 (15)	(Henn et al., 2011)	Removed NGIB007 and NGIB004 (missingness > 5%).
Kaba	Ka	17	(Henn et al., 2011)	
Kongo	Ko	9	(Henn et al., 2011)	
Luhya	Lu	23 (25)	(Henn et al., 2011)	Removed NA19027, NA19046 (not in HAP1117).†
Maasai	Ma1	21 (30)	(Henn et al., 2011)	Removed NA21528, NA21634, NA21447, NA21384, NA21382, NA21576, NA21616, NA21435, NA21405 (not in HAP1117).†
Mada	Ma2	12	(Henn et al., 2011)	
Mandenka	Ma3	22	(Henn et al., 2011)	
Nguni	Ng	9	(Xing et al., 2010)	
Pedi	Pe	10	(Xing et al., 2010)	
Sotho/Tswana	ST	8	(Xing et al., 2010)	
Xhosa	Xh	11	(Henn et al., 2011)	
Yoruba	Yo	21	(Henn et al., 2011)	Samples also in the Human Genetic Diversity Project.

Table Description of the Sub-Saharan Africa dataset, which combines data from (Xing et al., 2010) and (Henn et al., 2011). It comprises 21 ethnic groups (their names and abbreviations are given in the first and second column). The number of samples from each population are given in the third column and the source dataset in the fourth column. We excluded some samples because of issues with genotype quality or close familial relatedness. The final column indicates the excluded samples and the reason for exclusion. † The HAP1117 subset of HapMap3 excludes first- and second-degree relationships (Pemberton et al., 2010). ‡ These two Hema individuals – AFH7 and AFH10 – are classified as likely relatives and outliers in the analysis of Sub-Saharan Africa reported in (Wang et al., 2012).

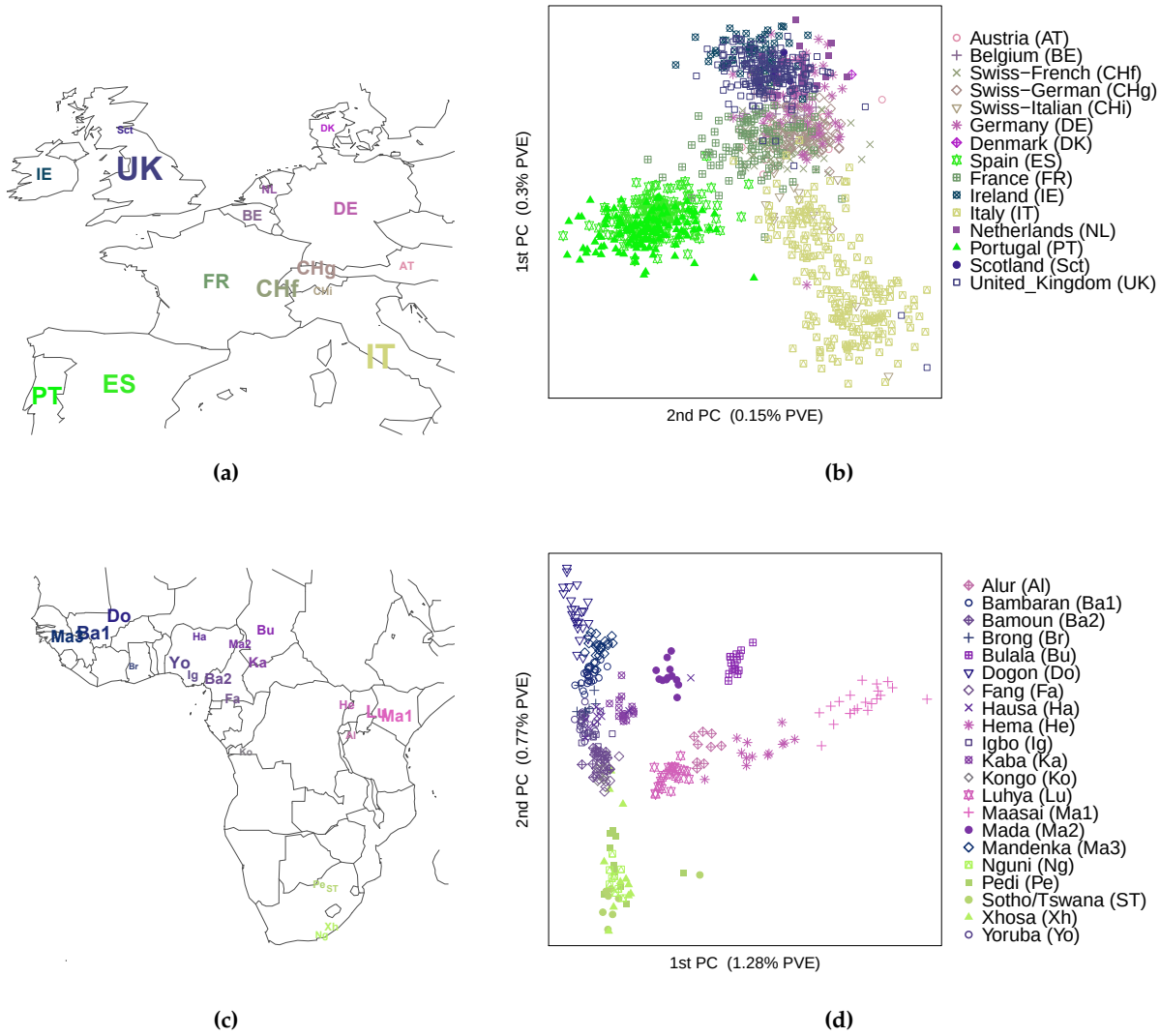
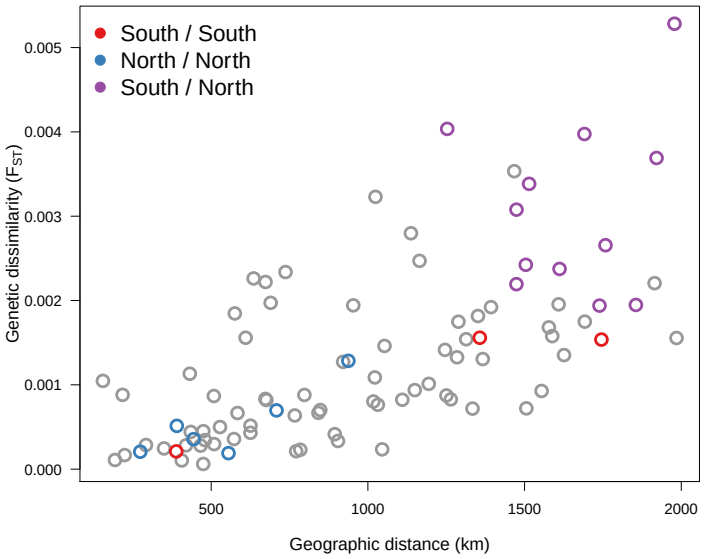
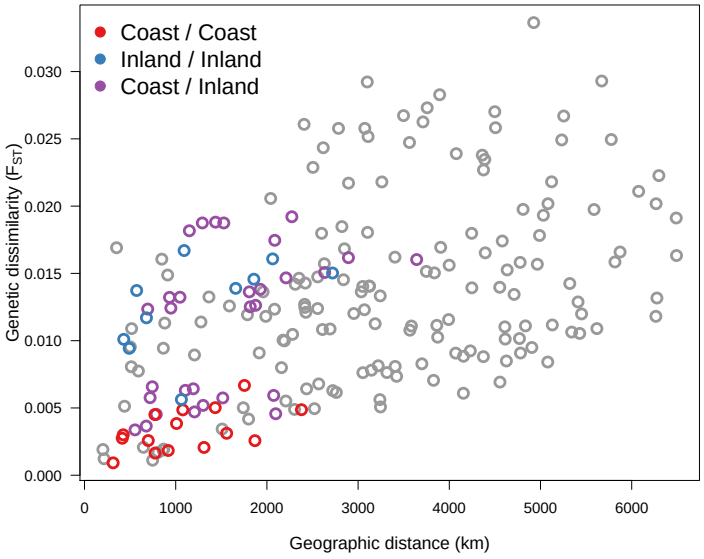


