bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Visualizing spatial population structure with estimated effective migration surfaces

Desislava Petkova[1], John Novembre[2] & Matthew Stephens[1,2]

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

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

Genetic data often exhibit patterns that are broadly consistent with "isolation by distance" – a phenomenon where genetic similarity tends to decay with geographic distance. In a heterogeneous habitat, decay may occur more quickly in some regions than others: for example, barriers to gene flow can accelerate the genetic differentiation between groups located close in space. We use the concept of "effective migration" to model the relationship between genetics and geography: in this paradigm, effective migration is low in regions where genetic similarity decays quickly. We present a method to quantify and visualize variation in effective migration across the habitat, which can be used to identify potential barriers to gene flow, from geographically indexed large-scale genetic data. Our approach uses a population genetic model to relate underlying migration rates to expected pairwise genetic dissimilarities, and estimates migration rates by matching these expectations to the observed dissimilarities. We illustrate the potential and limitations of our method using simulations and data from elephant, human, and *Arabidopsis thaliana* populations. The resulting visualizations highlight important features of the spatial population structure that are difficult to discern using existing methods for summarizing genetic variation such as principal components analysis.

All natural populations exhibit some degree of "structure": some individuals are more closely related than others. Population structure is shaped by many factors, but probably most influential are the barriers to gene flow that the population has experienced over its evolutionary history – barriers that may be due to extrinsic factors (such as topography or environment) or intrinsic factors (such as mate recognition, reproductive compatibility, or more complex interactions in social species such as humans). Studying the genetic structure of a population can therefore yield important scientific insights into the demographic and evolutionary processes that have shaped the population [1, 2], and help answer questions related to, for example, adaptation [3], speciation [4], hybridization [5], introgression [6] and recombination [7]. Understanding the genetic structure of a population may also be useful in contexts other than population genetics – for example, to identify subsets of genetically distinct groups that may require special conservation status [8], or to help correct for confounding in genetic association studies [9, 10].

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

These important questions have motivated the development of many statistical methods for analyzing population structure. Among these, admixture-based clustering [11, 12, 13] and principal components analysis (PCA) [14, 15] are most widely used. A valuable feature of both approaches is that they summarize the main patterns of population structure in explicit and intuitive visual representations. Visual summaries are especially useful tools: not only can they help generate and refine hypotheses about the biological and evolutionary processes that have shaped the population, but they can also help identify sample outliers or other unexpected patterns, which are key steps in any analysis. Aside from this shared feature, admixture-based and PCA-based methods have their own distinct strengths and limitations. Clustering methods are particularly useful when the population under study is well represented by a small number of relatively distinct groups, possibly with recent admixture among groups. However, clustering methods are less well adapted to settings that exhibit more "continuous" patterns of genetic variation, such as "isolation by distance" where genetic similarity tends to decay with geographic distance [16]. In comparison, PCA is arguably better adapted to more continuous settings [14], and has proven helpful in diagnosing isolation by distance as a feature of the data [17]. However, certain properties of PCA complicate its interpretation and limit the insights it can provide. For example, PCA is heavily influenced by sampling biases: more data being collected preferentially from some regions than others [18, 19, 20]. And while PCA projections are often interpreted post hoc with geographic information in hand, PCA itself ignores the sampling locations even if they are known – information that can be particularly helpful if the data exhibit some degree of isolation by distance.

Motivated by this, we have developed a novel tool for visualizing population structure in an important setting that is not ideally served by existing methods: the setting where individuals are sampled from known locations across a spatial habitat (the samples are "geo-referenced"), and where the population structure is broadly, but perhaps not entirely, consistent with isolation by distance. We aim to produce visualizations which highlight regions that deviate from exact isolation by distance, and thus identify corridors or barriers to gene flow, if they exist. Our work shares goals with several previous methods [21, 22, 23], although the approach we take here is less algorithmic, as it explicitly represents genetic differentiation as a function of the migration rates in an underlying population genetic model. Our model-based approach is conceptually related to early work on inferring migration rates from genetic data [24], although the details – and particularly the use of "resistance distance", a concept introduced in population genetics by [25] – are much closer to recent work on landscape connectivity [26] (see Discussion).

Results

Outline of the EEMS method

Figure 1 provides a schematic overview of our approach. In brief, the method is based on the "stepping stone" model [27], in which individuals migrate locally between subpopulations ("demes"), with symmetric migration rates that can vary by location. In order to capture "continuous" population structure, we use a dense regular grid of demes spread across the habitat, with each deme exchanging migrants only with its

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

neighbours. Under the stepping stone model, expected genetic dissimilarities depend on the locations of samples and on the migration rates between connected demes. The expected genetic dissimilarity between a pair of individuals can be computed by integrating over all possible migration histories in their genetic ancestry, and we approximate it using a distance metric from circuit theory which integrates all possible migration routes between a pair of demes [25]. Our method effectively involves adjusting the migration rates so that the genetic differences *expected* under the model closely match the genetic differences *observed* in the data, while at the same time respecting the fact that nearby edges will often tend to have similar migration rates. The end result is an estimate of the migration rate on every edge in the graph, which we interpolate across the habitat to produce an "Estimated Effective Migration Surface" or EEMS. The EEMS provides a visual summary of the observed genetic dissimilarities among samples, and how they relate to geographic location. For example, if genetic similarities tend to decay faster with geographic distance in some parts of the space, this will be reflected by a lower value of the EEMS in those areas. If, on the other hand, genetic similarities tend to decay in the same way with distance throughout the habitat, the EEMS will be relatively constant. We use the term "effective" because the model makes assumptions – most importantly, equilibrium in time – that may preclude interpreting the EEMS as representing historical rates of gene flow. Nonetheless, as we illustrate on several examples, the method provides an intuitive and informative way to quantify and visualize patterns of population structure in geographically structured samples.

Simulations under the stepping stone model

We illustrate the benefits and limitations of the EEMS method with several simulations. Using the program `ms` [28], we simulated data from a stepping stone model under two different migration scenarios: a "uniform" scenario, in which migration rates do not vary throughout the habitat, intended to represent a pure isolation by distance situation (Fig. 2a); and a "barrier" scenario, in which a central region with lower migration rates separates the left and right sides of the habitat (Fig. 2b). We applied both EEMS, and – for comparison – PCA, to data generated under these scenarios and under three different sampling schemes (Fig. 2c). The results illustrate two key points. First, whatever the sampling scheme, the underlying simulation truth is much easier to discern from the EEMS contour plots (Fig. 2e) than from the PCA projections (Fig. 2d). For the pure isolation by distance setting, the EEMSs are approximately uniform under all three sampling schemes, and for the barrier scenario the EEMSs highlight the barrier as an area of lower effective migration. In contrast, the simple nature of the underlying structure is not obvious from the PCA projections for either scenario, and indeed, the PCA results for the different scenarios do not differ in an easily identifiable, systematic way. Second, EEMS is much less sensitive to the underlying sampling scheme than is PCA. Indeed, the inferred EEMSs are qualitatively unaffected by sampling scheme, except in the extreme case where there are no samples taken on one side of the migration barrier (which renders the migration rates on that side of the barrier inestimable from the data, so that estimates in that region are driven by the prior: no heterogeneity in migration rates). In contrast, PCA shows its known proclivity to be heavily influenced by irregular sampling [18, 19, 20]. For example, both biased sampling or the presence of a barrier can produce clusters in the PCA results (top row,

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 2d).

Effective migration vs migration probabilities

Population genetics makes extensive use of the notion of "effective population size", which can be informally defined as the size of an idealized (random-mating, constant-sized) population that would produce similar patterns of genetic variation as those observed in the population. The effective population size is typically quite different from the census size of the population. Similarly, the EEMS should be interpreted to represent a set of "effective migration rates" that, within an idealized stepping stone model evolving under equilibrium in time, would produce similar genetic dissimilarities as those observed in the data. Therefore, the effective migration rates may be different from actual migration rates of individuals in the population.

To illustrate this idea we present simulations under two different migration scenarios, each producing an EEMS with an "effective barrier" to migration, but for different reasons. In the first simulation the effective barrier results from a lower population density in the central region (Fig. 3a); in the second simulation the effective barrier results from the populations splitting some time in the past (Fig. 3b). In both cases the EEMS correctly reflects the structure in the observed genetic differentiation: individuals on either side of the central region are less genetically similar than expected based on distance alone (under pure isolation by distance). Indeed, in both cases the EEMS qualitatively reflects average rates of historical gene flow. However, it should be clear that care is warranted in linking the EEMS to inferences about the actual underlying migration processes.

These results also illustrate the obvious but important fact that, because the EEMS method characterizes expected genetic differentiation under migration, it cannot distinguish between different scenarios that produce similar expectations for the pairwise genetic dissimilarities. This is also true of PCA [19]. In some cases it may be possible to distinguish among such scenarios using other aspects of the data, but we do not pursue this here.

