Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis

Chengzhen L. Dai,[1] Mohammad M. Vazifeh,[2] Chen-Hsiang Yeang,[3] Remi Tachet,[2] R. Spencer Wells,[4] Miguel G. Vilar,[5] Mark J. Daly,[6,7,8,9] Carlo Ratti,[2,10] and Alicia R. Martin[7,8,9,10,*]

The population of the United States is shaped by centuries of migration, isolation, growth, and admixture between ancestors of global origins. Here, we assemble a comprehensive view of recent population history by studying the ancestry and population structure of more than 32,000 individuals in the US using genetic, ancestral birth origin, and geographic data from the National Geographic Genographic Project. We identify migration routes and barriers that reflect historical demographic events. We also uncover the spatial patterns of relatedness in subpopulations through the combination of haplotype clustering, ancestral birth origin analysis, and local ancestry inference. Examples of these patterns include substantial substructure and heterogeneity in Hispanics/Latinos, isolation-by-distance in African Americans, elevated levels of relatedness and homozygosity in Asian immigrants, and fine-scale structure in European descents. Taken together, our results provide detailed insights into the genetic structure and demographic history of the diverse US population.

Introduction

The United States population is a diverse collection of global ancestries shaped by migration from distant continents and admixture of recent migrants and Native Americans. Throughout the past few centuries, continuous migration and gene flow have played major roles in shaping the diversity of the US. Mixing between groups that have historically been genetically and spatially distinct have resulted in individuals with complex ancestries, while within-country migration has led to genetic differentiation.[1–13]

Deeply characterizing population history is important for understanding human evolution and demographic history, as well as for adequate study design when associating genotypes to phenotypes.[14–17] Earlier population genetic studies in the US broadly characterized this structure, typically using a limited set of ancestry-informative markers or uniparental mtDNA and Y chromosome DNA data.[18] As the cost of genetic technologies have dropped, more recent studies have inferred population history with more complete genome-wide data, typically using more than 100,000 SNPs ascertained via sequencing or genotyping.

Previous genetic studies of the US population have sought to infer genetic ancestry and population history primarily in European Americans, African Americans, and Hispanics/Latinos.[7–9,19,20] European American ancestry is characterized by substantial mixing between different ancestral European populations and, to a lesser extent, admixture with non-European populations.[6,8,9] Isolation among certain European population, such as Ashkenazi Jewish, French Canadian, and Finnish populations, have also resulted in founder effects.[21–24] The mixing of European settlers with Native Americans has contributed to large variations in the admixture proportions of different Hispanic/Latino populations.[1,4,9] Among Hispanics/Latinos, Mexicans and Central Americans have more Native American ancestry; Puerto Ricans and Dominicans have more African ancestry; and Cubans have more European ancestry.[1,4] In African Americans, proportions of African, European, and Native American ancestry vary across the country and reflect migration routes, slavery, and patterns of segregation between states.[2,3,7,9,25] Although much effort has been made to understand the genetic diversity in the US, fine-scale patterns of demography, migration, isolation, and founder effects are still being uncovered with the growing scale of genetic data, particularly for Latin American and African descendants with complex admixture histories.[26,27] At the same time, there has been little research on the population structure of individuals with East Asian, South Asian, and Middle Eastern ancestry in the US.

Many previous studies have investigated specific population histories in the US at relatively small scales—on the order of hundreds to thousands of individuals. These studies have provided deep insights into many specific populations, with some well-powered to infer population history across a breadth of ancestries. Some of these

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [2]Senseable City Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [3]Institute of Statistical Science, Academia Sinica, Nankang, Taipei, Taiwan; [4]Insitome, Inc, Austin, TX 78701, USA; [5]Genographic Project, National Geographic Society, Washington, DC 20036, USA; [6]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; [7]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; [8]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [9]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[10]These authors contributed equally to this work
*Correspondence: armartin@broadinstitute.org
https://doi.org/10.1016/j.ajhg.2020.02.002.

insights have been made by applying methods that are computationally tractable only at smaller scales.[28,29] More recently, however, important insights highlight the need for broader and more comprehensive investigations of population history. For example, recent studies have shown that population structure is inaccurately captured in small sample sizes.[14,17] Additionally, millions of Americans have been interested enough in their genetic ancestry to pay direct-to-consumer companies for individual-level genetic ancestry reports.[8,9] The reliability of these reports is high for many individuals, but they are dependent on (1) the representativeness of their reference panel or customer database, (2) completeness and accuracy of multigenerational birth origin data, and (3) the application of multiple approaches to gain holistic insights into population history.