Figure Principal component analysis of humans in Western Europe and Sub-Saharan Africa. **(a)** We analyze 1201 individuals from 15 Western European populations. The data is part of the POPRES (Population Reference Sample) project (Nelson et al., 2008). The populations and their abbreviations are given in the legend on the right. The sampling is uneven and the size of the symbol indicates the relative number of individuals from each population. Colors are assigned according to latitude and longitude. **(b)** The first and second principle components are strongly correlated with latitude and longitude, as reported in (Novembre et al., 2008), and explain 0.3% and 0.15% of the individual genetic variation, respectively. **(c)** We analyze 314 individuals from 21 Sub-Saharan ethnic groups. The data is a compilation of two published SNP array datasets described in (Xing et al., 2010) and (Henn et al., 2011). Again, the symbols and colors are assigned according to geographic location and sample size. **(d)** The first and second principle components are strongly correlated with longitude and latitude, as reported in (Wang et al., 2012), and explain 1.3% and 0.8% of the individual genetic variation, respectively.

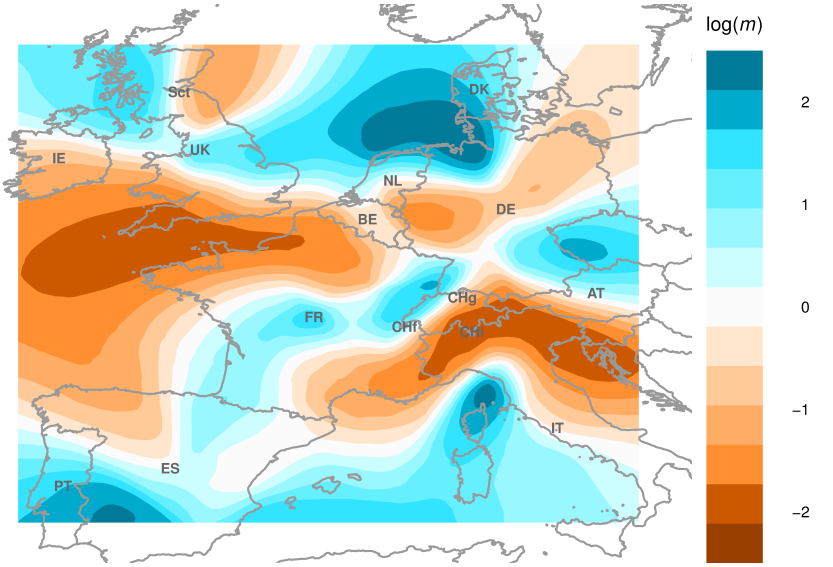


(a)

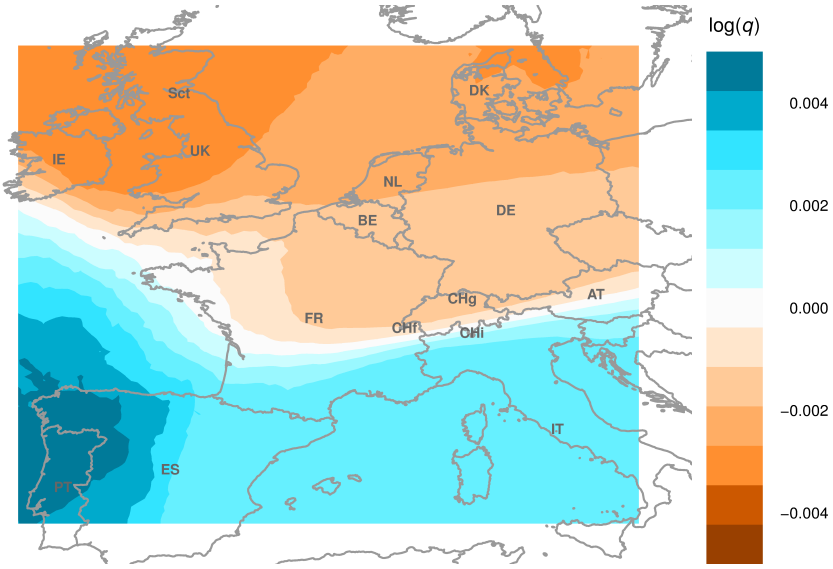


(b)

Supplementary Figure 11 Genetic dissimilarity (F_{ST}) as a function of geographic (great circle) distance, for pairs of human populations in Western Europe and Sub-Saharan Africa. On both continents, genetic differentiation increases with distance and this suggests that spatial variation is consistent with isolation by distance. The colors are chosen to emphasize comparisons between two groups of populations. **(a)** In Western Europe, the "south" group consists of Portugal (PT), Spain (ES), Italy (IT); the "north" group consists of Ireland (IE), Scotland (Sct), United Kingdom (UK), Holland (NL). Comparisons within the "south" and "north" groups are colored red and blue, respectively; comparisons between the two groups are colored purple. There is greater similarity within each group than between the groups. **(b)** In Sub-Saharan Africa, the "coast" group consists of Brong (Br), Yoruba (Yo), Igbo (Ig), Bamoun (Ba2), Fang (Fa), Kongo (Ko); the "inland" group consists of Hausa (Ha), Mada (Ma2), Bulala (Bu), Kaba (Ka), Hema (He). Coastal populations are more similar genetically than inland populations, even though some coastal populations are further apart.