Empirical results

To illustrate EEMS in practice we present results for four diverse empirical datasets: an African elephant dataset with strong differentiation between two geographically divided subspecies; two human datasets with individuals sampled from across Europe and across Sub-Saharan Africa where genetic differentiation has been shown to vary (somewhat) continuously with latitude and longitude; and an *Arabidopsis thaliana* dataset whose genetic variation is characterized by strong genetic similarity between Europe, where the plant is native, and North America, which it colonized in the last three hundred years [29].

African elephants in Sub-Saharan Africa

The African elephant (*Loxodonta africana*) has two recognized subspecies – the forest elephant (*L. a. cyclotis*) and the savanna (or bush) elephant (*L. a. africana*). Both subspecies are under threat, partly from poaching, and a large sample was collected and genotyped at 16 microsatellite loci to help assign contraband tusks to

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

their location of origin, and thus facilitate conservation efforts [30]. We analyze the augmented reference data from [31], which contains 211 forest and 913 savanna elephants.

These data provide a helpful illustration of the EEMS method because the subspecies structure is clear and strongly correlated with geography, so we know the primary structure that we would like our method to highlight: the low effective gene flow between forest and savanna elephants despite their geographic proximity. And, indeed, the EEMS for these data is dominated by a strong barrier separating forest and savanna locations (Fig. 4b). Thus, the African elephant provides an empirical example of an EEMS barrier due to a non-equilibrium history of drift after divergence, as in the simulation for Figure 3b. Notably, the EEMS successfully captures some of the winding shape of the geographic barrier between forest and savanna habitats, despite the fact that our method, based on Voronoi tilings, seems better adapted to capture barriers with simpler structure.

For the African elephant, one of the sixteen genotyped loci is extremely informative: the EEMS inferred from this locus alone is similar to that from all sixteen loci (Supplementary Fig. 3). However, the EEMS from the remaining fifteen loci is also similar to the EEMS for all sixteen loci (Supplementary Fig. 4a), demonstrating that this strongly differentiated locus is consistent with the others. (In principle, differences in inferred EEMSs among loci could provide a test for selection [32], but we do not pursue this further here.)

Because the strong differentiation between forest and savanna elephants dominates the EEMS, we also separately analyzed forest and savanna samples to examine subtler structure within each group. The presence of substructure has been detected previously [30] as the African elephant habitat can be divided into five broad biogeographic regions: West and Central forest regions, and North, East and South savanna regions (Fig. 4a). The EEMS for forest elephants (Fig. 4c) shows no strong deviation from uniform migration, suggesting that the genetic structure of the forest elephant is broadly consistent with isolation by distance. The EEMS for savanna elephants (Fig. 4d) shows a barrier separating elephants in the North region from the rest, and a "corridor" of higher effective migration connecting the South and the East regions. The barrier coincides with forest habitat that forms a known barrier to migration for the savanna elephant. The corridor is consistent with previous observations, from mitochondrial data, that the South and East regions are genetically more similar than their geographic distance would suggest [33]. These patterns are harder to recognize in the corresponding PCA plots (Supplementary Fig. 5) or admixture and cluster-based analyses (Supplementary Figures 6 and 7).

In addition to the migration rates, our method also estimates an effective diversity parameter within each deme, which reflects the expected genetic dissimilarities of two individuals sampled from that deme. For the African elephant, the inferred effective diversities are higher in the forest regions than in the savanna regions (Supplementary Fig. 4b). This represents in a direct, visual way the observation that forest elephants have higher heterozygosity than savanna elephants [34].

Humans in Europe and Sub-Saharan Africa

We analyze two large-scale genome-wide datasets to visualize the genetic structure of human populations on two continents: a collection of 1201 individuals from 13 Western European countries genotyped at 197K SNPs [17, 35] and a collection of 314 individuals from 21 Sub-Saharan African ethnic groups genotyped at 28K SNPs

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[36].

Previous PCA-based analyses of both datasets [17, 37, 36] have found that the two leading PCs are correlated with geographic location. This suggests that genetic similarity tends to decay with geographic distance (Supplementary Fig. 8), and therefore, in broad terms, the data are consistent with isolation by distance. EEMS analysis on the other hand highlights patterns that deviate from *stationary* (exact) isolation by distance (Fig. 5). However, for both datasets, one must exercise caution with detailed interpretations of the EEMS results as the geographic point locales for each individual are known only coarsely (see Discussion).

In Europe (Fig. 5a), the areas of highest effective migration span the North Sea and the Mediterranean, likely due to historic contacts between populations bordering these bodies of water; other areas of high migration span central France and Austria. Some regions of lower effective migration align with topographic barriers: Northern Italy and the Atlantic. An area of low migration also spans Germany. Overall the results are consistent with the idea that population structure in Europe is characterized by a north/south cline [38], but the EEMS also elucidates more complex patterns of differentiation in the north/south direction. While the PCA plot may visually suggest a simple relationship between genetics and geography, the EEMS results are far from a constant surface that one would expect from stationary isolation by distance; the EEMS helps highlight patterns in the genetic dissimilarity matrix that are difficult to discern in the corresponding PCA projections [17].

In Africa (Fig. 5b), the EEMS highlights a corridor of higher effective migration along the Atlantic coast, relative to lower migration rates inland. This indicates that – at a given distance apart – the coastal populations are more genetically similar than the inland populations. The U-shaped tail of the corridor moving inland could highlight higher than expected (from the geographic distance alone) genetic similarity between some ethnic groups in the west and those in the east (Supplementary Fig. S14). Non-genetic information about the subpopulations can help clarify this pattern: in this case, the Fang (Fa) and the Kongo (Ko) in the west, and the Luhya (Lu) in the east speak Bantu languages, so we might hypothesize that the link is partly due to shared ancestry between Bantu speaking groups. In an EEMS analysis after excluding the Luhya, the definition of the corridor that connects the east with the west is greatly decreased (Supplementary Fig. S12), which supports the hypothesis that this signal is driven partly by the genetic similarity between Bantu speaking peoples.

Our EEMS method attempts to explain observed genetic dissimilarities using a stepping stone model with varying migration rates. Some data sets may contain features that are not captured by this model – for example, recent long-distance migrants. As a model-checking diagnostic we suggest comparing the fitted genetic dissimilarities and the observed genetic dissimilarities. For both human datasets, the fitted and observed dissimilarities generally agree well (Supplementary Figures 10 and 13). Furthermore, they agree much better than under a constant migration model: for example, the proportion of variance explained increases from 14.2% to 97.8% for the European data, and from 16.4% to 91.4% for the African data. This illustrates that the estimated non-stationary migration pattern is a better explanation for the observed patterns of spatial differentiation than exact isolation by distance.

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Arabidopsis thaliana in Europe and North America

Arabidopsis thaliana is a small flowering plant with natural range in Europe, Asia and North Africa, and which is now found in North America as well. Although *A. thaliana* is a selfing plant with low gene flow, its genetic variation has significant spatial structure [39, 40]. On the continental scale, in Europe the data exhibit patterns consistent with isolation by distance, with an east/west gradient that has been interpreted as evidence for post-glaciation colonization [39]. In North America there is less spatial structure, genome-wide linkage disequilibrium and haplotype sharing, likely due to recent human introduction from Europe [39].

We analyze a large geo-referenced dataset from the Regional Mapping (RegMap) project [41]. The data include 980 accessions from Europe and 180 accessions from North America, genotyped at 220K SNPs.

In a combined analysis of the North American and European data (Fig. 6a), the EEMS shows a corridor of high effective migration across the Atlantic ocean, relative to lower effective migration within each continental group. This highlights the fact that the European and North American samples are more genetically similar than their distance would suggest under a simple isolation by distance scenario. Although the EEMS model assumes that migration is symmetric, and so it cannot infer a direction for gene flow, the EEMS is consistent with the hypothesis that recent directed migration introduced *A. thaliana* from Europe to North America.

Analyzing the North American samples alone, the EEMS has an area of high migration connecting the two sampled regions, Lake Michigan and the Atlantic coast (Fig. 6b). This indicates that samples from these two regions are similar genetically even though they are distant geographically – probably again due to human-assisted long-range "migration" rather than natural dispersal, and consistent with the observation that there is extensive haplotype sharing not only within but also between sampling locations [39].

For the European samples the EEMS highlights a number of regions of higher and lower effective migration (Fig. 6c). For example, in the British Isles, a region of lower migration separates the N British Isles from the rest of Britain, which in turn is connected to NW France by an area of high migration. (In PCA analysis in [41] most accessions from the British Isles are projected closest to France.) The lower effective migration area in Central and Northern France highlights the fact that the NW France samples are more similar to British samples than they are to other French samples a similar distance away. An area of higher effective migration covers much of Germany, and links it with E France in the South, and to parts of Norway and Sweden in the North. In contrast, Spain and Italy show substantially lower effective migration rates, suggesting that genetic similarity decays more quickly with distance in these areas than in other parts of Europe.

