# Genetic landscapes reveal how human genetic diversity aligns with geography

Benjamin Marco Peter[1], Desislava Petkova[2,3] & John Novembre[1,4]

[1] Department of Human Genetics, University of Chicago [2] Wellcome Trust Center for Human Genetics, University of Oxford, UK [3]Present Address: Procter & Gamble, Brussels, Belgium [4] Department of Ecology & Evolution, University of Chicago

**Geographic patterns in human genetic diversity carry footprints of population history[1,2] and need to be understood to carry out global biomedicine[3,4]. Summarizing and visually representing these patterns of diversity has been a persistent goal for human geneticists[5–9]. However, most analytical methods to represent population structure[10–14] do not incorporate geography directly, and it must be considered *post hoc* alongside a visual summary. Here, we use a recently developed spatially explicit method to estimate "effective migration" surfaces to visualize how human genetic diversity is geographically structured (the EEMS method[15]). The resulting surfaces are "rugged", which indicates the relationship between genetic and geographic distance is heterogenous and distorted as a rule. Most prominently, topographic and marine features regularly align with increased genetic differentiation (e.g. the Sahara Desert, Mediterranean Sea or Himalaya at large scales; the Adriatic, inter-island straits in near Oceania at smaller scales). We also see traces of historical migrations and boundaries of language families. These results provide visualizations of human genetic diversity that reveal local patterns of differentiation in detail and emphasize that while genetic similarity generally decays with geographic distance, there have regularly been factors that subtly distort the underlying relationship across space observed today. The fine-scale population structure depicted here is relevant to understanding complex processes of human population history and may provide insights for geographic patterning in rare variants and heritable disease risk.**

In many regions of the world, genetic diversity "mirrors" geography in the sense that genetic differentiation increases with geographic distance ("isolation by distance" [16–18]); However, due to the complexities of geography and history, this relationship is not one of constant proportionality. The recently developed analysis method EEMS visualizes how the isolation-by-distance relationship varies across geographic space[15] Specifically, it uses a model based on a local "effective migration" rate. For several reasons, the effective migration rates inferred by EEMS do not directly represent levels of gene flow[15]; however they are useful for conveying spatial population structure: high values of effective migration reflect genetic isolation accrues gradually with distance, and low values imply isolation accrues rapidly with distance. In turn, a map of inferred patterns of effective migration can provide a compact visualization of spatial genetic structure for large, complex samples.

We apply EEMS on a combination of 26 existing single nucleotide polymorphism (SNP) datasets. In total, these comprise 5372 individuals from 348 locations across Eurasia and Africa (Extended Data Table 1), which we organize in six analysis panels: an overview Afro-Eurasian panel (AEA), four continental-scale panels, and a panel of Southern African Hunter-Gatherers. As our focus is on isolation-by-distance patterns, we identified samples that are known to be admixed from distant sources, significantly displaced, and/or from hunter-gatherer groups, as these *a priori* should not fit an isolation-by-distance model. These samples are labelled on the EEMS maps, but were not included in the model fit (see Methods and Table S1). For all analysis panels, the inferred EEMS surfaces are "rugged", with numerous high and low effective migration features (Fig 1a, Fig 2) that are strongly statistically supported when compared to a uniform-migration model (Extended Data Table 2). The regions of depressed effective migration often align in long, connected stretches that are present in more than 95% of MCMC iterations. To facilitate discussion, we annotate these stretches with dashed lines and refer to them as "troughs" of effective migration (Figs. 1a, 2, Extended Data Figs. 2-4). Conversely, intermediate- and high-migration areas between troughs are referred to as corridors.

In the broad overview Afro-Eurasia panel (Fig. 1; n=4,002 samples; 219 locales; $F_{ST}$ = 0.061) we see that troughs often align with topographical obstacles to migration, such as deserts (Sahara), seas (Mediterranean, Red, Black, Caspian, South China Seas) and mountain ranges (Ural, Himalayas, Caucasus). None of these troughs completely surround large regions, as corridors intersperse among them. Among the main features are several large regions that have mostly high effective migration, such as Europe, East Asia, Sub-Saharan Africa and Siberia. Several large-scale corridors are inferred that represent long-range genetic similarity, for example: India is connected by two corridors to Europe (a southern one through Anatolia and Persia 'SC', and a northern one through the Eurasian Steppe 'NC'); East Asia (EA) is connected to Siberia and to southeast Asia and Oceania. The island populations of the Andaman islands (Onge) and New Guinea, as well as the populations of far northeastern Siberia, show troughs nearly contiguously around them – possibly reflecting a history of relative isolation [19–21].

Analyses on a finer geographic scale highlights subtler features (e.g. compare Europe in Fig. 1 vs Fig. 2a). At these finer scales we continue to see troughs that align with landscape features, though increasingly we see troughs and corridors that coincide with historical contact zones of language groups and proposed areas of human migrations. For example, in Europe (Fig. 2b) we observe troughs (NS, CE) roughly between where Northern Slavic speaking peoples currently reside relative to west Germanic speakers, and relative to the linguistically complex Caucasus region. In India (Fig. 2e), troughs demarcate regions with samples of Austroasiatic and Dravidian speakers, as well as central India (CI) relative to Northwestern India (Sindhi, Punjabi) and Pakistan. In Southeast Asia (Fig. 2k), troughs align with several straits in the Malay Archipelago, but we also observe a corridor from Taiwan through Luzon to the Lower Sunda Islands (LSI), and further to Melanesia, perhaps reflecting the Austronesian expansion. In Sub-Saharan Africa (Fig. 2g), we find corridors perhaps reflecting the Bantu expansion from West- into Southern and Eastern Africa, where contact with Nilo-Saharan speakers resulted in complex local structure. In Southern Africa, the structures in Bantu and Khoe-San speakers

(Fig. 2g/h) appear entirely uncorrelated, illustrating that in some cases, different language groups can maintain independent genetic structure in the same geographic region.

We contrast EEMS with principal component analysis (PCA), a widely used, non-spatial method for visualizing population structure. Quantitatively, performance is evaluated by comparing the fit of EEMS and PCA (using the first 10 components) to the observed genetic distances. EEMS performs better for small-scale panels, but PCA provides a better fit on the larger-scale AEA and CEA panels (Extended Data Figure 5). We hypothesize EEMS tends to represent local genetic differences relatively well, and this is supported by an analysis where we stratify the residuals of genetic distances (Fig. 3): In most panels EEMS fits best in the lowest percentiles (corresponding to local differences), and the fit quality tends to decrease for larger genetic distances. Qualitatively, we find repeatedly that the PCA-biplots mirror large-scale geography by reflecting the strongest gradients of diversity in a panel, such as the Out-of-Africa expansion in the AEA panel (Fig. 1b), the circum-Mediterranean and circum-Saharan distribution of diversity in Western Eurasia and Africa, respectively, and gradients from Europe into East Asia and South Asia in the Central/Eastern Eurasian panel (Fig. 2). PCA easily identifies outlier or admixed individuals (e.g. in Africa) that are not made apparent in EEMS but which are revealed when exploring model fit. Isolate populations such as the Sardinians and Basques strongly shape the PCA results (compare Fig. 2d to e.g. ref [16]), whereas they are simply placed in low-migration regions in EEMS. Also, many of the fine-scale distortions identified by EEMS are not directly apparent in the PCA-biplots. There are two likely reasons: first, using geographical information allows EEMS to discern subtle structure from effects of uneven sampling[15]. A second reason is how EEMS emphasizes local features - some patterns missing in the PCA-biplots may possibly be teased out in PCA by either investigating higher PCs, or by focusing analysis on an appropriate subset of the data.

Overall, the maps we present provide a compact summary of the complex relationship of genes and geography in human populations. In contrast to methods that identify short bursts of gene flow ("admixture") between diverged populations[22–24], EEMS models the local migration expected between nearby groups as a tool to represent heterogeneous isolation-by-distance patterns. This leads to the first of a few limitations that must be considered in interpretation: Processes that lead to non-spatial patterns of differentiation are not efficiently modelled in the EEMS framework, and resulted in the exclusion of 6.8% of samples (hunter-gatherers, admixed and displaced groups). Second, the results need interpretation in light of the sampling configuration. When there is a feature inferred in a region with few samples, the exact positioning of the inferred change on the map will be imprecise (e.g. the trough presumably associated with the English Channel in Fig 2b). The maps of posterior variance (Extended Data Figures 2 and 4) partly convey where there is uncertainty in positioning, but caution is still warranted as the modelling assumptions will introduce further uncertainty. Third, the maps inferred here represent a model of gene flow that predicts genetic diversity in humans sampled today – a fuller representation would represent genetic structure dynamically through time. This is especially relevant as ancient DNA data have recently suggested human population structure can be surprisingly dynamic (e.g. ref. [25]). Finally, the effective migration rates and their scales needs be interpreted with care. Low effective migration between a pair of populations does not imply an absence of migration nor large levels of absolute differentiation. In each of our maps the overall levels of differentiation are consistently low across all populations.