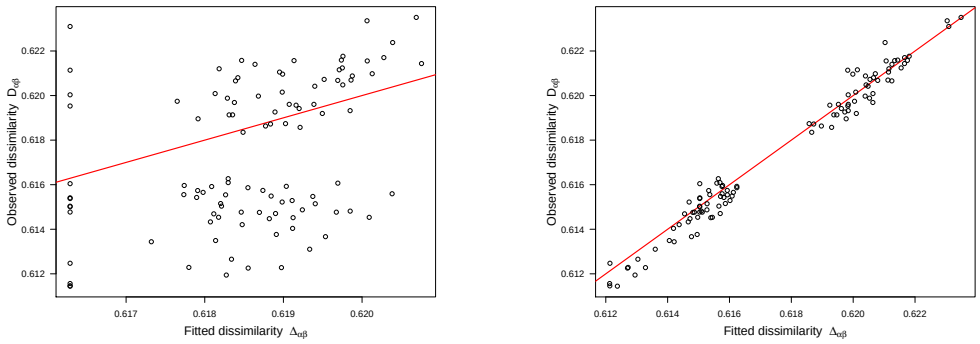


(a)

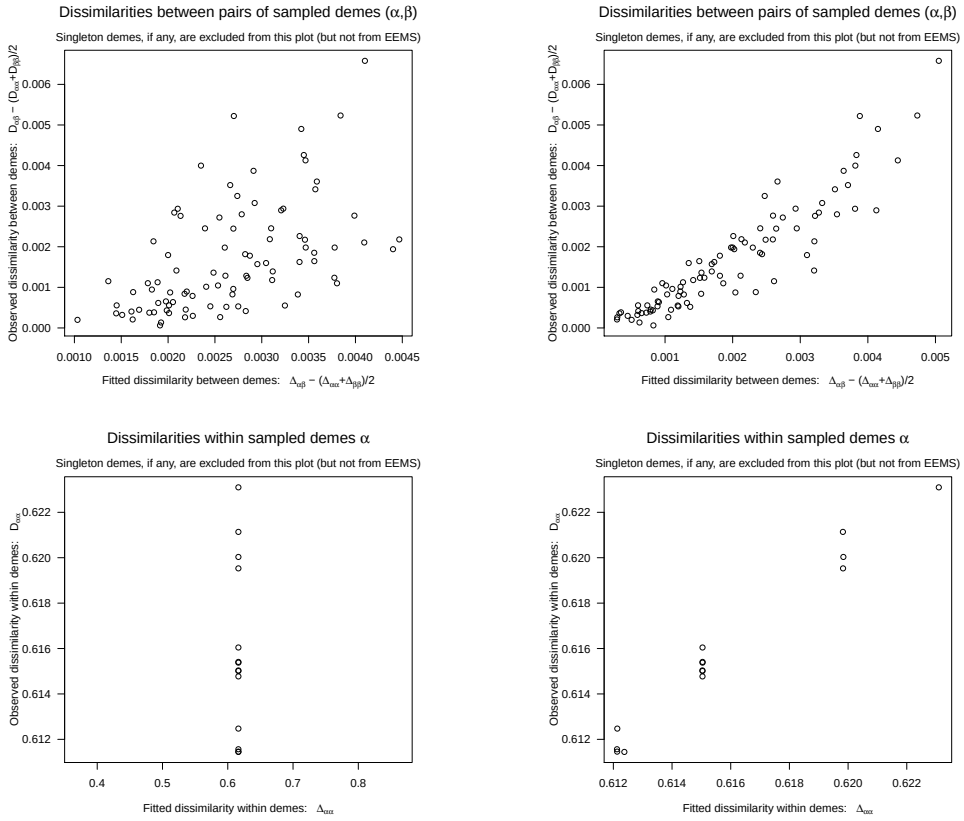


(b)

Figure EEMS analysis of the spatial structure of genetic variation in Western European, using data from the POPRES project (Nelson et al., 2008). EEMS estimates both the effective diversity rates within demes and the effective migration rates between connected demes. The fitted diversity rates can be interpolated across the habitat to produce an "estimated effective diversity surface", which is complementary to the "estimated effective migration surface". **(a)** Effective migration rates. This contour plot highlights effective barriers to migration in the north-south direction. **(b)** Effective diversity rates. This contour plot highlights the previously noted north-south gradient in human genetic diversity in Europe (Lao et al., 2008; Auton et al., 2009).