Discussion

We have developed EEMS (Estimated Effective Migration Surfaces), a new method for analyzing population structure from geo-referenced genetic samples. Our method produces an intuitive visual representation of the underlying spatial structure in genetic variation, which highlights potential regions of higher-than-average and lower-than-average historic gene flow. EEMS is specifically applicable to data that conform roughly to "isolation by distance" (IBD), i.e., in settings where genetic similarity tends to decay with geographic distance,

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

but where this decay with distance may occur more quickly in some regions than in others.

EEMS uses the concept of "isolation by resistance" (IBR), which aims to model how genetic differentiation accumulates in non-homogeneous landscapes [25]. IBR integrates over all possible migration paths between two points, providing a computationally convenient approximation to the coalescent process in structured populations, and better prediction of genetic differentiation than alternatives such as the least-cost path distance [42].

As originally introduced, IBR is used to build up a resistance map from known landscape/habitat features and thus determine whether resistance distance can improve prediction of observed genetic differentiation over Euclidean or least-cost path distances [25, 43]. More recently, IBR has been incorporated into a formal inference procedure designed to test whether resistance distances are correlated with specific, known landscape features such as altitude or river barriers [26]. In contrast, EEMS estimates an effective migration surface from genetic data without requiring observation of underlying landscape variables, providing an exploratory tool for spatial population structure. We view the hypothesis-driven and exploratory approaches as complementary, and we anticipate that each could be useful in many applications.

EEMS also complements existing methods for detecting barriers to gene flow in continuous spaces, such as Monmonier's maximum difference algorithm [21], Wombling methods [22], and LocalDiff [23]. Monmonier's algorithm analyzes observed genetic differences, but it requires that genetically distinct groups are specified *a priori* before attempting to find 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 [23] 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 and uses isolation by resistance to do so.

Although EEMS is aimed at settings where the underlying population structure is somewhat *continuous*, our method is built on a dense regular grid of *discrete* demes, with migration between neighboring demes. Since the demes do not correspond to *a priori* defined subpopulations, the size and registration of the grid are somewhat arbitrary. The choice of grid may be influenced by factors such as the density of sampling locations (one might want a grid sufficiently dense so that different sampling locations typically correspond to different demes), and computational tractability (computation scales cubically with number of demes in the grid). Although we have found that in practice results are qualitatively robust across a range of grids (Supplementary Fig. 1), some details can change with the specific grid, and so we suggest reducing the potential influence of the grid choice by averaging results over several different grids. In principle, it would seem attractive to dispense with the grid altogether, and use models of continuous migration. However, such models present theoretical and mathematical challenges [44], and at present we do not know how to achieve this. Furthermore, EEMS assumes that migration rates are symmetric, and therefore, it cannot detect anisotropic migration, i.e., different migration rates in the two opposite directions of the same edge.

In addition to estimating effective migration rates, our model also fits a deme-specific parameter, q, that describes the within-deme genetic diversity. For our purpose to characterize dissimilarities between demes,

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

q is a nuisance parameter, but in some applications spatial variation in diversity may be of interest in itself, and so visualizing the effective diversity rates may be useful for some analyses. In our examples, visualizing q highlights the previously noted north-south gradient in diversity in human Europeans (Supplementary Fig. 9b, [37, 45]), higher diversity in ethnic groups from East Africa compared to groups from South or West Africa (Supplementary Fig. 11b), and higher diversity in forest versus savanna elephants (Supplementary Fig. 4b, [34]).

EEMS requires that each sample has a specified geographic origin, but in some cases this information may be known only imprecisely. For example, in our analysis of human genetic variation in Europe, we use the coordinates previously described in [17], where samples are assigned to the geographic center of their ancestral country of origin, with the exception of samples from the Russian Federation, Sweden and Norway, which are assigned to the capital. To better deal with imprecision in geographic origin, the uncertainty could be incorporated into the model, as in [12]: the actual location of each individual will be treated as an unobserved latent variable, given a prior distribution, and integrated out in the MCMC estimation scheme. This approach might also improve robustness to data errors such as sample switches, and identify individuals whose genetic origins differ appreciably from their physical sampling locations. A similar extension could also help address the "spatial assignment problem" in non-stationary isolation by distance settings, i.e., attempt to infer the origin of individuals with unknown location, given reference samples with known location [30, 46].

Since the primary output of EEMS is a visual display of spatial patterns, we have paid attention to the details of this display. For example, we have selected a color scheme that is colorblind friendly [47], and which is "balanced" with respect to high vs low migration. (For example, we have attempted to give similar visual prominence to regions that are 10 times higher than average in their effective migration as to regions that are 10 times lower than average.) We have chosen the scale so that small differences in effective migration rates – say, less than a factor of two – tend not to be emphasized. These choices could undoubtedly be improved upon with further experimentation, and indeed any given scale or color scheme may work better for some datasets than for others. Users of the EEMS method may therefore wish to experiment with display settings, and should be aware of the impact that display parameters can have on the message conveyed by an EEMS.

Although we have not emphasized it here, in addition to a point estimate of the effective migration surface, EEMS fits a posterior distribution for the migration parameters at each location. Given that our modeling assumptions are simplistic we would not advocate interpreting these posterior distributions too literally; nonetheless, they may provide a useful assessment of the uncertainty in the estimated surface. It could be helpful to incorporate this information into the visual display: for example, shading the EEMS only in regions where the posterior 95% credible interval excludes 0 could reduce the danger of over-interpreting patterns that could have easily arisen by chance.

Like PCA, EEMS works directly with a matrix that summarizes the (dis)similarities between all pairs of samples, by averaging across genotyped markers. This makes it computationally tractable for large numbers of SNPs: once this matrix is computed, the complexity per MCMC iteration does not depend on the number of markers. Moreover, like PCA, EEMS is widely applicable: it can be applied to visualize dissimilarities between geo-referenced samples that have been computed from non-genetic features, e.g., language. (The

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

results from EEMS will likely only be useful in settings where similarity tends to decay with geographic distance, but this is easily assessed.) Summarizing the data by a pairwise distance matrix does, however, result in some loss of information. In particular, as highlighted in Figure 3, it limits the demographic scenarios that can be distinguished from the data; see [19] for detailed discussion. It may be fruitful to explore the use of other dissimilarity matrices to emphasize different aspects of the data, perhaps different historical timescales. For example, ChromoPainter [48] produces a measure of genetic similarity ("number of chunks copied"), which will tend to emphasize the most recent coalescent events between samples (rather than the average coalescence times). Similarly, distance matrices based on rarer SNPs should tend to emphasize more recent dispersal history [49]. It is possible that the pairwise sequentially Markov coalescent [50], extended to deal with pairs of individuals [51], could also provide useful alternative distance matrices which could be visualized as an EEMS.

Software implementing the EEMS method is available at `http://www.github.com/dipetkov/eems`.

Acknowledgements

This work was supported in part by US National Institutes of Health (NIH) grants R01HG007089 and R01GM108805 to J.N. and grant HG02585 to M.S. We thank Samuel Wasser for access to the elephant data analyzed in [31] and Ida Moltke for compiling the human dataset from Sub-Saharan Africa as described in [36]. We also acknowledge Brad McRae for helpful early discussions on resistance distances.

Online methods

EEMS uses a population genetic model that involves migration on an undirected graph $G = (V, E)$, with vertices (demes) V, connected by edges E. We fix the graph G to be a regular triangular grid embedded in a two-dimensional plane, so that each deme has known location and only neighboring demes are directly connected (Figure 1b). The density of the grid is pre-specified by the user, and depends on both computational considerations – computational complexity scales cubically with number of vertices – and the resolution of the available spatial data.

The EEMS model has parameters (m, q), where $m = \{m_e : e \in E\}$ specifies an effective migration rate on every edge in E and $q = \{q_v : v \in V\}$ specifies an effective diversity for every deme in V. Intuitively, the migration rates m characterize the genetic dissimilarities between distinct demes, while the diversities q characterize the genetic dissimilarities between distinct individuals from the same location. The EEMS model is a particular application of the more general stepping stone model [27], which allows directed migration as well as migration between demes that are not located close in space.

We use Bayesian inference to estimate the EEMS parameters (m, q). The key components to this inference are i) the *likelihood* $\ell(m, q)$, which measures how well the parameters explain the observed data; and ii) the *prior distribution* $p(m, q)$, which captures the expectation of a spatial structure in the parameters m and q: for example, it captures the idea that the migration rates on adjacent edges of the graph are often similar. The

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

following sections describe the likelihood and the prior in detail.