In this study, we comprehensively evaluate the population history of more than 32,000 genotyped individuals in the US who partook in the National Geographic Genographic Project, a not-for-profit public participation research initiative to study human migration history. This project has several distinct advantages compared to other large-scale population genetics datasets. Participants were genotyped with the GenoChip, a validated array of ~150,000 markers designed for genetic anthropology that excludes medically related SNPs to protect the health privacy of participants.[30] Individual-level genetic data are accessible to researchers around the world to answer anthropological questions. Additionally, most participants report birthplace and ethnicity data for themselves, their parents, and their grandparents, enabling fine-scale insights into recent history. Furthermore, participants report their postal code when they participated in the study, enabling analysis of intragenerational migration. These data therefore enable high spatiotemporal resolution into historical migration patterns. While these trends are consistent with US history at the population scale, we note that genetic ancestry patterns are not commensurate with individual-level ethnicity (i.e., cultural identity).

Here, we leverage these advantages over existing data to identify patterns of genetic ancestry by studying pairwise sharing among the project participants. We combine these comparative patterns with ancestral birth origin records and geographic information to uncover recent demographic and migration trends. By comprehensively analyzing these data to learn about recent migration events, we gain deeper insights into ancestral origins than in many existing studies, especially into Latin America. We also provide early insights into Asian Americans often ignored in genetic studies of the US, including South Asians, East Asians, and Middle Easterners. We also identify detailed patterns among European and African American populations, recapitulating some similar trends reported previously. Taken together, we use accessible individual-level genetic and birth record data to provide insights into the ancestral origins and complex population histories in the US.

Material and Methods

Human Subjects

The Genographic Project and Geno 2.0 Project received full approval from the Social and Behavioral Sciences Institutional Review Board (IRB) at the University of Pennsylvania Office of Regulatory Affairs on April 12, 2005. The IRB operates in compliance with applicable laws, regulations, and ethical standards necessary for research involving human participants. All DNA samples included in this study came from customers of the National Geographic Genographic Project, who have consented to have their results used in scientific research. To participate in the Genographic Project, participants would first order a DNA Ancestry Kit through the Genographic Project website. To ensure anonymity, each DNA Ancestry Kit is encoded with a randomly generated, nonsequential, Genographic Participant ID number. Prior to providing a sample, participants must read an IRB-approved consent form and provide written consent. Participants would then give a saliva sample, mix the saliva sample with a stabilization buffer solution, and return it along with their completed consent form via postal mail. DNA is then extracted from the saliva sample and genome-wide genotyping was performed (Genotyping and Quality Control). Once participants obtain their results, they can voluntarily provide an additional separate consent on the Genographic Project website to make their genotype data anonymously available for qualified anthropological and genetic research.

In addition to providing a DNA sample, participants also provided geographic location (postal code) data and, optionally, family history information in the form of ancestral birth origin and ethnicity (up to grandparental level). All data of individuals who consented to research were deidentified prior to its inclusion in the Genographic research database. We limited our study to those individuals who provided valid geographic locations in the United States. Approximately 75% of individuals selected provided complete pedigrees and family history data (see Supplemental Material and Methods for further detail).

Genotyping and Quality Control

Participants of the Genographic project were genotyped with the GenoChip array,[30] an Illumina iSelect HD custom genotyping bead array with approximately 150,000 markers that are Ancestry Informative Markers. It excludes markers that are medically related to protect the health privacy of participants and minimize the improper translation of direct-to-customer genetic ancestry results to clinical care.[31] The ability of the Genochip array to discern subpopulations was validated by producing concordant ancestry patterns with samples from the 1000 Genomes Project and demonstrating similar F_{ST} distributions and higher mean F_{ST} values when compared to the Affymetrix Axiom Human Origins array (used in HGDP-CEPH) and the Illumina Human660W-Quad BeadChip.[30] Raw genotype data were quality controlled (QC) using PLINK v1.90b3.39.[32] We filtered to keep samples with ≤ 0.1 missingness, sites with $= 0.0$ missingness, and MAF ≥ 0.05. A total of 32,589 individuals and 108,003 SNPs passed quality control.