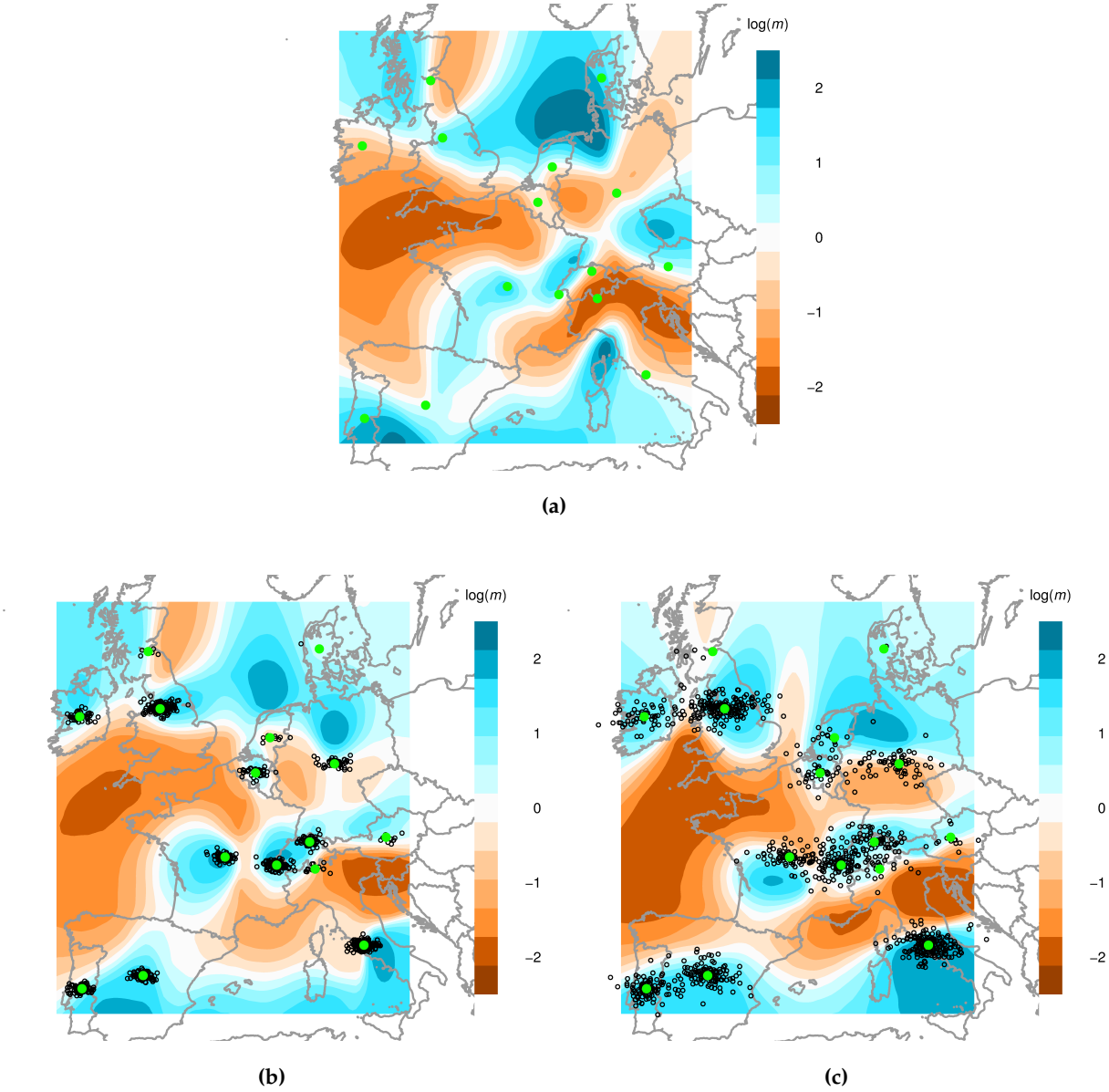


(a) **(b)**

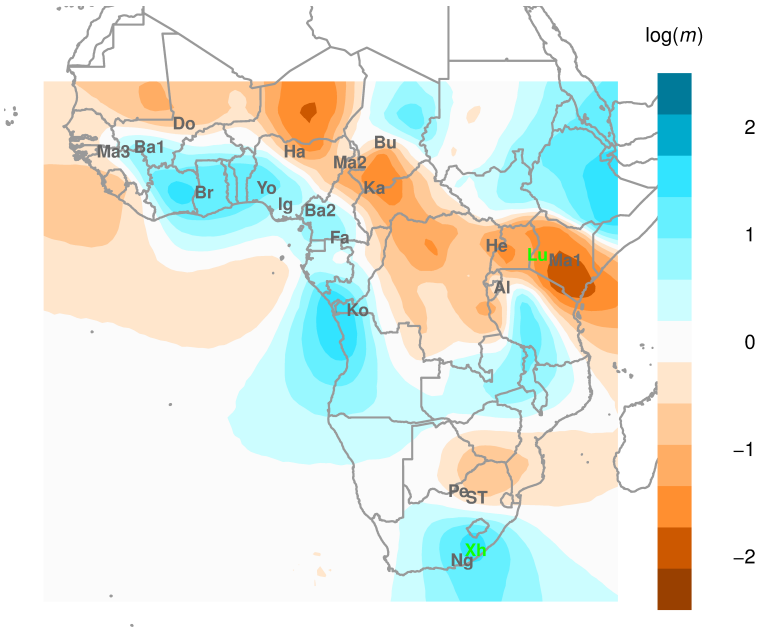


(c) **(d)**

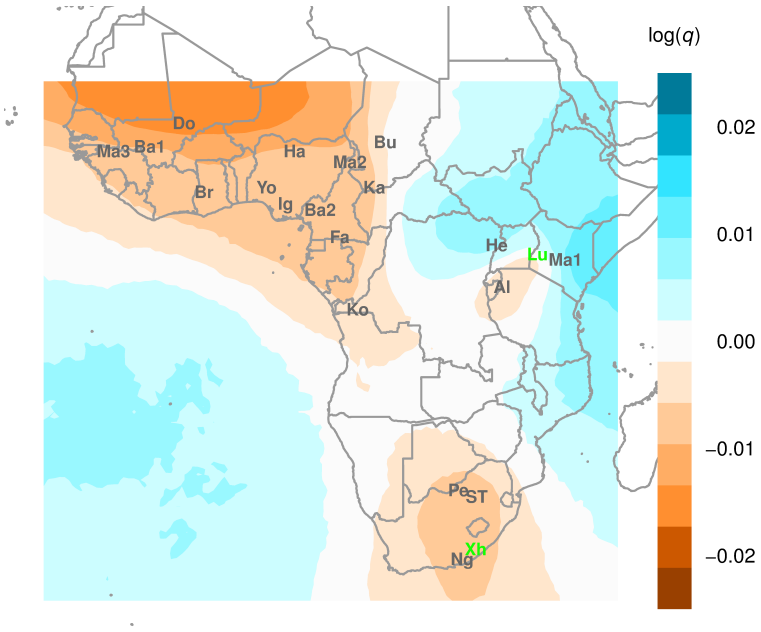
Supplementary Figure 12 Observed versus fitted dissimilarities between the 14 Western European populations. (We excluded Denmark (DK) because it has a single sampled individual.) EEMS attempts to estimate the effective migration and diversity rates so that the fitted genetic distances closely match the observed genetic differences; therefore, for a specific population grid, the coefficient of determination, r^2, between the fitted and observed values indicates the goodness-of-fit. For the results presented here we used a 19×15 grid. **(a)** Dissimilarities are modeled under the assumption of uniform migration, a setting which simulates exact isolation by distance ($r^2 = 0.142$). **(b)** Dissimilarities are modeled with EEMS, which estimates both the effective migration rates and the effective diversity rates, assuming equilibrium in time ($r^2 = 0.978$). Genetic dissimilarities can be further decomposed into a between-demes and a within-demes component, and the fitted and observed values in each component are plotted in **(c)** for IBD and **(d)** for EEMS. These two diagnostic plots are automatically generated by the EEMS software, to help assess the EEMS model fit.