Nonetheless, the maps presented here provide a useful representation of human genetic diversity, that complements results from geography-agnostic methods. Our results emphasize the importance of geographical features on shaping human genetic history and help explain fine-scale patterns of human genetic diversity[26]. By using recent large-scale SNP data and a novel analysis method, our work expands beyond previous studies of gene flow barriers in humans[27–29]. Our rugged migration landscapes suggest a synthesis of the clusters versus clines paradigms for human structure[6,7,30]: By revealing both sharp and diffuse features that structure human genetic diversity, our results suggest that more continuous definitions of ancestry in human population genetics should complement models of discrete populations with admixture. As rare disease variants are commonly geographically localized[31], the maps presented here may help predict regions where clustering of alleles should be expected. They also annotate present-day population structure that ancient DNA and historical/archaeological studies should aim to explain.

# References

1. Veeramah, K. R. & Hammer, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. *Nat. Rev. Genet.* **15,** 149–162 (2014).

2. Schraiber, J. G. & Akey, J. M. Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* **16,** 727–740 (2015).

3. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475,** 163–165 (2011).

4. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538,** 161–164 (2016).

5. Roberts, L. How to sample the world's genetic diversity. *Science* **257,** 1204–1205 (1992).

6. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298,** 2381–2385 (2002).

7. Serre, D. & Pääbo, S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* **14,** 1679–1685 (2004).

8. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).

9. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

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

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

12. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).

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

14. Lipson, M. *et al.* Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol. Biol. Evol.* **30,**

1788–1802 (2013).

15. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48,** 94–100 (2016).

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

17. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15942–15947 (2005).

18. 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).

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

20. Pugach, I., Delfin, F., Gunnarsdóttir, E., Kayser, M. & Stoneking, M. Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 1803–1808 (2013).

21. Pugach, I. *et al.* The Complex Admixture History and Recent Southern Origins of Siberian Populations. *Mol. Biol. Evol.* **33,** 1777–1795 (2016).

22. Patterson, N. J. *et al.* Ancient Admixture in Human History. *Genetics* genetics.112.145037 (2012).

23. Loh, P.-R. *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193,** 1233–1254 (2013).

24. Hellenthal, G. *et al.* A Genetic Atlas of Human Admixture History. *Science* **343,** 747–751 (2014).

25. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).

26. Baker, J. L., Rotimi, C. N. & Shriner, D. Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Sci. Rep.* **7,** 1572 (2017).

27. Barbujani, G. & Sokal, R. R. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 1816–1819 (1990).

28. Barbujani, G. & Belle, E. M. S. Genomic boundaries between human populations. *Hum. Hered.* **61,** 15–21 (2006).

29. Pagani, L. *et al.* Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538,** 238–242 (2016).

30. Rosenberg, N. A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* **1,** e70 (2005).

31. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet.* **10,** e1004528 (2014).

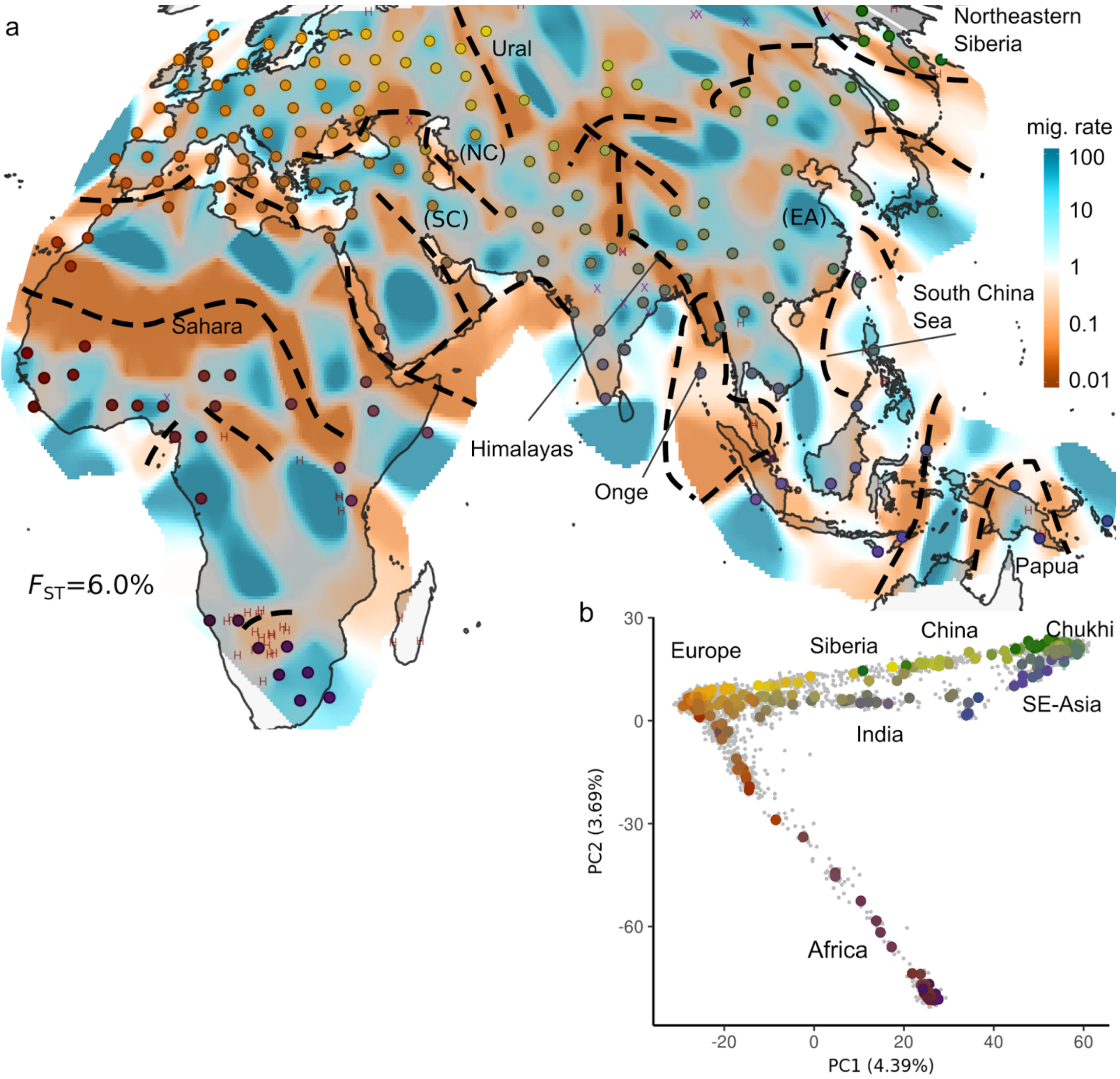


**Figure 1: Large-scale patterns of population structure. a:** EEMS posterior mean effective migration surface for Afro-Eurasia (AEA) panel. ‘X’ marks locations of samples excluded as displaced or recently admixed. ‘H marks locations of excluded hunter-gatherer populations. Regions and features discussed in the main text are labeled. Approximate locations of troughs are annotated with dashed lines (see Extended Data Figure 4). **b:** PCA plot of AEA panel: Individuals are displayed as grey dots, colored dots reflect median of sample locations; with colors reflecting geography and matching with the EEMS plot. Locations displayed in the EEMS plot reflect the position of populations after alignment to grid vertices used in the model (see methods). For exact locations, see annotated Extended Data Figure 2 and Table S1. The displayed value of $F_{ST}$ emphasizes the low absolute level of differentiation in human SNP data.