Ancestry Reference Panels

We leveraged a variety of reference populations to help better infer and interpret the genetic ancestry, admixture proportions, and population structure in the Genographic cohort. Data from the 1000 Genomes Project was used to help identify genetic ancestry and estimate admixture proportions.[1] 108,003 SNPs were shared

between the Genographic samples and the 1000 Genomes Project samples. We also used data from the Population Reference Sample (POPRES) to help understand the population structure of individuals with European ancestry in the Genographic cohort.[33] All analysis with the POPRES data was limited to the 46,710 SNPs that are shared between the two datasets. We also leveraged recently released sequence data for the Human Genome Diversity Project (HGDP) to expand the available set of ancestral populations from Asia.[34] All analyses using the HGDP data was performed using the 105,944 SNPs shared between the samples in Genographic Project and HGDP.

Principal Component Analysis

We performed principal component analysis (PCA) on the quality-controlled samples using FlashPCA v.2.0.[35] For PCA of all Genographic Project individuals, we used the genotypes of all 2,504 individuals from the 1000 Genomes Project as reference samples. We first computed PCs across the 108,003 shared sites for 1000 Genomes Project individuals. We then projected the Genographic Project individuals on the same principal component space using the flag: --project.

For PCA analyses of East Asian and South Asian populations, we used samples from 1000 Genomes Project that correspond to the East Asian and South Asian super populations. Similar to above, we first compute PCs for the 1000 Genomes Project samples separately for East Asians and South Asians. We then projected East Asian and South Asian Genographic Project individuals onto the respective principal component space using: --project.

Continental Ancestry Assignment

We assigned continental ancestry to each Genographic sample by using a random forest classifier. Using the PCs and known super population assignments (AFR, African; EUR, European; EAS, East Asian; AMR, admixed American; and SAS, South Asian) from the 1000 Genomes Project samples as training data, we applied the classifier to assign ancestry to each Genographic sample at 90% probability. We considered unassigned ancestries as "other" (OTH).

Comparison of Continental Ancestry Assignment with Self-Reported Data

To evaluate the concordance between continental ancestry assignments based on genetics and self-reported ethnicity, we standardized self-reported ancestral ethnicities and estimated the proportion of assigned individuals within each continental ancestry groups that have at least one grandparent with a matching continental ancestry. Since ancestral ethnicity data were provided in the form of free text and was therefore not standardized across participants, we manually cleaned and mapped the reported ethnicities to continent level ancestries. For example, African ancestry can include a country (e.g., Jamaican, Nigerian, Cape Verdean), an ethnic group (e.g., Amhara or Tigray from Ethiopia), a historical term used to describe African descendants in America (e.g., Melungeon, Maroon, Mulatto), or the commonly used terms of African American or Black.

Genetic Ancestry Proportion Estimation

We estimated admixture proportions using ADMIXTURE by first analyzing samples from the 1000 Genomes Project in unsupervised mode to learn allele frequencies.[36] Then, we projected the learned allele frequencies onto the Genographic samples to obtain the admixture proportions using the flag: -P. We ran ADMIXTURE with K = 2–9 and chose K = 5 as the most stable representation based on cross-validation.

For the analysis of East Asian and South Asian, we combined samples from HGDP and 1000 Genomes Project together to build more comprehensive reference panels. Specifically, we combined 1000 Genomes Project populations under the East Asian (EAS) super population label with HGDP samples that have the East Asian and Oceania region label, and we combined 1000 Genomes Project samples under the South Asian super population label with Central South Asia labeled populations in HGDP. Similar to above, we first ran ADMIXTURE on the ancestral reference panels for East Asians and South Asians, separately. We then projected the learned allele frequencies onto the Genographic samples to obtain admixture proportions using the flag: -P. We tested a variety of clusters, K = 2–9, and chose K = 4 for East Asians and K = 3 for South Asians as the most stable representations.