The likelihood

We first specify the likelihood for SNP data (on n individuals at p SNPs), and then extend to microsatellites. The key initial step is to summarize the observed genetic data by the matrix of average genetic differences between every pair of sampled individuals, D (defined precisely below). This approach – using D as a sufficient statistic for the population parameters – is motivated by the assumption that D contains most of the information about m and q. This may not be completely true, but the idea of performing inference using pairwise genetic similarities or dissimilarities has a long history in both population genetics and phylogenetics [52, 53, 54], and many existing methods are based on a similar assumption. For example, PCA [14] and TreeMix [55] both work with the genetic covariance matrix.

Therefore, let D_{ij} denote the observed genetic dissimilarity between individuals i and j. The expected value of D_{ij} is determined, up to a constant of proportionality that reflects the mutation rate, by how closely related i and j are, or more precisely, by the expected coalescence time between their gametes. The pairwise expected coalescence time in turn depends on the sampling locations $\delta(i)$ and $\delta(j)$, and on the population parameters (m, q): individuals sampled from demes that are connected by many short paths containing edges with high migration rates will tend to be more closely related, and hence more similar genetically, than individuals sampled from demes connected only by paths that are long and/or contain edges with low migration rates. The expected coalescence times can be computed, at some computational expense, by solving a large set of simultaneous equations; alternatively, they can be approximated – at less, but still nontrivial computational cost – using the concept of "resistance distance" [25]. We implemented both approaches, and found them to produce qualitatively similar EEMSs, and so all results presented here were obtained using the resistance distance.

Letting σ^2 denote the constant of proportionality mentioned above, we can write:

$$\mathrm{E}\{D \,|\, m, q, \sigma^2\} \;=\; \sigma^2 \Delta(m, q), \tag{1}$$

where $\Delta(m, q)$ is a matrix of expected dissimilarities that can be computed for any (m, q). Our modelling approach, detailed below, assigns high likelihood to values for (m, q, σ^2) such that $\sigma^2\Delta(m, q) \approx D$, while taking some account of dependencies among elements of D and of linkage disequilibrium among markers.

To make our specification precise we introduce the following notation:

- Z denotes the $n \times p$ matrix of genotypes: Z_{il} is the genotype of individual i at locus l, which is coded as 0, 1 or 2 copies of the minor allele.
- D denotes the $n \times n$ matrix of average genetic differences between individuals: $D_{ij} := (1/p) \sum_{l=1}^{p} (Z_{il} - Z_{jl})^2$.
- L is the $(n-1) \times n$ matrix such that $L_i = e_i - e_{i+1}$ where L_i is the ith row of L and e_i is a row vector with 1 in the ith component and 0s elsewhere.
- $W := -LDL'$, which is an $(n-1) \times (n-1)$ symmetric, positive definite matrix, as $W = -LDL' = 2(LZ)(LZ)'$ is a quadratic form [26].

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The matrix L is chosen because it forms a basis for the space of contrasts on n elements: for example, $e_i - e_{i+1}$ is a contrast between the ith and $(i+1)$st elements. Since L is a basis, W is a one-to-one mapping of D and we can specify a model for D by specifying a model for W [26]. (It turns out that, for technical reasons, it is easier to work with W than with D directly [56].) Furthermore, since the statistic W is positive definite, a natural model for W is the Wishart distribution, which can be parametrized by the expectation, $\mathrm{E}\{W\}$, and a scalar degrees of freedom parameter, k. Using equation (1), the expectation is given by:

$$\mathrm{E}\left\{W \,|\, m, q, \sigma^2\right\} \;=\; -L\mathrm{E}\left\{D \,|\, m, q, \sigma^2\right\}L' \;=\; -\sigma^2 L\Delta(m,q)L'. \tag{2}$$

We treat the degrees of freedom k as an additional free parameter to be estimated.

Putting this together yields a closed form for the density of W and thus – for the likelihood of the parameters k, m, q, σ^2 since $\ell(k, m, q, \sigma^2) := p(W \,|\, k, m, q, \sigma^2)$. Specifically:

$$W \,|\, k, m, q, \sigma^2 \;\sim\; \mathrm{W}_{n-1}\left(k, -\frac{\sigma^2}{k} L\Delta(m,q)L'\right). \tag{3}$$

We make the following observations:

1. Although we defined Z_{ij} as the number of copies of the minor allele, the likelihood does not depend on the labeling of the alleles because the differences $(Z_{il} - Z_{jl})^2$ do not depend on the labelling.

2. If the genotypes Z were independent across loci and normally distributed, then standard Gaussian theory would imply that W has a Wishart distribution, with the degrees of freedom k equal to the number of SNPs p. However, genotypes are neither normal nor independent, and rather than fix $k = p$ (as in [26]), we estimate the degrees of freedom k, under the assumption $n \leq k \leq p$. The smaller k is, the higher the variance of W about its expectation (which does not depend on k in our parametrization). By allowing $k < p$ we can, to some extent, account for sources of model mis-specification such as linkage disequilibrium between loci.

3. In defining $W := -LDL'$ we took L to be a specific matrix. However, other choices for L would yield equivalent likelihoods as long as the $(n-1)$ rows of L form a basis for the contrasts of n elements. (A contrast is a linear combination whose coefficients add to 0.) This key property ensures that W is a one-to-one mapping of D, and therefore we would get exactly the same likelihood with any basis L (up to a constant of proportionality that does not depend on the parameters). In other words, the introduction of a specific transformation W should be regarded as a technical trick to facilitate the development of a model for the genetic differences D; see [56] for a more mathematical discussion.

Application to microsatellites

At microsatellite loci an allele is typically coded as the number of repeats of a specific motif. To apply our method to microsatellites we define the genotype Z_{il} to be the *average* of the two alleles individual i carries at locus l. (This approach could likely be improved upon, but it suffices for our analysis of the African elephant data.)

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We then define $D^{(l)}$ as the matrix of pairwise differences at locus l, and $W^{(l)}$ as the corresponding W matrix:

$$D_{ij}^{(l)} \;:=\; (Z_{il} - Z_{jl})^2, \tag{4}$$

$$W^{(l)} \;:=\; -LD^{(l)}L'. \tag{5}$$

Since different microsatellite loci have different mutation rates, we introduce locus-specific scale parameters $\sigma_l^2 : l = 1, \ldots, p$. For locus l equation (1) becomes:

$$\mathrm{E}\{D^{(l)} \,|\, m, q, \sigma_l^2\} \;=\; \sigma_l^2 \Delta(m, q). \tag{6}$$

And we define the likelihood by assuming that each $W^{(l)}$ has an independent Wishart distribution, with one degree of freedom:

$$W_l \,|\, m, q, \sigma_l^2 \;\overset{\text{ind}}{\sim}\; \mathrm{W}_{n-1}\Big(1, -\sigma_l^2 L\Delta(m, q)L'\Big). \tag{7}$$

The rank-1 matrices W_l are not positive definite but we can nevertheless compute the likelihood for the parameters m, q and σ_l^2. (Since the p microsatellite loci are assumed to be independent, the degrees of freedom are effectively fixed to $k = p$.)

The dissimilarity matrix $\Delta(m, q)$

In population genetics, the expected genetic dissimilarity between two samples is a function of their expected coalescence time. Indeed, for haploid samples at SNP markers, as the mutation rate tends to 0, it can be shown that (see Supplementary Methods and [19])

$$\mathrm{E}\Big\{D_{ij} \,|\, m, q\Big\} \;\propto\; T_{\delta(i)\delta(j)}(m, q), \tag{8}$$

where $\delta(i)$ denotes the deme from which sample i is drawn, and $T_{\alpha\beta}(m, q)$ is the expected coalescence time of two independent haploid samples taken from demes α and β, respectively. Thus in equation (1) we have:

$$\Delta_{ij}(m, q) \;=\; T_{\delta(i)\delta(j)}(m, q). \tag{9}$$

Similarly, for diploid samples, we have (see Supplementary Methods):

$$\Delta_{ij}(m, q) \;=\; 4T_{\delta(i)\delta(j)}(m, q) - T_{\delta(i)\delta(i)}(m, q) - T_{\delta(j)\delta(j)}(m, q). \tag{10}$$

For any particular value of the parameters (m, q), the matrix $T(m, q)$ – and hence $\Delta(m, q)$ – can be computed by solving a system of linear equations [57, 58]. (To obtain $T(m, q)$ in the stepping stone model, the parameters (m, q) are the between-deme migration rates m^* and the within-deme coalescence rates q^*, where we use the superscript $*$ to indicate that the EEMS parameters have slightly different interpretation from those in [57, 58].)

However, computing T is expensive because we have to solve a linear system with $d(d+1)/2$ unknowns to find all pairwise expected coalescence times in a graph with d demes – this has complexity $O((d^2)^3)$. To reduce the computational cost, for all results presented here, we approximate coalescence times using the idea of "effective resistances" – a distance metric for weighted undirected graphs [59]. Computing R is less intensive

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

because we can obtain all pairwise resistance distances by inverting a $d \times d$ matrix [60] – this has complexity $O(d^3)$. (Efficiency can be improved further by computing the subset of resistance distances between sampled demes only; see Supplementary Methods.)