Supplementary Figure 13 Robustness of estimated effective migration rates to unbiased location uncertainty, using data from the POPRES project (Nelson et al., 2008). In this dataset geographic information is imprecise: except for Switzerland, nationals from the same country are assigned to the central point of its area. Swiss individuals are categorized into Swiss-German, Swiss-French and Swiss-Italian, and assigned to three different locations within Switzerland. **(a)** Effective migration surface with the original assigned coordinates, indicated in green. **(b,c)** Effective migration surface after adding an unbiased random error to the assigned location of each individual. The "jittered" coordinates are in black, the original coordinates in green. The effective migration surface is robust to small unbiased location errors, except in sparsely sampled geographic regions. In this case, the effective migration estimates vary the most in the top right corner where there is a single individual from Denmark.

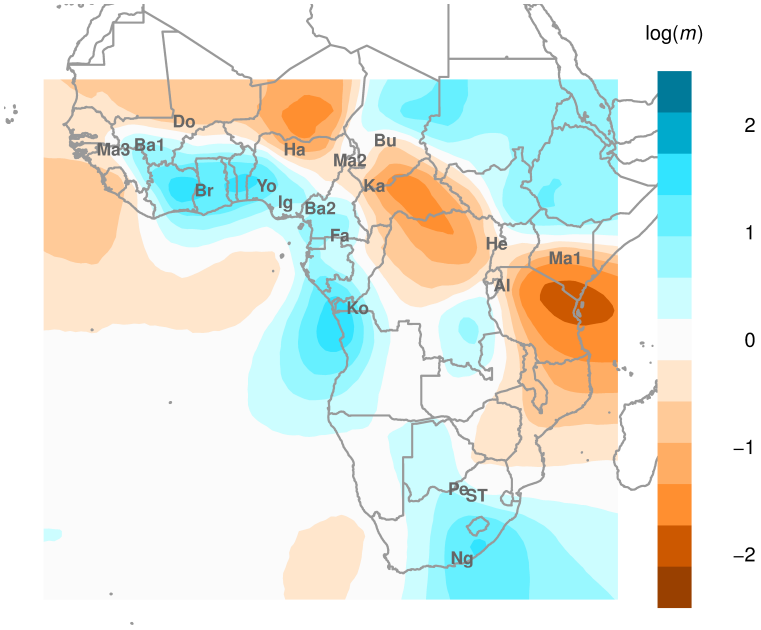


(a)

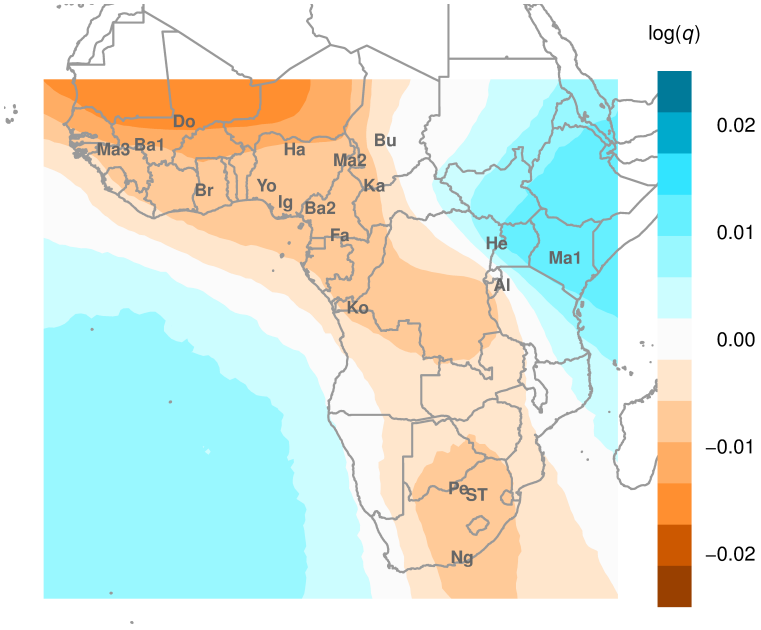


(b)

Figure EEMS analysis of 314 individuals from 21 Sub-Saharan African ethnic groups: Alur (Al), Bambaran (Ba1), Bamoun (Ba2), Brong (Br), Bulala (Bu), Dogon (Do), Fang (Fa), Hausa (Ha), Hema (He), Igbo (Ig), Kaba (Ka), Kongo (Ko), Luhya (Lu), Maasai (Ma1), Mada (Ma2), Mandenka (Ma3), Nguni (Ng), Pedi (Pe), Sotho/Tswana (ST), Xhosa (Xh), Yoruba (Yo). The Luhya and Xhosa populations, highlighted in green, are recent geographic migrants (Henn et al., 2011). We remove these two populations and re-analyze the Sub-Saharan data in **Supplementary Figure 14**. **(a)** Effective migration rates. This contour plot emphasizes that effective migration is higher along the Atlantic coast that it is inland. **(b)** Effective diversity rates. This contour plot emphasizes that effective diversity is higher in East Africa than it is in South or West Africa.

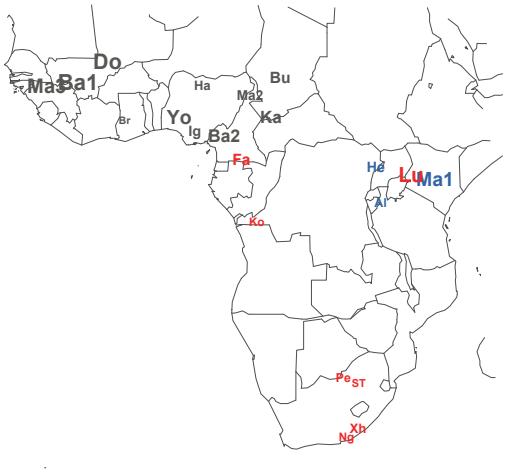


(a)

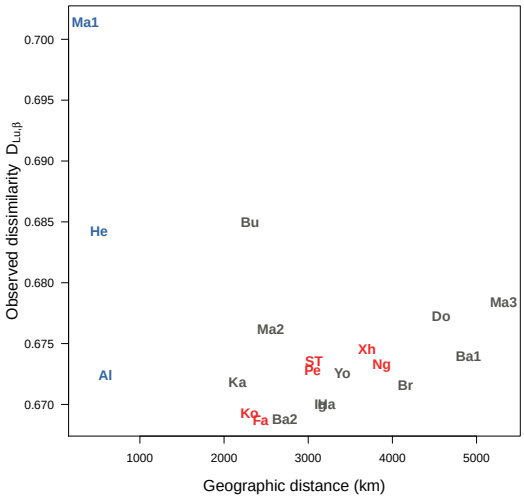


(b)

Supplementary Figure 14 EEMS analysis of the spatial structure in 19 Sub-Saharan African populations, after excluding the Luhya (Lu) and the Xhosa (Xh), two Bantu speaking populations considered recent geographic migrants in (Henn et al., 2011). EEMS approximates a spatial demographic model which evolves under equilibrium in time and recent migration deviates from this assumption. The other Bantu speaking populations are Pedi (Pe), Sotho/Tswana (ST) and Nguni (Ng) in the south and Fang (Fa) and Kongo (Ko) in the west. **(a)** Effective migration rates. **(b)** Effective diversity rates.