UMAP

We applied the Uniform Manifold Approximation and Projection (UMAP) method to visualize subcontinental structure.[37,38] We first combined the PCs of the Genographic samples and the 1000 Genomes Project samples into one dataset. We then applied UMAP on the first 20 PCs from the joint dataset to produce a two-dimensional plot. We tested various parameter choices for UMAP and found that the default nearest neighbor value of 15 and the minimum distance values of 0.5 delivered the clearest result. Coloring of UMAP plots are described in the Supplemental Material and Methods.

We further examined the subcontinental structure of Genographic Project individuals who were classified as European ancestry individuals with data from the Population Reference Sample (POPRES).[33] Similar to the analyses with the 1000 Genomes Project data, we performed dimensionality reduction with PCA and UMAP, keeping the same parameter values. Coloring of POPRES data was grouped by continental regions: Southeast Europeans = Croatia, Yugoslavia, Bosnia-Herzegovina, Serbia, Romania, Hungary, Albania, Macedonia; Central Europe = Switzerland, France, Germany, Germany, Swiss-Italian, Belgium, Swiss-French, Netherlands, Swiss-German; British Isle = Scotland, Ireland, United Kingdom; South Europe = Italy, Cyprus, Turkey, Greece; Iberian = Portugal, Spain; Eastern Europe = Austria, Czech Republic, Poland, Russia; Scandinavia = Sweden, Norway.

Phasing and Haplotype Estimation

Genographic genotypes were phased with the Sanger Imputation Service using EAGLE2[39] and the Haplotype Reference Consortium reference panel.[40] No genotype imputation was performed.

fineSTRUCTURE Analysis

For classified East Asian individuals and South Asian individuals, we inferred clusters of unrelated individuals with shared ancestries by applying the fineSTRUCTURE framework v.4.0.1, a model-based approach to estimate patterns of haplotype similarity and identify clusters of discrete populations.[29] We performed fineSTRUCTURE analysis separately for the two populations. The first part of the fineSTRUCTURE framework uses ChromoPainter to measure shared ancestry between individuals and estimate a coancestry matrix. This matrix is then used in fineSTRUCTURE's clustering and tree-building algorithm to hierarchically cluster individuals from fine levels of structuring to broader levels. We first applied ChromoPainter to phased genotypes to estimate the number of contiguous segments (chunks) shared and total amount of

genome (in cM) shared between each pair of individuals within each population, as well as the normalization parameter (*c*). Using the coancestry matrix and normalized parameter, we then ran the fineSTRUCTURE with 2 million Markov Chain Monte Carlo (MCMC) iterations, of which 1 million are "burn-in" iterations, and every 2,000 iterations was sampled. Finally, we used fineSTRUCTURE to infer a hierarchical tree using 100,000 hill-climbing moves. We used the scripts accompanying the fineSTRUCTURE software as well as the *ape* package in R to visualize the coancestry matrix and dendrogram results.

To examine the properties of the inferred clusters, we sought to examine structure at both the broad-scale and fine-scale. There is no definitively correct level of the dendrogram to pick for examination. We examined clades at various levels of the tree and assessed broad structure at the levels in which clades had sufficient number of individuals (on average 50 or more samples). We further used a combination of PCA and analysis of ancestral origins to assess and define these clades. Some of the clusters are small but genetically distinct as evident by the branch length and height of the split (i.e., Girmitiyas, Bangladesh), and therefore, they were kept as separate clades.

Unlike traditional PCA, PCA using the coancestry matrix (i.e., chunk counts matrix) can better discern fine-scale population structure and provide greater interpretability.[29] We performed PCA on the chunk counts matrix using in the Python library *scikit-learn*. Individual markers are colored and labeled based on their respective grouping.

Estimating Effective Migration Surfaces

We estimated migration and diversity relative to geographic distance using the estimating effective migration surfaces (EEMS) method for Genographic Project individuals that were classified under African, European, and admixed American ancestries.[41] We excluded East Asian and South Asian ancestries due to low sample size and density. We used unrelated individuals with available postal code data. We first computed pairwise genetic dissimilarities with the EEMS *bed2diffs* tool and then ran EEMS with *runeems_snps*, setting the number of demes to 250 and to 500. Per the recommendation in the manual, we adjusted the variance for all proposed distributions of diversity, migration, and degree-of-freedom parameters such that all were accepted 10%–40% of the time. We increased the number of Markov chain Monte Carlo (MCMC) iterations until it converged.