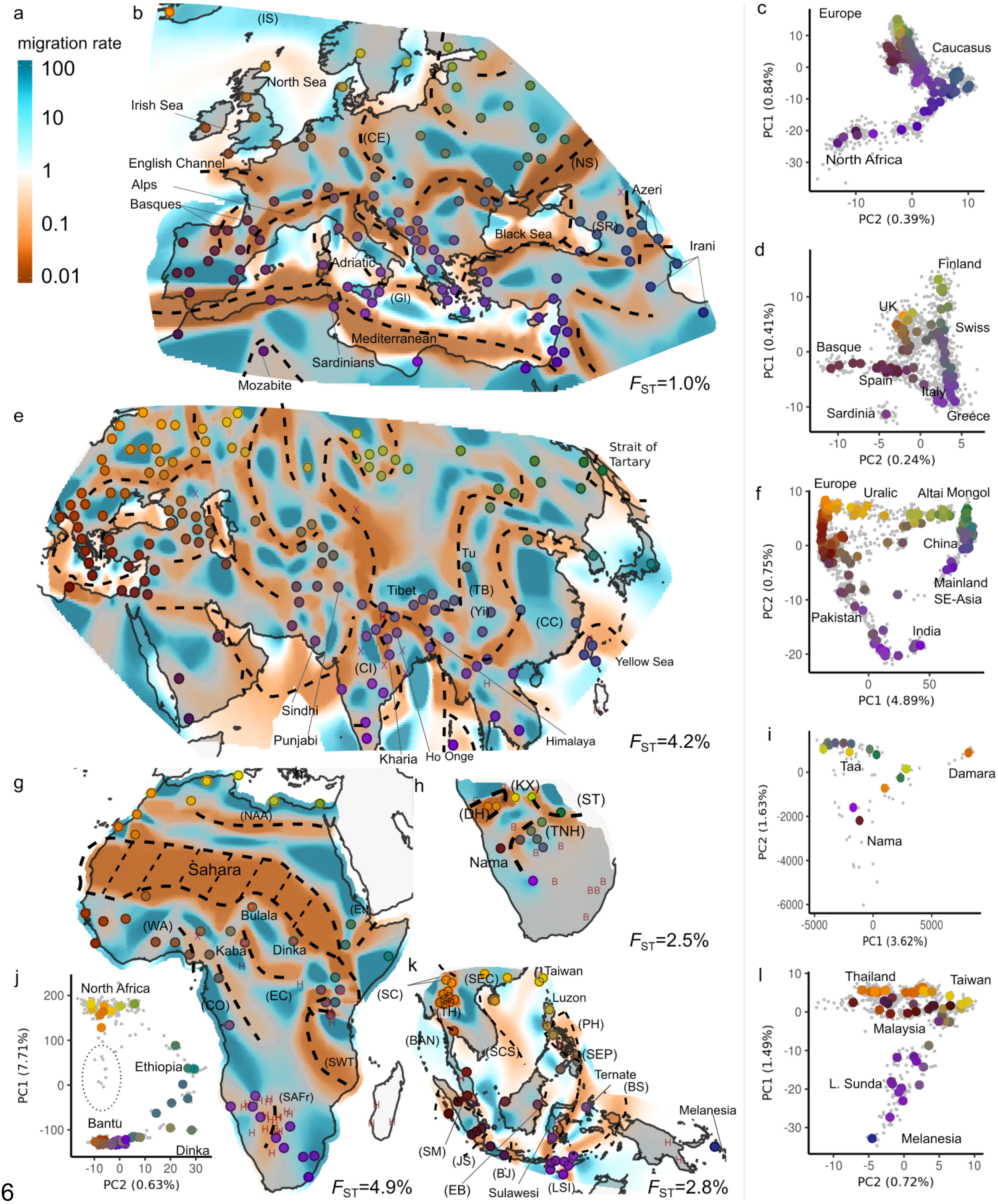


**Figure 2: Regional patterns of genetic diversity. a:** scale bar for relative effective migration rate. Posterior effective migration surfaces for **b**: Western Eurasia (WEA) **e**: Central/Eastern Eurasia (CEA) **g**: Africa (AFR) **h** Southern African hunter-gatherers (SAHG) **k**: and Southeast Asian (SEA) analysis panels. 'X' marks locations of samples noted as displaced or recently admixed, 'H' denotes Hunter-Gatherer populations (both 'X' and 'H' samples are omitted from the EEMS model fit); in panel g, red circles indicate Nilo-Saharan speakers and in panel h, 'B' denotes Bantu-speaking populations. Approximate location of troughs are shown with dashed lines (see Extended Data Figure 4). PCA plots: **c**: WEA d:Europeans in WEA **f**: CEA **i**: SAHG **j**: AFR **l**: SEA. Individuals are displayed as grey dots. Large dots reflect median PC position for a sample; with colors reflecting geography matched to the corresponding EEMS figure. In the EEMS plots, approximate sample locations are annotated. For exact locations, see annotated Extended Data Figure 4 and Table S1. Features discussed in the main text and supplement are labeled. $F_{ST}$ values per panel emphasize the low absolute levels of differentiation.

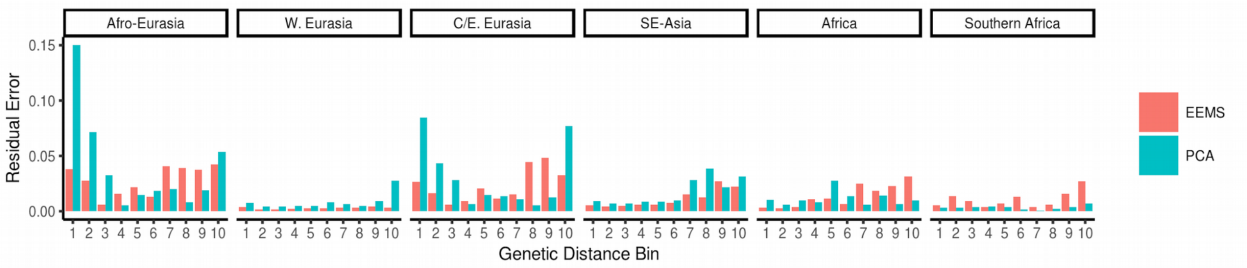


Figure 3: Comparing Fit of PCA and EEMS. We show the relative error of EEMS (red) and PCA(blue, first 10 PCs) for all pairs, stratified by genetic distance. For each panel, all pairwise genetic distances were distributed in ten bins of equal size, for which we then computed the median absolute error of the fitted model vs the observed distances. For W. Eurasia and SE-Asia, EEMS fits uniformly better than PCA. In the Afro-Eurasian, Central/Eastern Eurasian and African panel, EEMS fits better for smaller distances, but the fit is worse for larger distances. For the Southern African Hunter-Gatherers, EEMS fits worse than PCA for all distance bins.

# Material and Methods

## Merging pipeline

We obtained SNP genotype data from 26 different studies (Extended Data Table 1). Processing was done using a reproducible snakemake pipeline[32] available under http://github.com/NovembreLab/eems-merge, heavily relying on plink 1.9[33] for handling genotypes. The sources differ in the input format and pre-processing, however in general we performed the following steps:

1. Remove all non-autosomal, non-SNP variants
2. Map SNP to forward strand of human reference genome b37 coordinates using chip manufacturer metadata files or SNP identifiers
3. Remove strand-ambiguous A/T and G/C variants

The remaining SNPs were then merged using successive plink --bmerge commands into a single master dataset with 9,003 individuals and 1.9M SNPs but a total genotyping rate of only 20.6%. 46 SNPs were removed because different studies reported different alternative alleles. We used a relationship filter of 0.6 using the “--rel-cutoff 0.6” flag in plink to remove 667 closely related individuals or duplicates. After merging, each analysis panel had missingness rates <0.5% (AEA=0.2%, WEA=0.3%, CEA=0.2%, SEA=0.5%, AFR=0.2%, SAHG=0.1%). In all panels, all SNPs passed a one-sided HWE-test (p-value< $10^{-5}$), with the exception of SEA, where nine (out of 8507 SNPs) failed and were excluded.

## Data Retrieval and Filtering

### Human Origins data set[25]

Sampling location information was obtained from table S9.4 of ref. [25], and the data were shared by David Reich. We used the population information in the `vdata` subset of all ascertainment panels, except for the analysis where we assess ascertainment bias.The utility `convert` from `admixtools`[22] was used to convert the data into plink format.

### Estonian Biocentre data

The data generated by the Estonian Biocentre[34] were provided in plink format by Mait Metspalu on 10/30/15, along with location information where it was available. This data set contained 1,282,568 SNPs. Of those, 6770 SNPs had non-unique ids and were removed.

### HUGO Pan-Asian SNP consortium[35]

The data were downloaded on 6/24/15 from www.biotec.or.th/PASNP. Location-metadata were obtained on the same day from the map on the same website, and individuals were matched to populations using the individual identifiers. All individuals with the same tag were assigned the median of all locations from that tag. The data were first lifted onto hg19 (with 5 out of 54794 SNPs being removed), and then re-formatted into binary plink format. Due to the small size of the chip used and the low overlap with the human origins array in particular, we only consider this data in the South-East Asian panel.

### Uniform global sample [36]

This data were downloaded on 6/20/15 from http://jorde-lab.genetics.utah.edu/pub/affy6_xing2010/. Sampling locations were provided by Jinchuan Xing. We used version 32 of the annotation file obtained on 6/19/15 from affymetrix.com to map SNPs onto hg19, remove strand-ambiguous SNPs and to flip SNPs that were on the minus-strand.

### POPRES data[37]

POPRES data were obtained under dbGAP study accession phs000145 to John Novembre, and we used the data as processed in ref [16], and only retain individuals for which all grandparents were from the same country, and labelled the Swiss sample according to self-reported language. We used version 32 of the annotation file obtained on 6/19/15 from www.affymetrix.com ("Mapping250K_sp.na32.annot.csv" and "Mapping250K_Sty.na32.annot.csv") to filter SNPs that did not map onto hg19 and we removed strand-ambiguous AT and GC polymorphisms.

### African data

Data from refs [38,39] were obtained on 04/19/17 from David Comas' website under http://www.biologiaevolutiva.org/dcomas/?p=607. We used version 32 of the annotation file GenomeWideSNP_6.na32.annot.csv" obtained on 6/19/15 from affymetrix.com to map SNPs onto hg19, remove strand-ambiguous SNPs and to flip SNPs that were on the minus-strand.