To approximate T in terms of R, let $R_{\alpha\beta}(m)$ denote the resistance distance between demes α and β in the graph G. (Note that $R_{\alpha\beta}$ is not a function of only the *local* migration rate $m_{\alpha\beta}$, but is determined by the *global* migration pattern m.) The effective resistances R are approximately related to the expected coalescence times T through [25]:

$$R_{\alpha\beta}/4 \;\approx\; T_{\alpha\beta} - \left(T_{\alpha\alpha} + T_{\beta\beta}\right)/2. \tag{11}$$

This approximation is exact for isotropic migration (i.e., if the demes are equivalent with respect to the rate and pattern of migration), and for more general migration models the approximation gets better as the migration rates increase [25]. Using equation (11) we approximate the expected coalescence time between two haploid samples from demes α and β as:

$$T_{\alpha\beta} \;=\; \underbrace{T_{\alpha\beta} - \left(T_{\alpha\alpha} + T_{\beta\beta}\right)/2} + \underbrace{\left(T_{\alpha\alpha} + T_{\beta\beta}\right)/2} \;\approx\; R_{\alpha\beta}/4 + \left(q_\alpha + q_\beta\right)/2. \tag{12}$$

That is, we split the expected coalescence times, $T_{\alpha\beta}$, into a between-demes component, which is approximated by the effective resistances $R_{\alpha\beta}$, and a within-deme component, which is determined by the diversity rates q. The effective resistances $R_{\alpha\beta}$ depend on m, and the vector q is treated as a free parameter to be estimated. We then obtain $\Delta(m, q)$ by substituting $T(m, q)$ with its approximation according to equation (12).

Generally, we have found that the approximation based on the effective resistances R produces visualizations comparable with those obtained using the more computationally expensive coalescence times T.

Prior Distributions

The Voronoi prior on migration rates

Our prior for the migration rates m captures the idea that nearby edges will tend to have similar rates, while it also allows the rates to vary among edges. We parametrize the prior $p(m)$ using a Voronoi tessellation of the two-dimensional habitat H, which partitions H into C convex polygons (cells) as follows. First select C distinct points (seeds) $s_1, \ldots, s_C \in H$. Then define cell c to be the set of points in H that are closer to seed s_c than to any other seed. Given a Voronoi tessellation of H, we associate with each cell a migration rate m_c. We use these to induce a migration rate on each edge in the graph G, with the migration rate of the edge joining demes α and β given by:

$$m_{\alpha\beta} \;=\; \left(m_{c_\alpha} + m_{c_\beta}\right)/2, \tag{13}$$

where c_α denotes the cell that contains deme α. Migration rates are naturally positive and therefore we parametrize the m_c values on the $\log_{10}$ scale. Further, to capture the idea that the migration rates of different cells may be similar to one another we parametrize them as deviations from an overall mean rate μ:

$$\log_{10}(m_c) \;=\; \mu + e_c, \tag{14}$$

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where e_c denotes the "effect" of cell c and determines whether the local dispersal in cell c is faster or slower than the average.

In this formulation, migration rates on every edge in the graph are determined by the parameters $(C, s_1, \ldots, s_C, e_1, \ldots, e_C, \mu)$. To complete the Bayesian specification we place priors on the model parameters:

$$\text{(number of Voronoi cells)} \qquad C \,|\, a, b \;\sim\; \text{Neg-Bi}(r,\, u), \tag{15a}$$

$$\text{(cell locations)} \qquad s_1, \ldots, s_C \,|\, C \;\overset{\text{iid}}{\sim}\; \text{U}(H), \tag{15b}$$

$$\text{(cell effects)} \qquad e_1, \ldots, e_C \,|\, \omega_m^2, C \;\overset{\text{iid}}{\sim}\; \text{N}_{[-2,+2]}(0,\, \omega_m^2), \tag{15c}$$

$$\text{(overall migration rate)} \qquad \mu \;\sim\; \text{U}(-2.477, +2.477), \tag{15d}$$

$$\text{(variance between cells)} \qquad \omega_m^2 \;\sim\; \text{Inv-G}(c_\omega/2,\, d_\omega/2). \tag{15e}$$

Here $\text{U}(H)$ denotes the uniform distribution with support the habitat H; $\text{N}_{[a,b]}(\mu, \sigma^2)$ denotes the truncated normal distribution with mean μ, variance σ^2 and support $[a, b]$; $\text{Inv-G}(c, d)$ is an inverse gamma distribution with shape c and rate d; and $\text{Neg-Bi}(r, u)$ denotes a zero-truncated negative binomial distribution with shape (number of failures) r and probability of success u. [The zero-truncated negative binomial has support $\{1, 2, 3, \ldots\}$; we truncate the support at zero because the Voronoi tessellation should have at least one cell.] For all results reported here we have used $r = 10, u = 2/3, c_\omega = 0.001, d_\omega = 1$.

The prior on the number of Voronoi cells C was chosen because the negative binomial is an overdispersed Poisson (a continuous mixture of Poisson distributions): Suppose that C is a Poisson random variable with mean λ, and λ is a Gamma random variable with shape r and rate $u/(1-u)$. If we integrate λ, we obtain $C \sim \text{Neg-Bi}(r, u)$. For all analyses described here, we have used $r = 10$ and $u = 2/3$, which results in a diffuse prior on C, with prior mean 20 and prior variance 60.

The lower and upper bounds on the mean log migration rate μ are chosen so that on the original scale the mean migration rate varies in the range $[1/300, 300]$. The bounds are somewhat arbitrary, and chosen to reflect values that might be considered "very small" (approaching the limit of discrete demes evolving independently) and "very large" (approaching a panmictic population). The cell effects $e_1, \ldots, e_C$ are constrained to lie in the range $[-2, +2]$, so that the migration rate of a cell can vary within a factor of 100 from the mean migration rate.

Other priors

If there are more genotyped markers than sampled individuals, we can estimate the degrees of freedom k. The prior on k is uniform on the $\log_{10}$ scale, to reflect our uncertainty about the order of magnitude of this parameter:

$$p(k) \;\propto\; \frac{1}{k}. \tag{16}$$

The prior is proper because k is bounded: $n \leq k \leq p$ where n is the number of samples and p is the number of SNPs.

The prior on the Wishart scale parameter σ^2 is:

$$\sigma^2 \;\sim\; \text{Inv-G}(c_\sigma/2,\, d_\sigma/2). \tag{17}$$

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In all results presented here we have used $c_\sigma = 0.001$, $d_\sigma = 1$. For both variance scale parameters, σ^2 and ω_m^2, we have chosen the inverse gamma shape and rate hyperparameters, c and d, so that the prior distributions are weakly informative.

Markov Chain Monte Carlo estimation

We use Markov Chain Monte Carlo (MCMC) to estimate the EEMS parameters by sampling from their posterior distribution given the observed genetic differences D. The two Voronoi tessellations – which describe the spatial organization of the migration rates and the diversity rates, respectively – are independent of each other and are estimated with birth-death MCMC because the number of Voronoi cells is unknown. That is, every move proposes to either add a new cell or remove one of the existing cells (if there are at least two). The location and rate parameter of each cell are updated in turn with a random-walk Metropolis-Hastings step. (A cell in the migration Voronoi has an effective migration rate parameter, a cell in the diversity Voronoi has an effective diversity rate parameter.) See Supplementary Methods for further details about the computational methods implemented in EEMS.

Empirical datasets

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 [30, 61]. 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 [30]. We analyze the reference data from [31], which excludes putative hybrids and consists of 211 forest and 913 savanna elephants.

Human European and African data. The European dataset was collected and genotyped as part of the POPRES (Population Reference Sample) project [35] 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 1379 individuals analyzed in a previous publication [17], with the individual IDs and marker list available from `https://github.com/jnovembre/Novembre_etal_2008_misc`. We removed six samples identified as possible outliers based on their position in PC1-PC2 space: the single sample from Slovakia (13011) that might have Italian rather than Slovakian ancestry, and five samples from Italy (7623, 33242, 34049, 38532, 49500) that project outside of the main Italian cluster and thus might have insular Italian ancestry (Sardinian or Sicilian) [17]. In Figure 5a, we use the population abbreviations used in [17], which generally correspond to ISO country codes. (The samples from Switzerland (CH) are split into three subpopulations: French, Italian and German speaking Swiss, coded as CHf, CHi and CHg, respectively.)