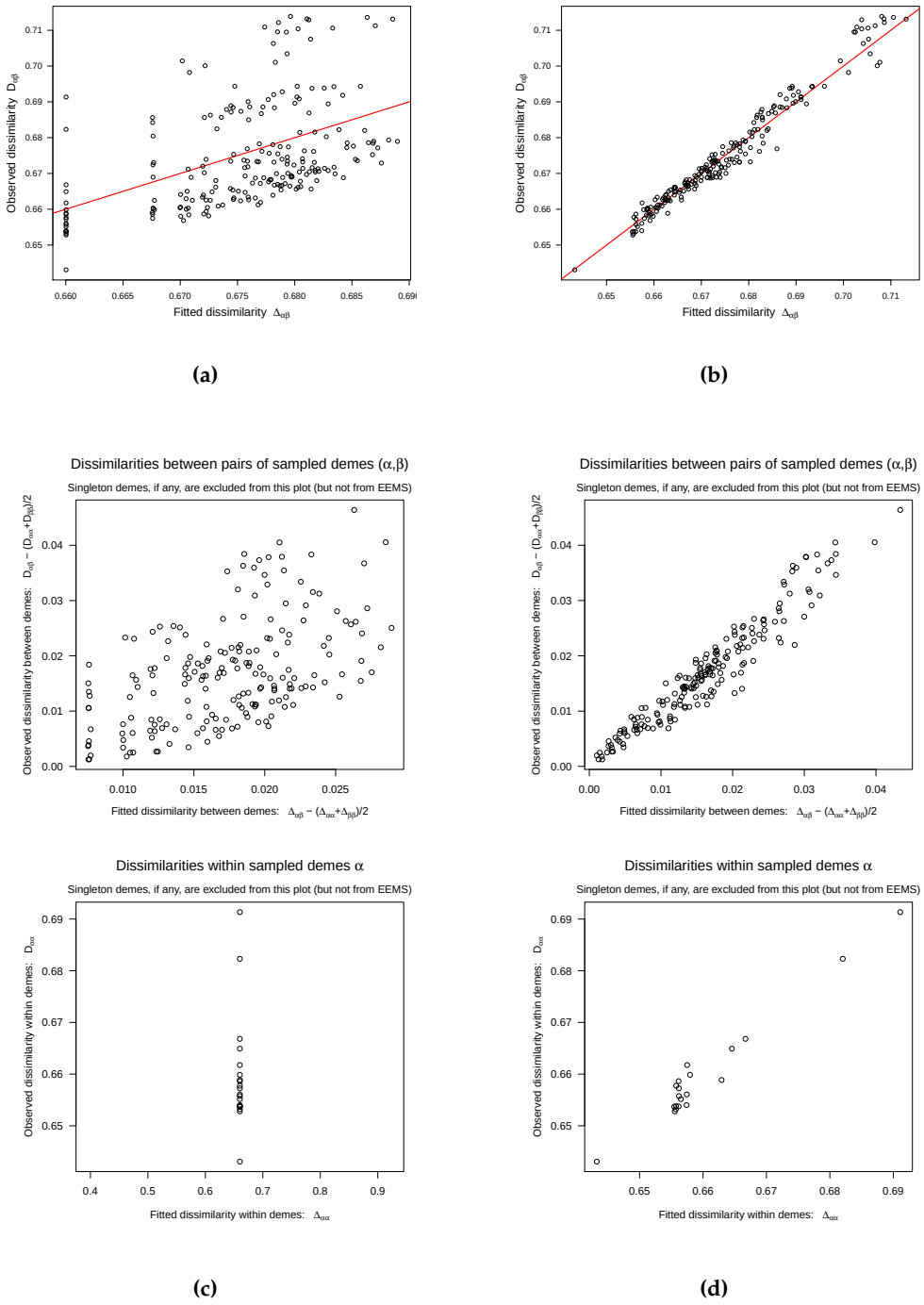


(a)



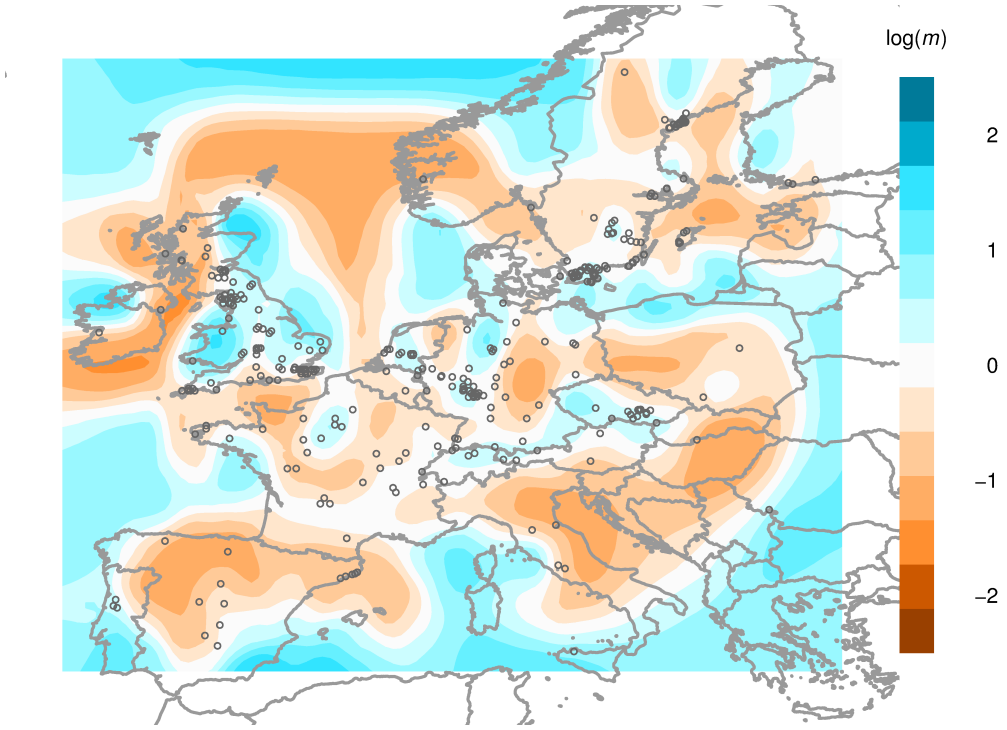
(b)