To evaluate the robustness of EEMS to sampling bias, we simulated three different sampling schemes. We used individuals classified with African ancestry as it is the smallest of the three ancestries and therefore more likely to be impacted by sampling bias. In the first sampling scheme, we randomly subsampled individuals to 80% of the original sample size. In the second scheme, we used the US Census Regions assignments for states and explored the impact of even sampling across the four major Census Regions. We subsampled African Americans so that each Census Region was represented in equal proportions. In the last sampling scheme, we explored the scenario of overrepresentation in the South by subsampling at 80% but this time with half of the subsamples being from the South and the remaining samples are evenly distributed across the three regions.

Haplotype Calling and Network Construction

We used IBDSeq v.r1206 to generate shared identity-by-descent (IBD) segments from genotype data for all unrelated individuals.[42] Unlike other IBD detection algorithms, IBDseq does not rely on phased genotype data and is less susceptible to switch errors in phasing that can cause erroneous haplotype breaks. We filter for IBD segments greater than 3 cM. We removed segments that overlapped with long chromosomal regions (1 Mb) that had no SNPs across all unrelated individuals. These sites can result in false positives IBD sharing and likely correspond to centromeres and telomeres. We calculate the cumulative IBD sharing between individuals by summing the length of all shared IBD segments. We then constructed a haplotype network of unrelated individuals by defining vertices an individuals and edge weights between vertices as the cumulative IBD sharing between individuals. We filtered for edges with cumulative IBD sharing is $\geq$12 cM and $\leq$72 cM, as previously described.[8]

Detection of IBD Clusters

While fineSTRUCTURE can identify population structure in admixed cohorts using haplotype similarity,[28] fineSTRUCTURE does not scale to large sample sizes and is not recommended for samples >10,000.[29] We therefore sought to identify clusters of related individuals in the haplotype network using the Louvain Method implemented in the igraph package for R. The Louvain Method is a greedy iterative algorithm that assigns vertices of a graph into clusters to optimize modularity (a measure of the density of edges within a community to edges between communities).[43] The Louvain Method begins by first assigning each node as its own community and then adds node *i* to a neighbor community *j*. It then calculates the change in modularity and places *i* in the community that maximizes modularity. The algorithm repeats this continuously and terminates when no vertices can be reassigned.

We partitioned the haplotype network into clusters by recursively applying the Louvain Method within subcommunities. At the highest level, we take the full, unpartitioned haplotype graph and identify a set of subcommunities. We isolate the vertices within each subcommunity, keeping only the edges between those vertices to create separate new networks. We then apply the Louvain Method to the new subgraphs. We repeat this process up to four levels. We combined subcommunities with low genetic divergence based on F_{ST} values of < 0.0001.

Annotation of IBD Clusters

We used a combination of ancestral birth origins and self-reported ethnicities to discern demographic characteristics of each cluster. For each cluster, we quantified the proportion of each birth origin (i.e., country of origin) among all four grandparents, treating each grandparent's origin equality. We use these proportions to inform population labels. Clusters in which a single non-US birth origin was in high proportions was labeled with that country. In cases where multiple non-US birth locations exists in approximately equally high proportions, we assigned a label representing the broader region (e.g., Eastern Europeans for Poland, Lithuania, Ukraine, and Slovakia; East Asia for Japan, China). For certain clusters, annotations could not be easily discerned by birth origin data. In these cases, we relied on self-reported ethnicities to label the clusters as these populations were found to be less associated with a non-US country (e.g., Ashkenazi Jews) or the population has resided in the US for generations (e.g., African Americans, Acadians).

Runs of Homozygosity

We used PLINK v.1.90b3.39 to infer runs of homozygosity with a window of 25 SNPs.[32] By default, PLINK reports only the runs of

homozygosity with lengths ≥1 Mb. For each individual, we calculated the sum of runs of homozygosity (sROH) by summing the lengths of homozygous segments. We compared ROH segments inferred by PLINK with homozygosity-by-descent (HBD) segments inferred using IBDSeq. The two approaches largely agreed in ROH lengths (Spearman correlation = 0.94; $p = 7.24 \times 10^{-12}$), with the exception that the median sROH lengths for the Greece-Italy and Italy clusters were lower in IBDSeq while the median sROH length for East Asians were higher in IBDSeq when compared to PLINK (Table S1).