The African dataset was compiled from a subset of two published SNP array datasets; one presented in [62] (and available from `http://jorde-lab.genetics.utah.edu/?page_id=23`), and the other presented in [63] (and available from `http://www-evo.stanford.edu/repository/`). From the Xing *et al.* dataset we extracted the

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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), Xhosa (Xh) and Yoruba (Yo) as well as the Luhya (Lu) and Maasai (Ma1) samples, in which no pair is more closely related than first cousins [64]. 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 2 Hema individuals that were classified as likely relatives and outliers in most analyses of Sub-Saharan samples in [36]. After these exclusions, we analyzed a dataset composed of 314 samples from 21 Sub-Saharan populations genotyped at 27,825 polymorphic SNPs.

***Arabidopsis thaliana* data**. The *Arabidopsis thaliana* dataset was collected and genotyped as part of the RegMap (Regional Mapping) project [41] and is available at `http://bergelson.uchicago.edu/regmap-data/`. We downloaded unimputed SNP genotypes for 1193 samples with high-quality geographic coordinates (latitude and longitude), categorized into 12 geographic regions. From these we analyzed 1160 accessions from North America and Europe, genotyped at 214,051 SNPs using the Affymetrix Arabidopsis 250K SNP chip [41]. These include 180 accessions from the region *Americas* and 980 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.

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).

[2] Reich, D., Thangaraj, K., Patterson, N. J., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).

[3] Beaumont, M. A. & Balding, D. J. Identifying adaptive genetic divergence among populations from genome scans. *Proc. Natl. Acad. Sci. U.S.A.* **13**, 969–980 (2004).

[4] Becquet, C. & Przeworski, M. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* **17**, 1505–1519 (2007).

[5] Teeter, K. C. *et al.* Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res.* **18**, 67–76 (2008).

[6] Kronforst, M. R., Young, L. G., Blume, L. M. & Gilbert, L. E. Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. *Evolution* **60**, 1254–68 (2006).

[7] Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170–175 (2011).

[8] Gonder, M. K. *et al.* Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4766–4771 (2011).

[9] Campbell, C. D. *et al.* Demonstrating stratification in a European American population. *Nat. Genet.* **37**, 868–72 (2005).

[10] Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).

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

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

[13] Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–64 (2009).

[14] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

[15] Patterson, N. J., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[16] Rousset, F. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219–1228 (1997).

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

[18] Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40**, 646–649 (2008).

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

[20] DeGiorgio, M. & Rosenberg, N. A. Geographic sampling scheme as a determinant of the major axis of genetic variation in principal components analysis. *Mol. Biol. Evol.* **30**, 480–488 (2013).

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

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

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

[24] Beerli, P. & Felsenstein, J. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4563–4568 (2001).

[25] McRae, B. H. Isolation by resistance. *Evolution* **60**, 1551–1561 (2006).

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

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

[28] Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).

[29] O'Kane, S. L. & Al-Shehbaz, I. A. A synopsis of *Arabidopsis* (Brassicaceae). *Novon* **7**, 323–327 (1997).

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

[31] Wasser, S. K. *et al.* Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots. *Submitted* (2014).

[32] Beaumont, M. A. & Nichols, R. A. Evaluating loci for use in the genetic analysis of population structure. *Proc. Biol. Sci.* **263**, 1471–2954 (1996).

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[33] Georgiadis, N. *et al.* Structure and history of African elephant populations: I. Eastern and southern Africa. *J Hered.* **85**, 100–104 (1994).

[34] Comstock, K. E. *et al.* Patterns of molecular genetic variation among African elephant populations. *Molecular Ecology* **11**, 2489–2498 (2002).

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

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

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

[38] Tian, C. *et al.* Analysis and application of European genetic substructure using 300K SNP information. *PLoS Genet.* **4**, e4 (2008).

[39] Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, e196 (2005).

[40] Platt, A. *et al.* The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000843 (2010).

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

[42] McRae, B. H. & Beier, P. Circuit theory predicts gene flow in plant and animal populations. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19885–90 (2007).

[43] McRae, B. H., Dickson, B. G., Keitt, T. H. & Shah, V. B. Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology* **89**, 2712–2742 (2008).

[44] Felsenstein, J. A pain in the torus: Some difficulties with models of isolation by distance. *Am. Nat.* **109**, 359–368 (1975).

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

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

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

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

[49] Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[50] Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).

[51] Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).

[52] Cavalli-Sforza, L. L. & Edwards, A. W. F. Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.* **19**, 233–257 (1967).

[53] Felsenstein, J. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25**, 471–492 (1973).

[54] Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–25 (1987).

[55] Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).

[56] McCullagh, P. Marginal likelihood for distance matrices. *Stat. Sin.* **19**, 631–649 (2009).

[57] Bahlo, M. & Griffiths, R. C. Coalescence time for two genes from a subdivided population. *J. Math. Biol.* **43**, 397–410 (2001).

[58] Hey, J. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theor. Popul. Biol.* **39**, 30–48 (1991).

[59] Klein, D. J. & Randić, M. Resistance distance. *J. Math. Chem.* **12**, 81–95 (1993).

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

[61] Wasser, S. K. *et al.* Combating the illegal trade in African elephant ivory with DNA forensics. *Conserv. Biol.* **22**, 1065–1071 (2008).

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure legends

Figure 1 A schematic overview of the EEMS method, using geo-referenced African elephant data [31] for illustration. **(a-c)** Setting up the grid of demes: **(a)** Samples are collected at known locations across a two-dimensional habitat; here green and orange represent two elephant subspecies – forest and savanna. **(b)** A dense triangular grid is chosen to span the habitat. **(c)** Each sample is matched to the closest deme on the grid. **(d-f)** Estimated Effective Migration Surface (EEMS) analysis: **d**) Migration rates are allowed to vary according to a Voronoi tessellation which partitions the habitat into "cells" with constant migration rate; colors represent relative rates of migration in each cell, ranging from low (orange) to high (blue). **(e)** The migration rate of an edge is determined by the cell this edge falls into. The cell locations and migration rates are then adjusted, using a Bayesian inference scheme, so that the expected genetic dissimilarities under the EEMS model closely match the observed genetic dissimilarities. **(f)** An EEMS is produced by averaging samples from the posterior distribution on the migration rates, interpolating between grid points. Here, and in all other figures, e_m denotes the effective migration rate, on the $\log_{10}$ scale, relative to the overall migration rate across the habitat. (Thus $e_m = 1$ corresponds to an effective migration rate that is 10-fold larger than the average.) The main feature of the elephant EEMS is a "barrier" of low effective migration that separates the habitats of the two subspecies: forest elephants to the west of the barrier, and savanna elephants to the north, south and east of the barrier.

Figure 2 Simulation results juxtaposing EEMS and PCA analysis. For each method we show results for two migration scenarios, representing "uniform" migration and a "barrier" to migration, and three different sampling schemes. **(a,b)** The true underlying migration rates under the two scenarios; colors represent relative migration rates. **(c)** The three sampling schemes used; the size of the circle at each node is proportional to the number of individuals sampled at that location, and locations are color-coded to facilitate cross-referencing the EEMS and PCA results. **(d)** PCA results. **(e)** EEMS results. In contrast to PCA, EEMS is robust to the sampling scheme and shows clear qualitative differences between results for the different migration scenarios, reflecting the underlying simulation truth.

Figure 3 Simulations results illustrating that EEMS infers effective migration rates, rather than actual steady-state migration rates. **(a)** Individuals have uniform migration rates but the central area has lower population density (those demes have fewer individuals, which is represented by smaller circles in gray), so that a lower number of migrants are exchanged per generation. The difference in population density produces an effective barrier to gene flow, which is reflected in the EEMS. **(b)** A simple "population split" scenario: migration was initially uniform, but at some time in the past a complete barrier to migration arose in the central area (which is represented by dashed edges). Under this scenario the groups on either side of the central region diverge, which creates a barrier in the EEMS.