### South-East Asian data[40]

The data were obtained on 7/14/15 from Mark Stoneking in three different source files. After merging the three different source files, SNPs not mapping to hg19 using the annotation file

"GenomeWideSNP_6.na32.annot.csv" were removed, as were AT and GC SNPs. Sampling locations were extracted from Figure 1 of ref [40]

### Mediterranean Panel[41]

Data were obtained on 8/13/15 in binary plink format from http://drineas.org/Maritime_Route/RAW_DATA/PLINK_FILES/MARITIME_ROUTE.zip. Sampling location information was obtained from Supplementary Table 3 in ref. [41]. SNPs not mapping to hg19 using the annotation file "GenomeWideSNP_6.na32.annot.csv" were removed, as were AT and GC SNPs.

### Tibetan and Himalayan data

Data from refs [42–44] were obtained from Choongwon Jeong and Anna Di Rienzo. We used the same filtering as in the [42] study, but only added the samples originating from these three studies with permission from the respective authors.

## Combining Meta-information

All sources with the exception of the Estonian Biocentre data provided (approximate) sampling coordinates. However, the level of accuracy varied between sources, with some providing specific ethnicities, some (such as POPRES) only providing country information and others just providing city- or state-level information. For POPRES-derived data, and most countries, we assigned individuals to the country's centerpoint, with the exception of Sweden and Finland, which were assigned their capital.  For the Estonian Biocentre data, sampling location data were highly heterogeneous. Samples that could not be confidently assigned to a region with an accuracy of around 100km were excluded. For populations with samples from multiple studies, the most accurate source location was used. For locations covered with different accuracy, only the most accurate samples were retained. For example, we dropped all Spanish individuals from POPRES (only country level data), as the Human Origins data provided higher resolution, with samples from eleven different regions in Spain. The resulting table is given as Table S1.

### Samples omitted from model fitting

Besides samples whose geographic origin we could not unambiguously assign (n=682), we chose to label on the maps a number of samples that would violate some assumptions of the EEMS model (n=593) without including them in the model fitting. As EEMS assumes an isolation-by-distance model, populations that have recently migrated a long distance have undue influence on the results. As EEMS is also multivariate and analyzes all data sets jointly, these events can have an undesirable disproportionate effect on the estimated surface. As a consequence, we chose to a priori omit known displaced or recent migrant populations (where we define recent as approx. the last 500 years). Similarly, we omit populations that are known *a priori* to be admixed between source groups that are clearly distinct. The resulting populations are denoted as "ADMIX" in Table S1. These include the Han-Chinese in Singapore and Han-Chinese in Taiwan, who both are recent migrant populations to those locales, as well as the Uygurs, who are admixed between East Asian and Europeans. In India, we omitted the Bhunjia, Dhurwa and Gond samples, who were denoted as admixed by the primary publication[45].

Furthermore, we omitted the Kusundas, who have both Indian and Tibetan ancestry, and the Kalmyks, who moved from present-day China to the Caspian Sea in the 17th century. Finally, we omitted the Yakut, who have both Turkic and East Asian ancestry, and all Jewish samples, due to complexity of the diaspora and subsequent local admixture[46].

In addition, we label but omit from model fitting most hunter-gatherer populations because they frequently live in and around other human groups with limited interaction, giving thus two layers of structure that EEMS does not model. (Extended Data Figure 6c). An exception was made for Onge, since they are geographically isolated from other subsistence populations, and have been interpreted as fundamental to understand Indian population structure.[19] In addition, hunter-gatherers make up a very small proportion of modern human genetic diversity, but are well-studied genetically, and combining the samples would include a bias that would be difficult to control. We do, however, analyze the South-African hunter gatherer samples, as we have very dense samples from a single region, but do not include them in our Afro-Eurasian and African panels. Other African samples we omit are the Mbuti and Biaka, the Hazara and Sandawe and all Malagasy samples. We also omitted several high latitude samples, as most of these samples contain groups that have largely hunter-gatherer ancestries, were nomadic or were recently displaced and thus are difficult to place in an explicit geographic setting. This included the Saami in Europe as well as the Arctic Karelian, Nenets, Nganasans, Chukchi, Dolgan and Aleuts. In South-East Asia, we omit all Negrito samples as well as the Aeta and Ati on the Phillippines.

Finally, to avoid any possible distortion due to uneven sampling, we downsampled all single locales to at most 50 individuals, drawn independently for different panels. This resulted in a total of 5372 individuals used in at least one panel (Supplementary Data Table 1).

## Visualization pipeline

We developed a second pipeline using snakemake[32] to perform all subsetting and demographic analyses, available under github.com/NovembreLab/eems-around-the-world. The pipeline allows for defining panels using a flexible set of features, latitudinal and longitudinal boundaries, continent or country of samples, source study, as well as the addition and exclusion of particular samples or populations. Based on these subsets, different modules allow performing EEMS and PCA analyses, as well as generating all the figures, that were then annotated using inkscape. All configuration variables are stored in json and yaml config files. We perform EEMS and PCA for each panel independently. Structural variants are a potential confounding factor for genome-wide SNP based analysis. In PCA, these variants may result in a number of neighboring SNP in high LD to have very high loadings, thus overemphasizing the effect of these variants. For this reason, it is advisable to remove regions containing SNP that have extremely high loadings on some Principal component. Thus, for each panel, we perform a preliminary PCA analysis using flashpca[47]. The loading-scores for each PC were normalized by dividing them by the standard deviations on each PC [outlier_score = L[i]/sd(L[i]) ], and then we removed a 200kb window around any SNP for which |outlier_score| > 5. We also dropped individuals with more than 5% missingness, and SNPs with more than 1% missing data from each panel.

### EEMS

To generate the map surfaces, we must choose a grid size and boundaries.  Choosing a coarse grid results in faster computation, but only produces a map with broad-scale patterns. A finer grid, on the other hand, is able to reveal more details, but at a steep increase in computational cost and with an increased danger of introducing patterns that are harder to interpret.  Grid density and sizes are given in Extended Data Table 1, along with a population level $F_{ST}$ calculated using plink.

We evaluated the impact of SNP ascertainment bias by running EEMS on the multiple, documented SNP ascertainment panels of the Human Origins data[25].  We found that while ascertainment bias has an effect on the heterozygosity surfaces that EEMS estimates, the migration surfaces remain relatively unaffected (Extended Data Fig. 1). Therefore, we restrict our presentation to the migration surfaces.

For each panel, we performed six pilot runs of 6 million iterations each. The run with the highest likelihood was then used for a second set of four runs of 4 million iterations each, with the first 1 million discarded as burn-in. Every 10,000th iteration was sampled. EEMS approximates a continuous region with a triangular grid, which has to be specified. We generated global geodesic graphs at three resolutions (approximate distance between demes of 100, 200 and 500km, respectively) using dggrid v6.1[48] and intersected these graphs with the area representing each panel (Extended Figures 2,3). All other (hyper-)parameters were kept at their default values[49].

We compared EEMS to an isolation-by-distance model with a constant migration rate by re-fitting EEMS allowing only a single migration rate tile, but arbitrary diversity rate tiles using the otherwise same settings. The resulting log Bayes Factors are given in Extended Table 2.

### Evaluating fit of EEMS and PCA to genetic distances

For EEMS, the posterior samples imply an expected distance matrix between populations. For PCA, the components and their loadings provide an approximation to the genetic distance matrix between individuals. We use the median PCA values of individuals across ten PC components to produce an expected genetic distance matrix between populations.  We use ten PC components as most investigators evaluate population structure based on only the first several PCs. For each method the expected genetic distance matrices are compared to the observed matrices using a simple linear correlation computed between all pairwise distances.

## Acknowledgements

We are grateful for helpful comments from Choongwon Jeong, Matthew Stephens, Anna Di Rienzo, Melinda A. Yang, Joshua G. Schraiber and members of the Novembre lab.  This research was supported by research grants NIH/NCI U01 CA198933 and NIH/NIGMS R01

GM108805 (J. N. and B. M. P) and by a Swiss National Science Foundation early postdoc mobility fellowship (B. M. P.). We dedicate the paper in memoriam of Brad McRae (1966-2017) whose work on resistance distances underlies the EEMS methodology.

Author contributions.

B.M.P. analyzed data. B.M.P., D.P., and J.N. interpreted results. B.M.P and J.N conceived the study and wrote the manuscript.

## Competing financial interests

The authors declare no competing financial interests.

## Correspondence to:

Benjamin Peter (benj.pet@gmail.com), or John Novembre (jnovembre@uchicago.edu)

## Data Availability

The source and availability of all data is outlined in the methods (“Data Retrieval and Filtering” subsection).

## Supplementary Text on Regional Scale Analyses

Here we provide a more expanded discussion of the regional-scale results. To help identify features that we discuss, we have added labels to discussed features in the figures, and refer to them in the text here in parentheses. The labels are typically capitalized abbreviations and in some cases are full words.