Local Ancestry Inference

We inferred local ancestry with RFMix v.1.5.4 for Genographic samples in clusters that were annotated as Hispanics/Latinos and African Americans.[44] We used samples of African (LWK, MSL, GWD, YRI, ESN, ACB, and ASW; n = 661), European (CEU, GBR, FIN, IBS, and TSI; n = 503), and Native American (MXL, PUR, CLM, and PEL; n = 347) ancestry from the 1000 Genomes Project to build the reference panel for classifying genomic segments. We ran RFMix with the default minimum window size (0.2 centimorgans, cM) and a node size of 5 with the flags: -w 0.2, -n 5. We then collapsed the output of RFMix, which denotes the classified ancestry of each SNP for each individual, into local ancestry segments/tracts (in cM) for each individual. We then derived global ancestry proportions for each individual using that individual's local ancestry tracts; we summed the length of local ancestry tracts for each ancestry (EUR, AFR, AMR) dividing by the total length of the genome to get the global proportion of each ancestry. Global ancestry proportions were visualized using the *python-ternary* package in Python (see Web Resources).

Genetic Divergence

We computed weighted Weir-Cockerham F_{ST} estimates for each pair of haplotype clusters using PLINK v.1.90b3.39.[32] Using the distance matrix of F_{ST} values between clusters, we constructed an unrooted phylogenetic tree using the neighbor joining method implemented in *scikit-bio* (see Web Resources). We visualized the tree using Interactive Tree of Life (see Web Resources).

Effective Population Size

We estimated effective population size with IBDNe.[45,46] Using the inferred IBD segments between individuals for each cluster, we ran IBDNe in default mode separately for each cluster to infer the effective population size over time along with confidence intervals.

Results

Genetic Ancestry and Diversity across the United States

To assess the diversity of ancestries among individuals in the Genographic Project, we first performed principal component analysis, projecting Genographic samples into the same principal component (PC) space as that of the 1000 Genomes Project samples (Figures 1A–1C, S1, and S2).[35,36] Since self-reported ancestry was not consistently provided across all Genographic Project individuals, we leveraged the 1000 Genomes Project data to assign continental ancestry to each Genographic sample (Material and Methods). We first trained a Random Forest classifier on the first 10 PCs of the 1000 Genomes Project samples with super population groupings as ancestry labels (EUR, European; AMR, Admixed American; AFR, African; EAS, East Asian; SAS, South Asian). We then used the trained model to assigned continental ancestry to each individual in the Genographic cohort at > 90% confidence. A total of 3,028 individuals (9.3% of total) did not meet the classification threshold (Table S2). The inability to classify these individuals may be due to variable levels of admixture not reflected in the 1000 Genomes reference populations. No particular bias was found in the ancestral birth origin records for these individuals, as the top non-US origins are Germany (3.0%), Italy (2.6%), Poland (2.5%), UK (2.5%), and Mexico (2.0%). Overall, the assigned continental ancestry was largely consistent with the self-reported ancestral ethnicity, as 95% of classified African-ancestry individuals and 85% of classified Hispanic-ancestry individuals who reported ancestral data had at least one grandparent of that ancestry (Material and Methods).

Regional differences in genetic ancestry correspond to historical demographic trends. We evaluated the distributions of classified individuals across the four designated US Census regions: South, Northeast, Midwest, and West (Table S2). Classified individuals of European descent make up the majority (78.5%) of the Genographic cohort and are the most prevalent in the Midwest (82.8% of individuals in the Midwest; $p < 0.01$, Fisher's exact test; Table S2). Admixed American ancestry individuals are most prominent in the West and South (9.7% and 7.8% of total individuals in the West and South, respectively; $p < 0.05$, Fisher's exact test). Individuals classified as having African ancestry are most common in the South (3.2%), followed by the Northeast (3.0%). East Asians mostly reside in the West (2.1%), while South Asians are most abundant in the Northeast (1.0%). While the proportion of individuals classified as of European descent in the Genographic cohort (78.5%) are similar to the proportions of individuals reported as "White" in the US Census Data (76.1%; Table S3), we note that genetic ancestry is not a direct measure of ethnicity and race, and the two are not fully comparable (Supplemental Material and Methods). The large proportion of unclassified individuals also hinders our ability to properly compare the Genographic cohort to the US Census and understand how representative the Genographic cohort is of the US population. Overall, the distribution of Genographic Project participants by state reflects the US population distribution reported in the Census (Spearman's $\rho = 0.91$, $p = 1.5 \times 10^{-20}$; Figure S3). However, the states of Washington, California, Virginia, Maryland, and Colorado have higher proportions (>1% difference) of participants when compared to the US population distribution while Texas and Ohio have lower proportions of participants (Table S4). For certain ancestries, some ascertainment bias exists. For example, individuals with African ancestry are overrepresented in California but are absent in Idaho, Maine, Nebraska, North Dakota, South Dakota, and Wyoming.