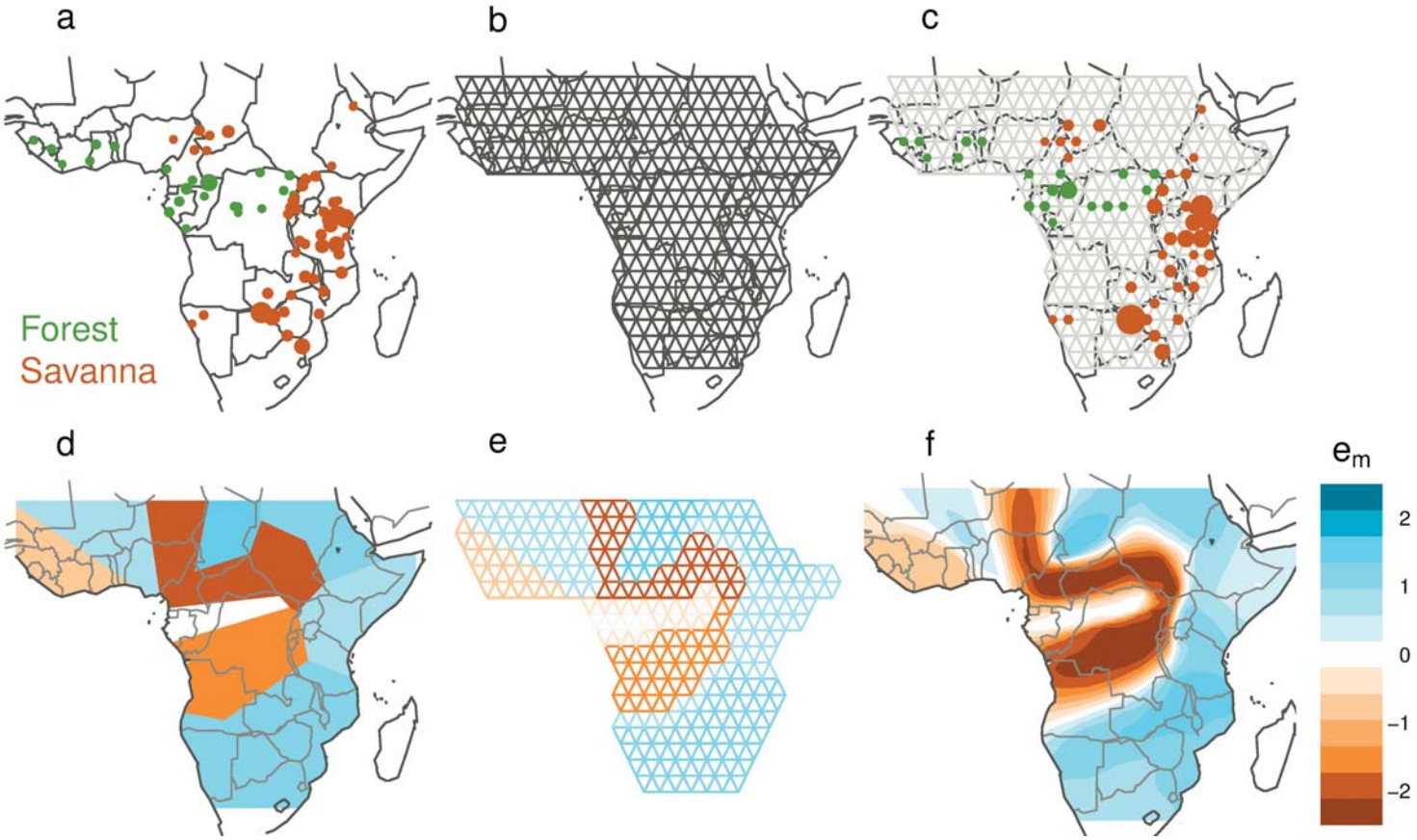
bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4 EEMS analysis of the population structure of the African elephant (data from [31]). **(a)** African elephant samples are collected from two subspecies in five biogeographic regions: the forest elephant subspecies inhabits the west and central regions (in green), the savanna elephant subspecies inhabits the north, east and south regions (in orange). **(b)** EEMS for forest and savanna samples analyzed jointly; **(c)** EEMS for forest samples only; **(d)** EEMS for savanna samples only.

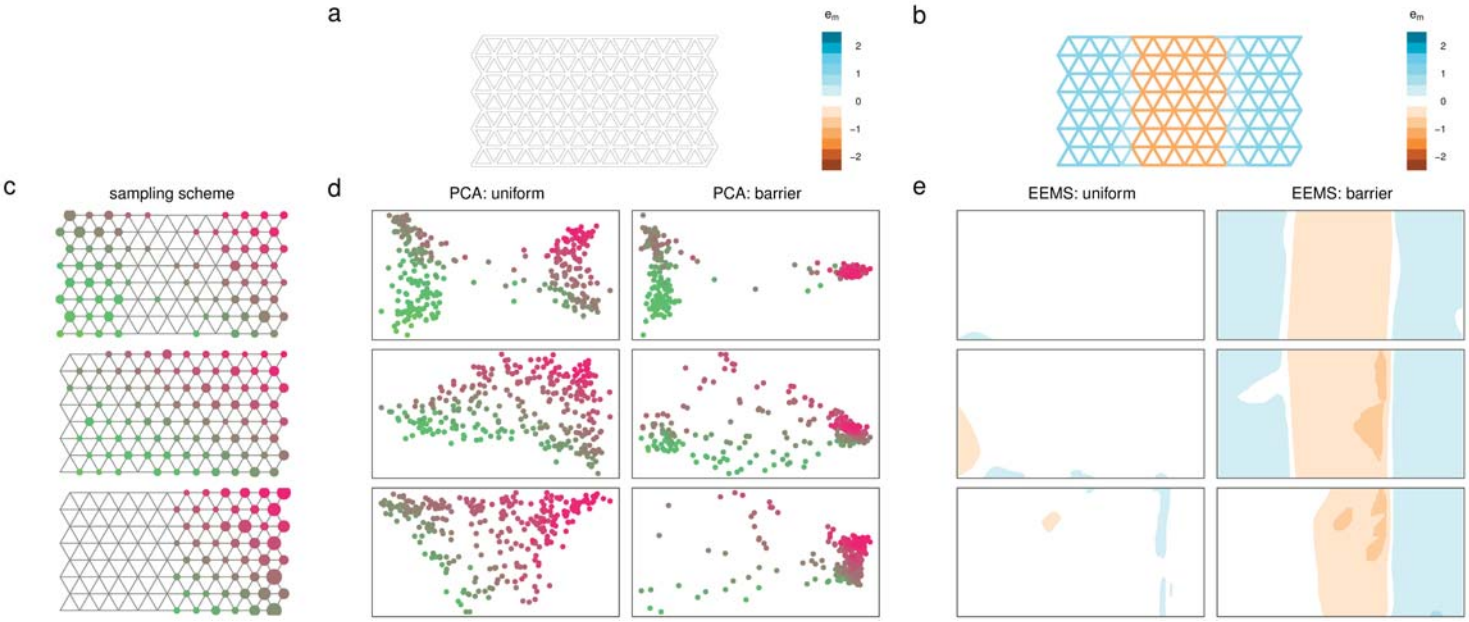
Figure 5 EEMS analysis of human population structure in Western Europe and in Sub-Saharan Africa. **(a)** Estimated effective migration rates in Western Europe (data from the POPRES project [35]). **(b)** Estimated effective migration rates in Sub-Saharan Africa (data from [36]).

Figure 6 EEMS analysis of the population structure of *Arabidopsis thaliana* (data from the RegMap project [41]): **(a)** North America and Europe combined; **(b)** North America; **(c)** Europe.

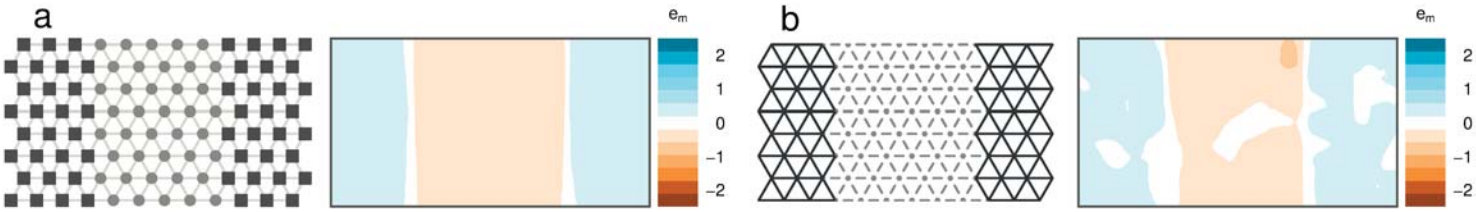
bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