**Europe.** Europe appears largely homogeneous in the Afro-Eurasia panel, but a finer-scale analysis (Western Eurasia panel, Fig. 2a; n=2,018; 119 locales, $F_{ST}$=0.010) reveals abundant fine-scale structure: bodies of waters are consistently covered by lower effective migration regions, with migration being lower in southern seas (Mediterranean, Adriatic, Black Sea) relative to those in northern Europe (North Sea, Irish Sea, English Channel). Terrestrial barriers are observed in: The Alps (and an adjacent region extending into Southern France), surrounding the Mozabites in Tunisia, the western and northern edges of the Arabian desert (though we note the region has few samples). Troughs reflecting historical domains are observed: between Germanic and Northern Slavic-speakers (CE), between domains of Slavic-speakers and the Caucasus (NS), and in the Caucasus in a region with Irani , Azeri, and Adygei in Southern Russia (SR). Remaining regions are generally inferred to have above average migration, with one obvious corridor being that between Iceland and Scandinavia, presumably due to the recent colonization of Iceland. One interesting feature is an area of East-West low migration between the Italian peninsula and Greece (GI). A corridor between Crete and Sicily is inferred south of it, and between mainland Greece and southern Italy north of it. This likely reflects a pattern of close genetic similarity among coastal Mediterranean populations observed previously[41] but suggests it may have north-south structure. Ancient DNA results suggest that the patterns we observe are recent[50,51] and have been shaped in the last 3,000-5,000 years with contributions from multiple sources. Strikingly, proposed expansion routes through the Eurasian Steppe and Levant into Europe partially align with corridors of high effective migration.

**Central/Eastern Eurasia.** The Central/Eastern Eurasia surface (Fig. 2e; n=2,411; 163 locales, $F_{ST}$=0.042) is overall similar to the patterns seen in the AEA panel, with a trough through the Himalayas/Tien-Shan and two corridors connecting Europe with Central Asia around the Caspian Sea. Particularly in India and East Asia the higher resolution EEMS analysis reveals additional details: Where the global analysis did not reveal any strong patterns in South Asia, at the higher resolution we observe troughs in the Indian subcontinent between central India (CI) and populations to the north (Sindhi, Punjabi), two Austroasiatic speaking populations to the east (Kharia, Ho), and Southern India, where Dravidian languages are most common.

In East Asia, we observe marine troughs in the East China Sea, strait of Tartary and the Andaman Sea (Onge). Terrestrially, we observe troughs between coastal China (CC), a central region with several Tibeto-Burman samples (TB, along with the Tu who speak a Mongolic language, and have been suggested to have received European admixture 1,200y ago[24]), and a western region anchored by Tibetan samples. The coastal Chinese region extends in a corridor into Korea and Japan.

Overall, the Central/East Asia panel is particularly complex with one of the lowest levels of $r^2$ between EEMS expected genetic distances and the observed distances ($r^2$ = 0.66, Extended Data Fig. 5). This is expected as the relatively open steppe has been the site of repeated long-range population movements and invasions, by e.g. Bronze Age Steppe populations, Mongols and Turkic speakers, that we expect are difficult to depict using the model of steady-state gene flow model fit by EEMS.

**South-East Asia.** In the South-East Asian panel (n=940, 53 locales; $F_{ST}$=0.028; Fig. 2k) troughs align with the many seas and channels in this region: the South-Chinese Sea (SCS), the waterway running east of the Philippines (PH) and Sulawesi south to the Flores Sea (SEP), the waterway between western New Guinea into the Banda Sea (BS), the Malacca strait between Sumatra and Malaysia (SM), the Sunda Strait between Java and Sumatra (JS), the Java Sea between Bali and Java (BJ), as well as the Makassar strait and Celebes Sea between Borneo and Sulawesi (EB). Two corridors, one from Taiwan/Luzon through Western Mindanao to Sulawesi, and one from Ternate through the Lower Sunda Islands (LSI) into Melanesia possibly reflect the Austronesian expansion that started roughly 3,000 years ago[52]. On the mainland, we find low effective migration north of Bangkok (BAN) and near samples from Northern Thailand (TH) (including the Southern Chinese Wa and Jinuo samples (SC)). These two samples have low inferred effective migration with South-Eastern Chinese samples (SEC).

**Africa.** In Africa, we analyze non-hunter-gatherers (AFR, n=521, 47 locales, $F_{ST}$=0.049; Fig. 2g) and South-African hunter-gatherers (SAHG, n=109, 16 locales, $F_{ST}$=0.025; Fig 2h) independently, as traditional hunter-gatherers and farmers are typically differentiated and it is difficult for EEMS to model large genetic dissimilarities at close geographic proximity (Extended Data Fig. 6a)[53,54]. In the AFR-panel, language group boundaries align with several troughs: a large one extends through the Sahara into Eastern Africa, roughly along the boundary of Niger-Congo and Afro-Asiatic language speakers[55]. In sub-Saharan Africa, west Africa appears as a high-gene-flow region (WA), and two corridors pass from Nigeria - one along the coast of Congo (CO) southwards and another further east (EC) connecting to Kenya and Tanzania. The South African samples (all Bantu speakers) are split into Eastern and Western Bantu groups by a single trough. In both Central and Eastern Africa Nilo-Saharan and Niger-Congolese speakers overlap, resulting in low effective migration imperfectly correlated with language groups: The Nilo-Saharan Dinka and Bulala are in a region of high gene flow, to the exclusion of the Kaba. Southern and Eastern Africans are separated by low effective migration through Mozambique and South-Western Tanzania (SWT).

In Northern Africa (NAA), we see a trough of low effective migration separating two latitudinal corridors; one following the Mediterranean coast and one inland (Fig 2g). The inland corridor disappears in our lower-resolution Afro-Eurasia panel (Figure 1a) and when we drop individuals from Western Sahara that appear intermediate between North Africa coastal populations and East African populations (Extended Data Fig. 6d). We suspect this corridor emerges as EEMS attempts to model ancestry in Western Sahara populations that is distinct from that found in coastal North Africa.

For the South African Hunter-Gatherers (Fig. 2h) most samples fall into a central region with high effective migration, including the Taa, Naro and Hoan (TNH). Troughs in the North separate this region from the Sua and Tswa (ST) and in the south-west from the Khomani and Nama (Nama), respectively. The remaining samples fall either into a Northern high migration area (Khwe and Xuun, KX) or a North-Western low migration area (Damara and Haiom, DH). These results are broadly consistent with existing work on African population structure[56–59], and emphasize African population structure appears largely determined by the Sahara desert, the Bantu and Arabic expansions, and the complex structure of hunter-gatherer groups specifically in South Africa.

# Extended Data

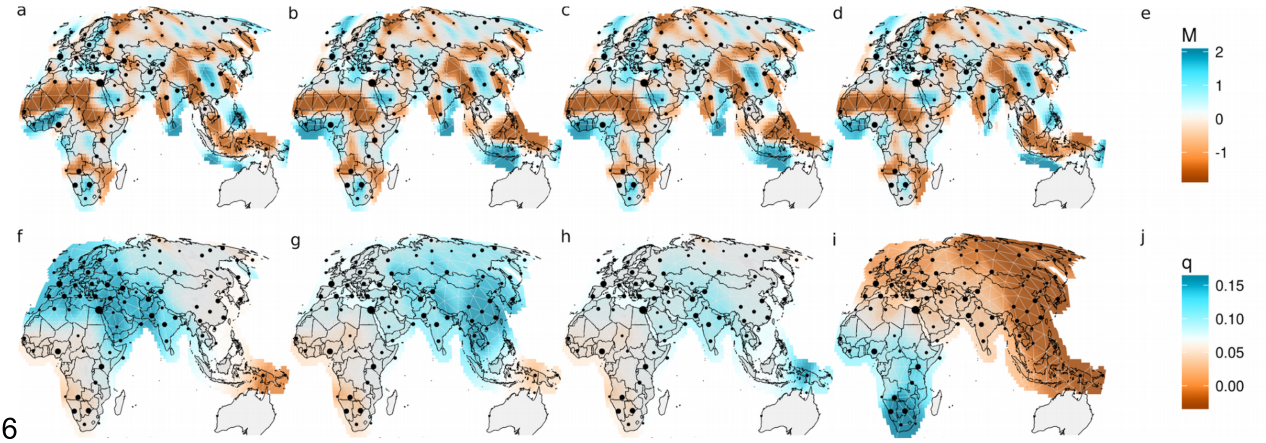
| Authors | Abbrev | Ind. | Loc. | Reference |
|---|---|---|---|---|
| Bryc et al. 2009 | B09 | 109 | 10 | Ref. 38 |
| Behar et al. 2010 | Be10 | 295 | 22 | Ref. 60 |
| Behar et al. 2013 | B13 | 130 | 20 | Ref. 61 |
| Bigham et al. 2010 | Bi10 | 45 | 3 | Ref. 43 |
| Cardona et al. 2014 | C14 | 120 | 16 | Ref. 62 |
| Chaubey et al. 2011 | C11 | 26 | 3 | Ref. 45 |
| Di Cristofaro et al. 2013 | D13 | 5 | 1 | Ref. 63 |
| Fedorova et al. 2013 | F13 | 24 | 3 | Ref. 64 |
| HUGO Pan-Asian SNP Consortium 2009 | H09 | 870 | 42 | Ref. 35 |
| Hunter-Zinck et al. 2010 | H10 | 86 | 1 | Ref. 39 |
| Jeong et al. 2017 | J17 | 53 | 2 | Ref. 42 |
| Kovacevic et al. 2014 | K14 | 70 | 6 | Ref. 65 |
| Lazaridis et al. 2014 | L14 | 1409 | 142 | Ref. 66 |
| Metspalu et al. 2011 | M11 | 120 | 7 | Ref. 67 |
| Migliano et al. 2013 | M13 | 49 | 3 | Ref. 68 |
| Paschou et al. 2014 | Pa14 | 621 | 29 | Ref. 41 |
| Pierron et al. 2014 | Pi14 | 16 | 1 | Ref. 69 |
| Nelson et al. 2008 | N08 | 542 | 29 | Ref. 37 |
| Raghavan et al. 2014 | R14 | 75 | 7 | Ref. 70 |
| Rasmussen et al. 2011 | Ra11 | 3 | 1 | Ref. 71 |
| Rasmussen et al. 2010 | R10 | 44 | 3 | Ref. 72 |
| Reich et al. 2011 | Re11 | 96 | 15 | Ref. 73 |
| Xing et al. 2010 | X10 | 92 | 4 | Ref. 36 |
| Xu et al. 2011 | X11 | 28 | 3 | Ref. 44 |
| Yunusbayev et al. 2012 | Y12 | 183 | 14 | Ref. 74 |
| Yunusbayev et al. 2015 | Y15 | 247 | 36 | Ref. 34 |