Supplementary Figure 15 Geographic distances and genetic dissimilarities between a recent migrant population (the Luhya, who speak a Bantu language) and 20 other ethnic groups in Sub-Saharan Africa. **(a)** The populations highlighted in red speak Bantu languages, the populations highlighted in blue speak Nilotic languages. The population names are given in the table on page 29. **(b)** Observed genetic dissimilarity vs geographic distance between one Bantu speaking population – the Luhya (Lu) in the east – and each of the other 20 ethnic groups in the Sub-Saharan Africa dataset. The Luhya are geographically close to the other ethnic groups in the east but are genetically distinct from the Hema (He) and the Maasai (Ma1). The Luhya are recent geographic migrants (Henn et al., 2011), which could explain the difference between the EEMS in Section S7, which includes the Luhya, and the EEMS in **Supplementary Figure 14**, which excludes the Luhya.

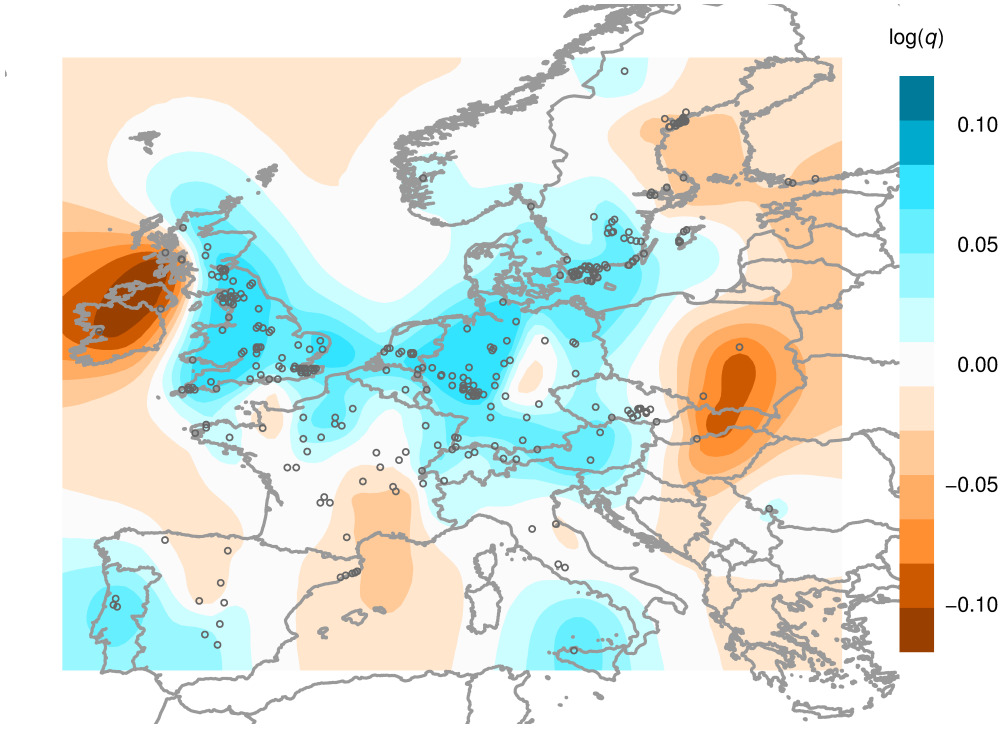


Supplementary Figure 16 Observed versus fitted dissimilarities between 21 Sub-Saharan ethnic groups. EEMS attempts to estimate the effective migration and diversity rates so that the fitted genetic distances closely match the observed genetic differences; therefore, for a specific population grid, the coefficient of determination, r^2, between the fitted and observed values indicates the goodness-of-fit. For the results presented here we used a 19×17 grid. **(a)** Dissimilarities are modeled under the assumption of uniform migration, a setting which simulates exact isolation by distance ($r^2 = 0.164$). **(b)** Dissimilarities are modeled with EEMS, which estimates both the effective migration rates and the effective diversity rates, assuming equilibrium in time ($r^2 = 0.914$). Genetic dissimilarities can be further decomposed into a between-demes and a within-demes component, and the fitted and observed values in each component are plotted in **(c)** for IBD and **(d)** for EEMS. These two diagnostic plots are automatically generated by the EEMS software, to help assess the EEMS model fit.

S8 Additional analysis of *A. thaliana* population structure

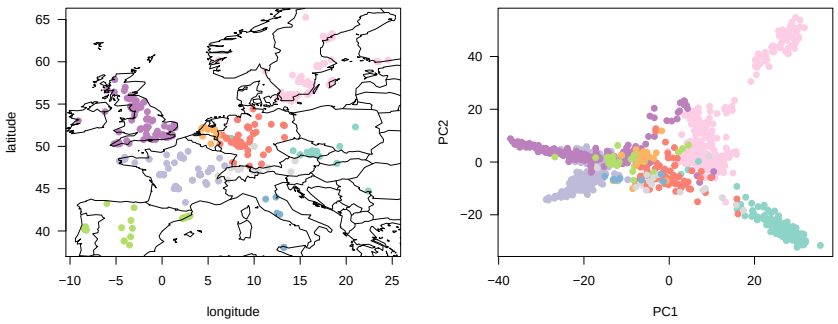


(a)

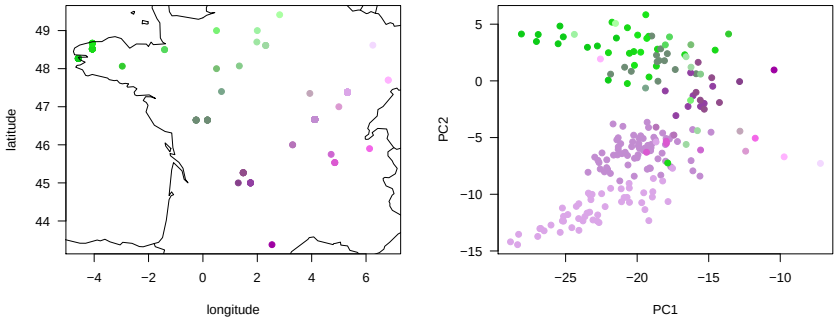


(b)