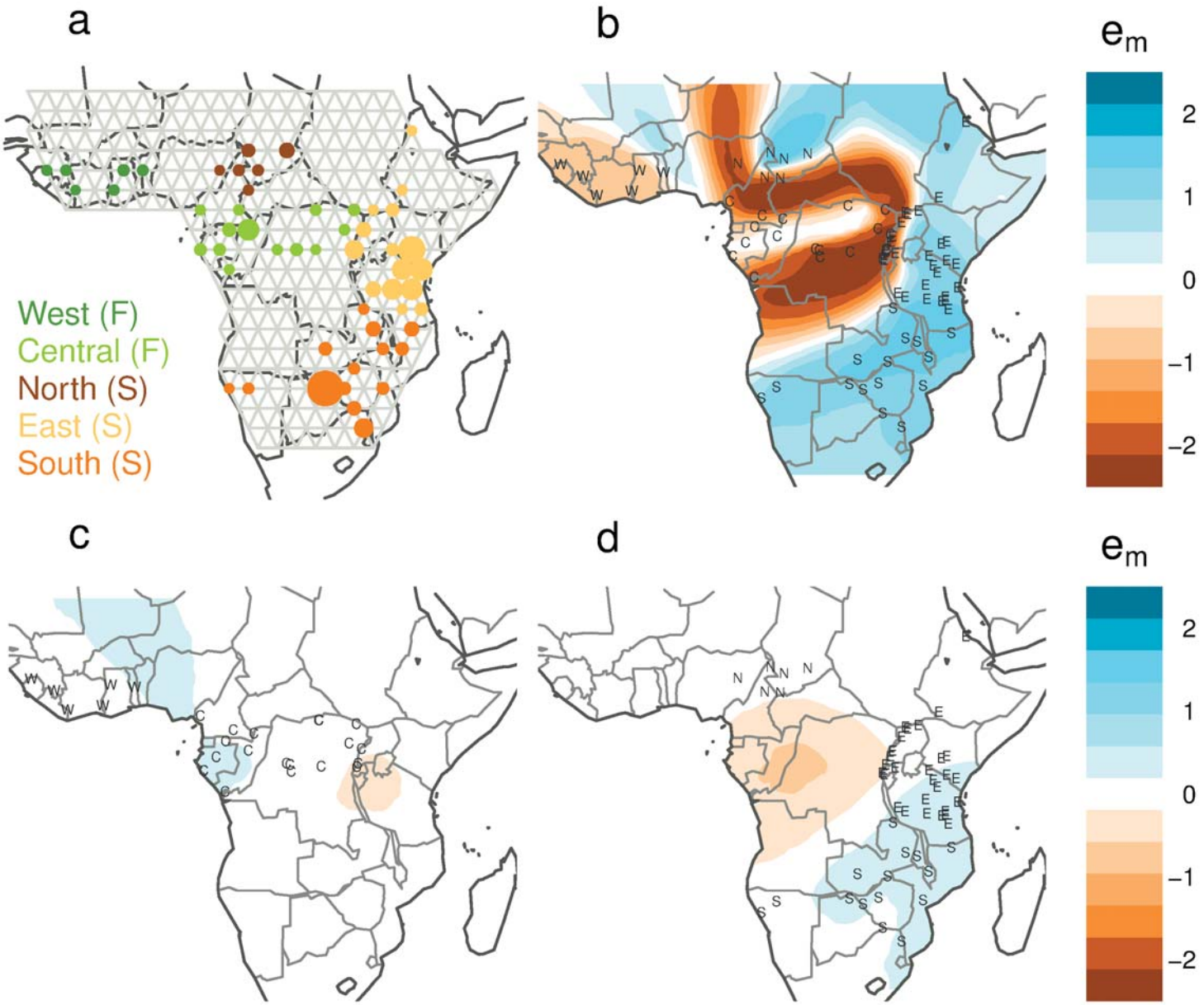
bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



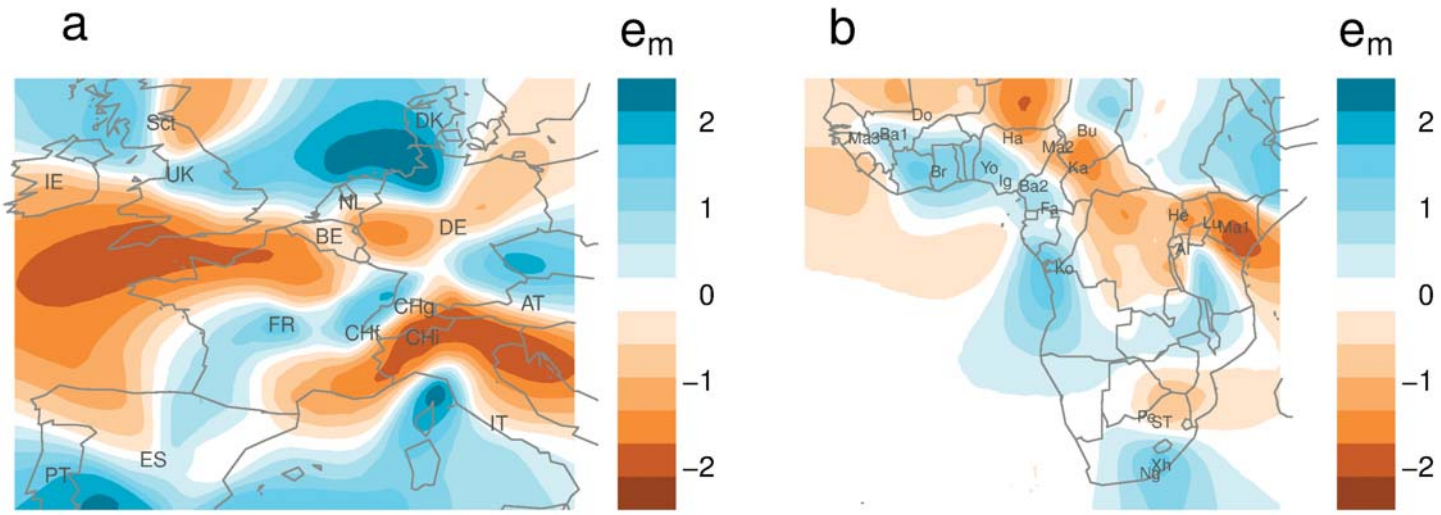
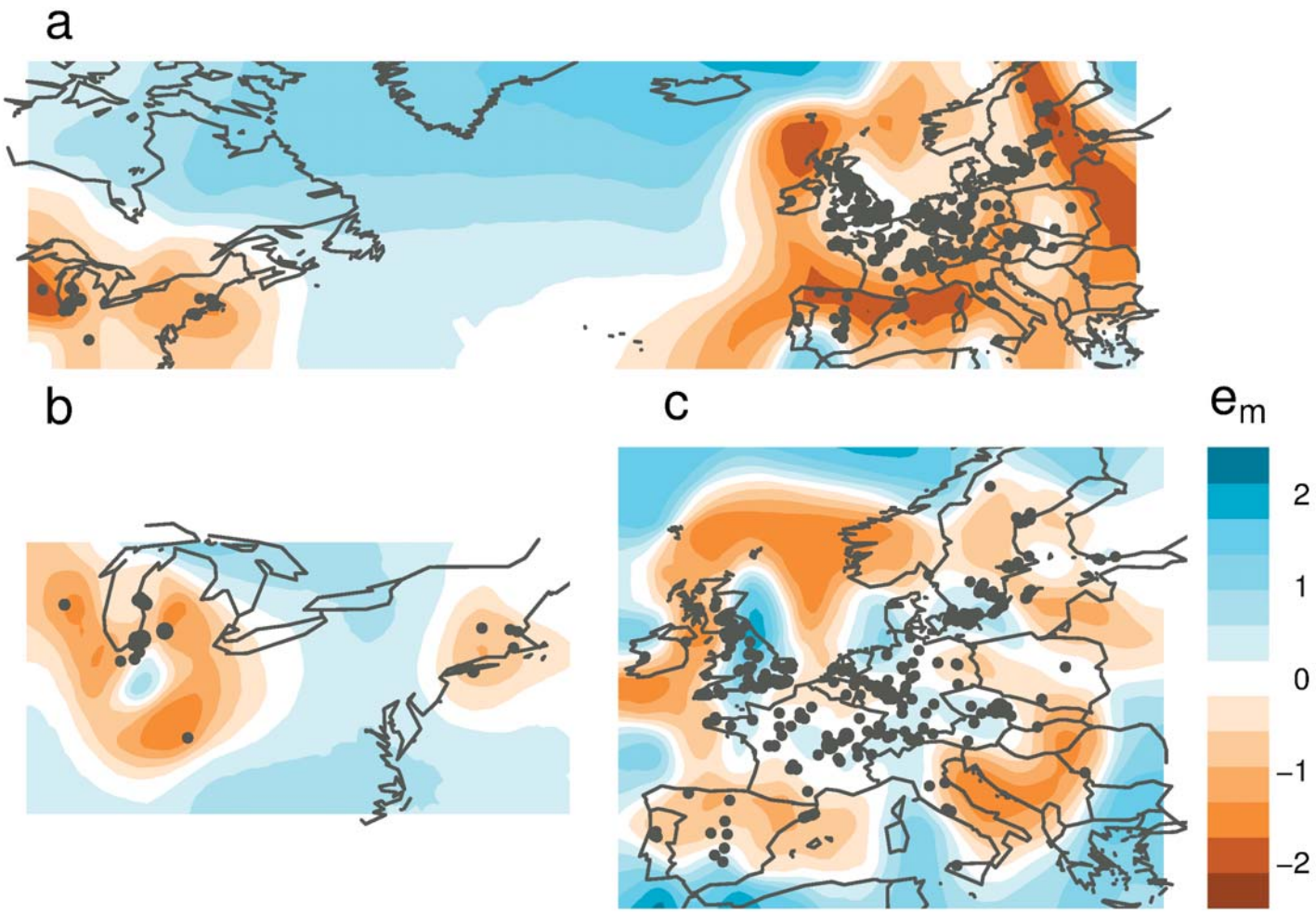
bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/011809; this version posted November 26, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.