**Extended Data Table 1:** Data Sources. Abbrev: Abbreviation; Ind: total number of individuals; Loc. Number of unique sample locations

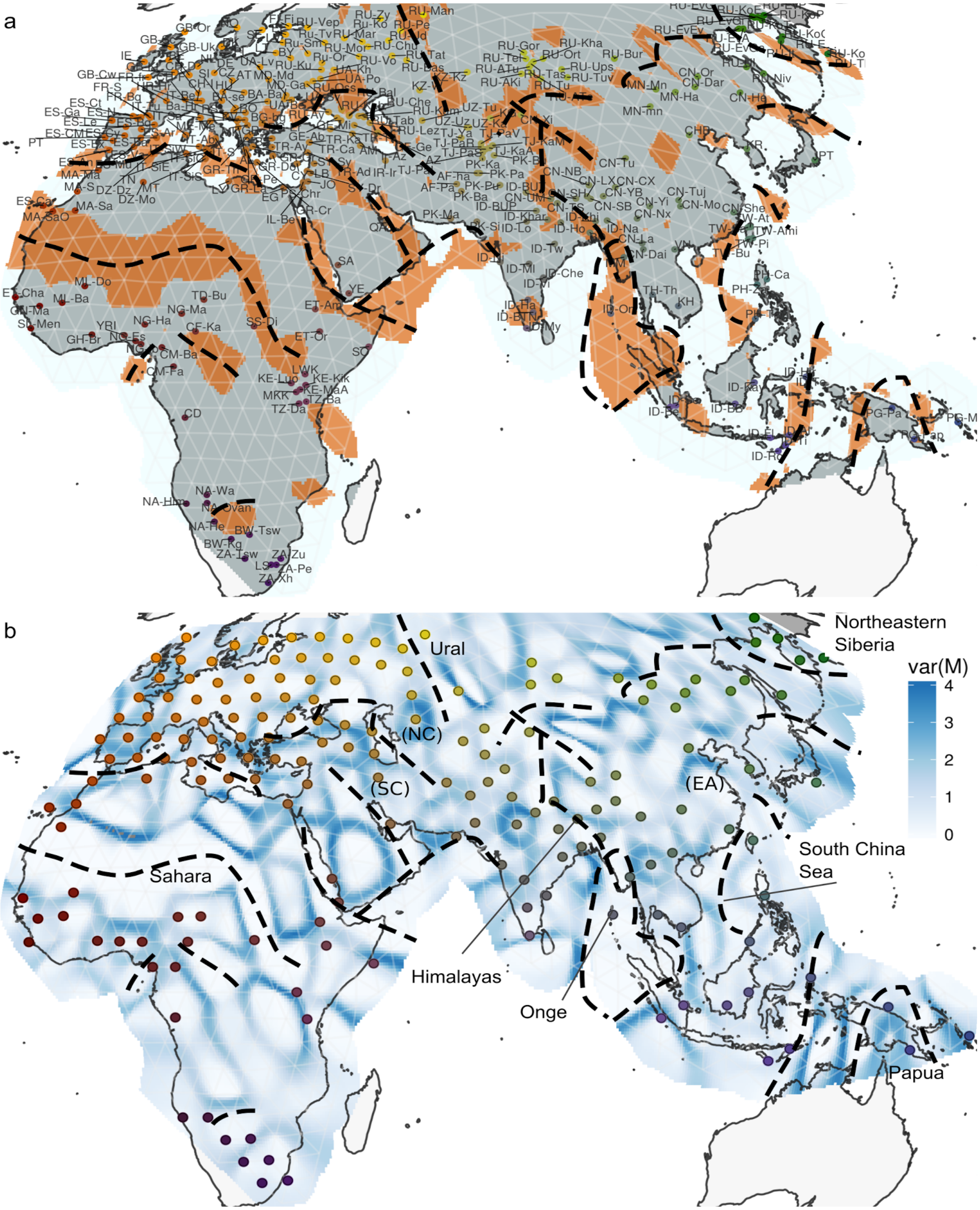
| Panel | Abb. | Individuals | Locations | SNPs | Grid Size (# of demes) | Resolution (km) | $F_{ST}$ | Support (log-BF) |
|---|---|---|---|---|---|---|---|---|
| Afro-Eurasia | AEA | 4006 | 291 | 20167 | 620 | 500 | 0.0605 | 232,047 |
| Western Eurasia | WEA | 2018 | 119 | 26358 | 1320 | 100 | 0.0097 | 41,371 |
| Central/Eastern Eurasia | CEA | 2411 | 163 | 21060 | 1078 | 200 | 0.0417 | 111,794 |
| South-East Asia | SEA | 939 | 52 | 8498 | 1388 | 100 | 0.0284 | 9,378 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Africa | AFR | 521 | 47 | 19493 | 647 | 200 | 0.0490 | 4,881 |
| Southern Africa | SAHG | 109 | 16 | 532343 | 227 | 100 | 0.0249 | 1,448 |

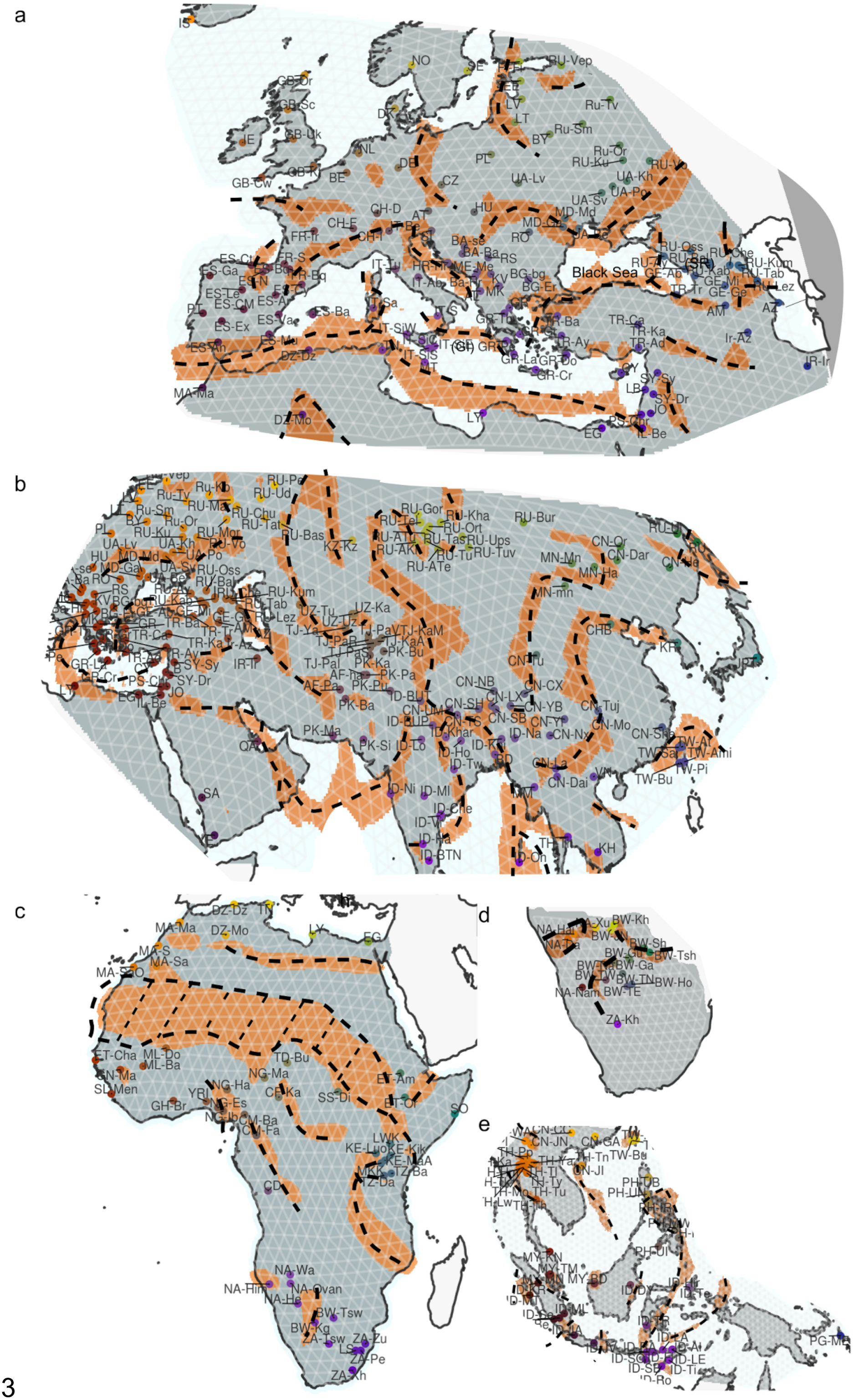
**Extended Data Table 2:** Analysis Panels. Abb. Panel Abbreviation. Res. Avg. distance between grid points (in km) ; Support: log Bayes factor in favor of complex vs constant migration model.