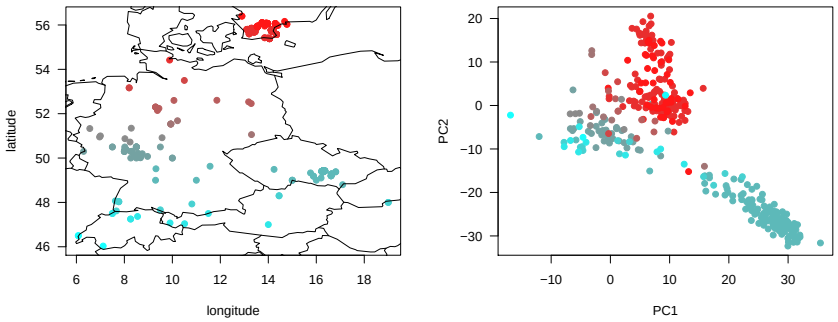
Figure EEMS analysis of 979 *Arabidopsis thaliana* accessions collected in Europe; geo-referenced data from the RegMap project (Horton et al., 2012). **(a)** Effective migration surface. **(b)** Effective diversity surface. There is less variation in effective migration and effective diversity is higher across the core natural range in Central Europe. Effective diversity tends to decrease in coastal regions and at the boundaries of the sampled habitat.



(a)



(b)



(c)

Supplementary Figure 17 Principal component analysis (PCA) of *Arabidopsis thaliana* in Europe. **(a)** Original sampling locations (left) and PCA projections (right). In both panels colors are assigned according to collection area, after grouping some neighboring countries together for clarity of presentation. **(b)** Original sampling locations (left) and PCA projections (right) of accessions collected in France. Colors are assigned according to latitude and longitude in order to emphasize the separation of NW samples from SE samples. The PCA plot shows the same projections as those in (a) but for the France subset of the accessions. **(c)** Original sampling locations (left) and PCA projections (right) of accessions collected in Central Europe and South Sweden. Colors are assigned according to latitude and longitude, using an altered rule, in order to emphasize a different spatial pattern.

References

Auton, A. *et al.* (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.*, 19:795–803.

Babić, D., Klein, D., Lukovits, I., Nikolić, S. & Trinajstić, N. (2002). Resistance-distance matrix: a computational algorithm and its application. *Int. J. Quantum. Chem.*, 90:166–176.

Bapat, R. B. & Raghavan, T. E. S. (1997). *Nonnegative matrices and applications*. Cambridge University Press, Cambridge, UK.

Baran, Y., Quintela, I., Ángel Carracedo, Pasaniuc, B. & Halperin, E. (2013). Enhanced localization of genetic samples through linkage-disequilibrium correction. *Am. J. Hum. Genet.*, 92:882–894.

Comstock, K. *et al.* (2002). Patterns of molecular genetic variation among African elephant populations. *Mol. Ecol.*, 11:2489–2498.

Duforet-Frebourg, N. & Blum, M. (2014). Nonstationary patterns of isolation-by-distance: inferring measures of local genetic differentiation with Bayesian kriging. *Evolution*, 68:1110–1123.

Epperson, B. K. & Li, T. (1996). Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proc. Natl. Acad. Sci. U.S.A.*, 93:10528–10532.

Guillot, G., Estoup, A., Mortier, F. & Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics*, 170:1261–1280.

Hanks, E. & Hooten, M. (2013). Circuit theory and model-based inference for landscape connectivity. *J. Am. Stat. Assoc.*, 108:22–33.

Henn, B. *et al.* (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc. Natl. Acad. Sci. U.S.A.*, 108:5154–5162.

Horton, M. *et al.* (2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.*, 44:212–216.

Hubisz, M., Falush, D., Stephens, M. & Pritchard, J. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.*, 9:1322–1332.

Hudson, R. (1990). Gene genealogies and the coalescent process. In Futuyma, D. & Antonovics, J., editors, *Oxford surveys in evolutionary biology*, volume 7, pages 1–44. Oxford University Press.

Kimura, M. & Weiss, G. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49:561–576.

Lao, O. *et al.* (2008). Correlation between genetic and geographic structure in Europe. *Curr. Biol.*, 18:1241–1248.

Lawson, D., Hellenthal, G., Myers, S. & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet.*, 8:e1002453.

Leslie, S. *et al.* (2015). The fine-scale genetic structure of the British population. *Nature*, 519:309–314.

Light, A. & Bartlein, P. (2004). The end of the rainbow? Color schemes for improved data graphics. *Eos*, 85:385.

Manel, S. *et al.* (2007). A new individual-based spatial approach for identifying genetic discontinuities in natural populations. *Mol. Ecol.*, 16:2031–2043.

Manni, F., Guerard, E. & Heyer, E. (2004). Geographic patterns of (genetic, morphologic, linguistic) variation: How barriers can be detected by using Monmonier's algorithm. *Hum. Biol.*, 76:173–190.

McRae, B. (2006). Isolation by resistance. *Evolution*, 60:1551–1561.

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

Nelson, M. *et al.* (2008). The population reference sample, POPRES: A resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.*, 83:347–358.

Novembre, J. *et al.* (2008). Genes mirror geography within Europe. *Nature*, 456:98–101.

Novembre, J. & Slatkin, M. (2009). Inference in the continuous isolation-by-distance model: Using low-frequency alleles to estimate scaled dispersal distance. *Evolution*, 63:2914–25.

Ohta, T. & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.*, 22:201–204.

Pemberton, T., Wang, C., Li, J. & Rosenberg, N. (2010). Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *Am. J. Hum. Genet.*, 87:457–464.

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

Rañola, J. M., Novembre, J. & Lange, K. (2014). Fast spatial ancestry via flexible allele frequency surfaces. *Bioinformatics*, 30:1847–1860.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components —an alternative to reversible jump methods. *Ann. Stat.*, 28:40–74.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Aust. J. Stat.*, 32:227–230.

Wang, C., Zöllner, S. & Rosenberg, N. (2012). A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet.*, 8:e1002886.

Wasser, S. *et al.* (2004). Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proc. Natl. Acad. Sci. U.S.A.*, 10:14847–14852.

Wasser, S. K. *et al.* (2015). Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots. *Science*, 349:84–87.

Xing, J. *et al.* (2010). Toward a more uniform sampling of human genetic diversity: A survey of worldwide populations by high-density genotyping. *Genomics*, 96:199–210.

Yang, W.-Y., Novembre, J., Eskin, E. & Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. *Nat. Genet.*, 44:725–731.