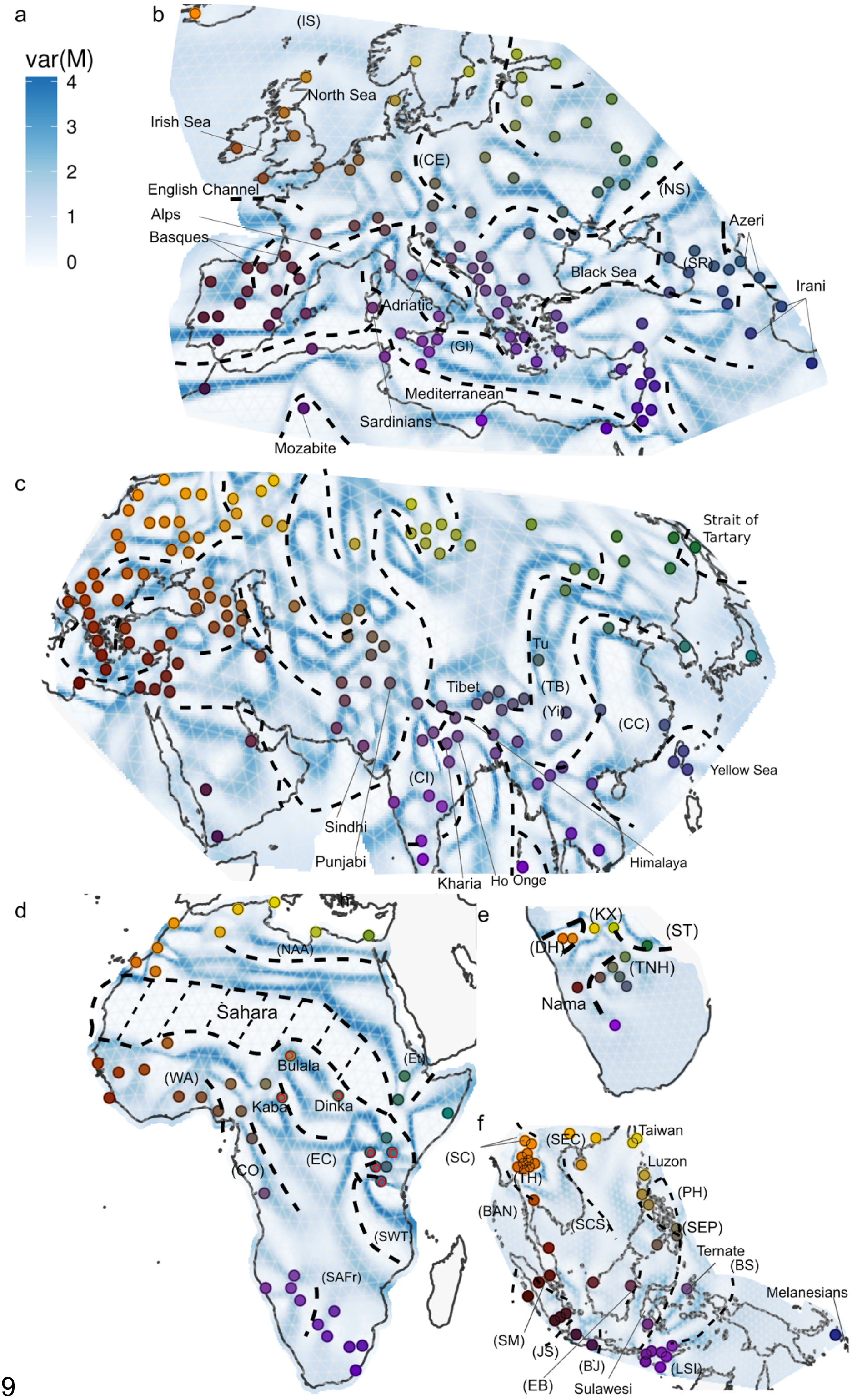
**Extended Data Figure 1**: Ascertainment bias. We run EEMS only using the Human Origin data[25], using SNPs ascertained in a French (a/f), Chinese (b/g), Papuan (c/h) and San(d/i) individual. Migration rate surfaces (a-d) remain robust, whereas the within-deme diversity surfaces (f-i) show highests diversity at the respective ascertainment location. e/j: scale bars for migration rates and within-deme diversity rate parameters, respectively.



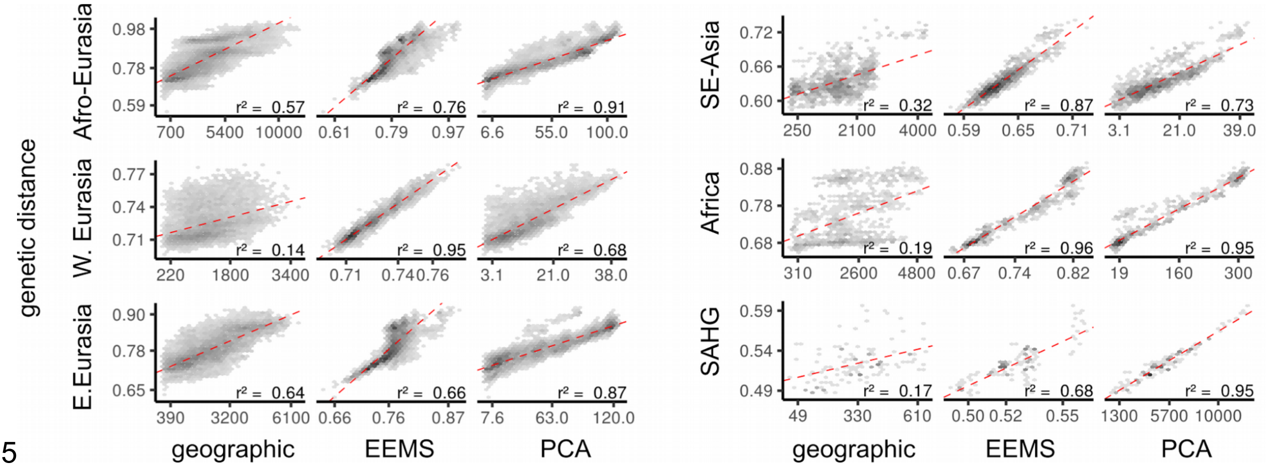
**Extended Data Figure 2: a:** Location of troughs (below average migration rate in more than 95% of MCMC iterations) are given in brown. Sample locations and EEMS grid are displayed. **b:** Posterior variance on migration rate parameters. Note that most significant features are in low variance regions, but that they are often surrounded by high-variance regions, implying the exact boundary of troughs is estimated with uncertainty. Grid-fitted sample locations are displayed. Annotation in both panels is identical to Figure 1a.



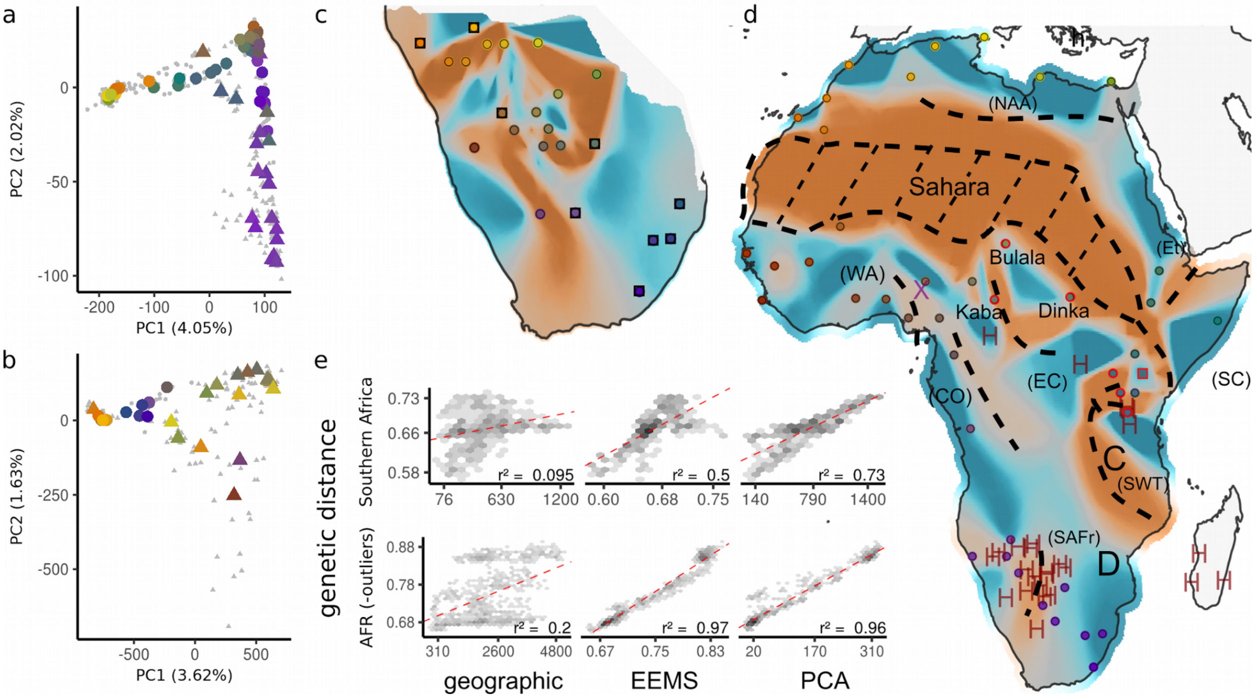
**Extended Data Figure 3:** Location of troughs (below average migration rate in more than 95% of MCMC iterations) are given in brown. Sample locations and EEMS grid are displayed for **a**: WEA **b**: CEA **c**: AFR **d:** SAHG and **e**: SEA analysis panels. Annotation in all panels is identical to Figure 2. $F_{ST}$ values are provided per panel to emphasize the low absolute levels of differentiation.



**Extended Data Figure 4:** Posterior variances in migration rate parameters. Grid-fitted sample locations are displayed .**a**: scale bar **b:** WEA **c**: CEA **d**: AFR **e:** SAHG and **f**: SEA analysis panels. Note that most significant features are in low variance regions, but that they are often surrounded by high-variance regions, implying the exact boundary of troughs is estimated with uncertainty. Annotation of troughs and select features is identical to Figure 2.



**Extended Data Figure 5:** Hex-binned scatterplots of genetic distance versus geographic distance (in km), predicted distance via EEMS model fit, and predicted distance via a ten-component PCA, for all panels. Darker areas correspond to bins with more points. The fit of a simple linear regression (red dashed lines) and r² are given.



**Extended Data Figure 6:** Results for Africa panels with all samples analyzed. **a**: PCA of all African samples **b**: PCA of all Southern African samples. In both samples, individuals annotated as hunter-gatherers are displayed as triangles. Colored dots reflect median of sample locations; with colors reflecting geography and matching in the corresponding EEMS posterior. Approximate sample locations are annotated. For exact locations, see annotated Extended Data Figure 4 and Table S1. **c**: EEMS posterior mean surface of all Southern African samples. Agriculturalist samples are marked with squares. The interspersed geography of Hunter-Gatherers and agriculturalists results in a poor fit with several very sharp boundaries. **d**: EEMS posterior mean of AFR samples with outlier individuals (circled in Figure 2j) removed. The horizontal barrier (NAA) observed in Fig 2g disappeared. **e**: Hex-binned scatterplots of genetic distance versus geographic distance (in km), predicted distance via EEMS model fit, and predicted distance via a ten-component PCA, for the data corresponding to the EEMS maps presented in this figure. Darker areas correspond to bins with more points. The fit of a simple linear regression (red dashed lines) and r² are given. In Southern Africa, geography and EEMS only weakly predict genetic diversity.

# Additional References

32. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

33. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).

34. Yunusbayev, B. *et al.* The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *PLoS Genet.* **11,** e1005068 (2015).

35. HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. *Science* **326,** 1541–1545 (2009).

36. 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).

37. 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).

38. Bryc, K. *et al.* Genome-Wide Patterns of Population Structure and Admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci. U. S. A.* (2009). doi:10.1073/pnas.0909559107

39. Hunter-Zinck, H. *et al.* Population genetic structure of the people of Qatar. *Am. J. Hum. Genet.* **87,** 17–25 (2010).

40. Reich, D. *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* (2011).

41. Paschou, P. *et al.* Maritime route of colonization of Europe. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 9211–9216 (2014).

42. Jeong, C. *et al.* A longitudinal cline characterizes the genetic structure of human populations in the Tibetan plateau. *PLoS One* **12,** e0175885 (2017).

43. Bigham, A. *et al.* Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. *PLoS Genet.* **6,** e1001116 (2010).

44. Xu, S. *et al.* A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol. Biol. Evol.* **28,** 1003–1011 (2011).

45. Chaubey, G. *et al.* Population Genetic Structure in Indian Austroasiatic Speakers: The Role of Landscape Barriers and Sex-Specific Admixture. *Mol. Biol. Evol.* **28,** 1013–1024 (2011).

46. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466,** 238–242 (2010).

47. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9,** e93766 (2014).

48. Sahr, K., White, D. & Kimerling, A. J. Geodesic Discrete Global Grid Systems. *Cartogr. Geogr. Inf. Sci.* **30,** 121–134 (2003).

49. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48,** 94–100 (2016).

50. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (2015).

51. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522,** 167–172 (2015).

52. Duggan, A. T. & Stoneking, M. Recent developments in the genetic history of East Asia and Oceania. *Curr. Opin. Genet. Dev.* **29,** 9–14 (2014).

53. Excoffier, L. & Schneider, S. Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 10597–10602 (1999).

54. Perry, G. H. & Verdu, P. Genomic perspectives on the history and evolutionary ecology of tropical rainforest occupation by humans. *Quat. Int.* **448,** 150–157 (2017).

55. Campbell, M. C. & Tishkoff, S. A. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.* **9,** 403–433 (2008).

56. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324,** 1035–1044 (2009).

57. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 786–791 (2010).

58. Pickrell, J. K. *et al.* The genetic prehistory of southern Africa. *Nat. Commun.* **3,** 1143 (2012).

59. Uren, C. *et al.* Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. *Genetics* **204,** 303–314 (2016).

60. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466,** 238–242 (2010).

61. Behar, D. M. *et al.* No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. *Hum. Biol.* **85,** 859–900 (2013).

62. Cardona, A. *et al.* Genome-Wide Analysis of Cold Adaptation in Indigenous Siberian Populations. *PLoS One* **9,** e98076 (2014).

63. Di Cristofaro, J. *et al.* Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. *PLoS One* **8,** e76748 (2013).

64. Fedorova, S. A. *et al.* Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol. Biol.* **13,** 127 (2013).

65. Kovacevic, L. *et al.* Standing at the Gateway to Europe - The Genetic Structure of Western Balkan Populations Based on Autosomal and Haploid Markers. *PLoS One* **9,** e105090 (2014).

66. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).

67. Metspalu, M. *et al.* Shared and Unique Components of Human Population Structure and Genome-Wide Signals of Positive Selection in South Asia. *Am. J. Hum. Genet.* **89,** 731–744 (2011).

68. Migliano, A. *et al.* Evolution of the Pygmy Phenotype: Evidence of Positive Selection from Genome-wide Scans in African, Asian, and Melanesian Pygmies. *Hum. Biol.* **85,** (2013).

69. Pierron, D. *et al.* Genome-wide evidence of Austronesian–Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 936–941 (2014).

70. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505,** 87–91

(2014).

71. Rasmussen, M. *et al.* An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. *Science* **334,** 94–98 (2011).

72. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463,** 757–762 (2010).

73. Reich, D. *et al.* Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89,** 516–528 (2011).

74. Yunusbayev, B. *et al.* The Caucasus as an Asymmetric Semipermeable Barrier to Ancient Human Migrations. *Mol. Biol. Evol.* **29,** 359–365 (2012).