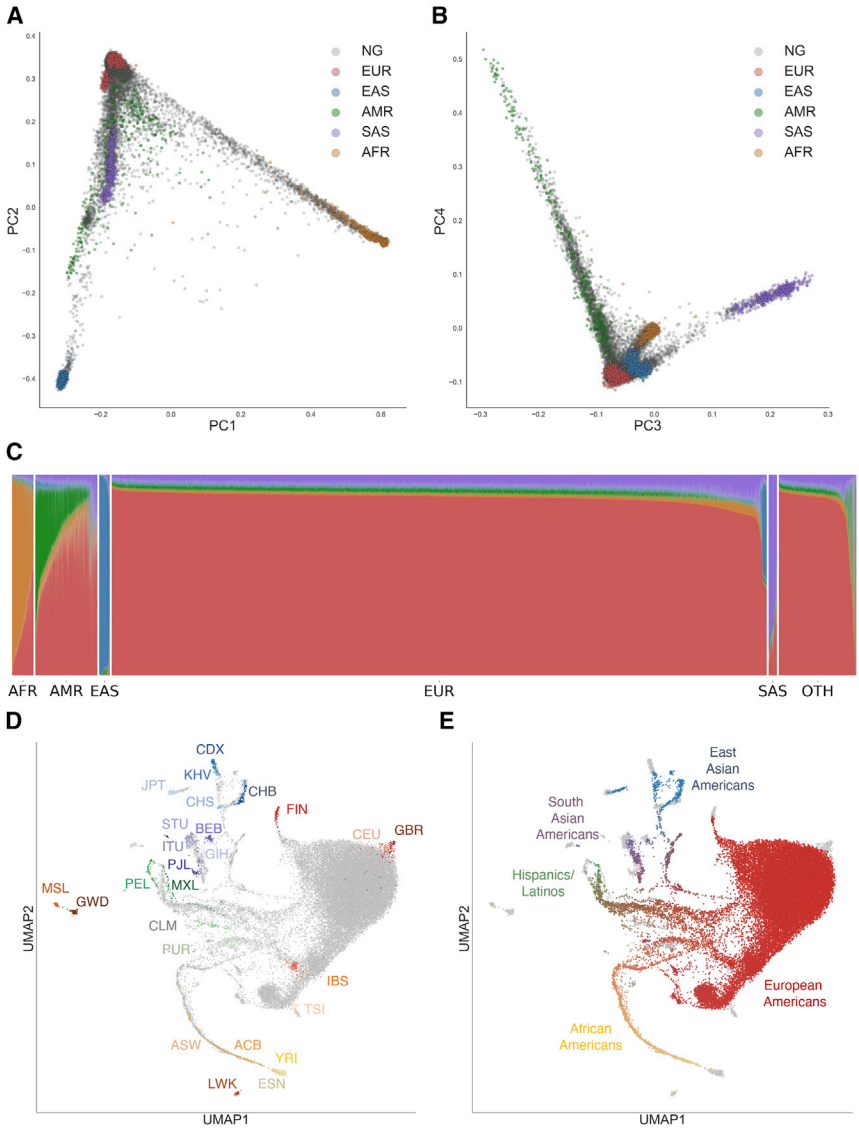


Figure 1. Genetic Diversity of the US Population

(A) Principal Components Analysis (PCA) of individuals in the United States and in the 1000 Genomes Project. Each individual is represented by a single dot. Individuals in this study (NG) are colored in gray while 1000 Genomes Project individuals are colored by super population (EUR, red; AFR, yellow; AMR, green; EAS, blue; SAS, purple). Principal components (PC) 1 and PC 2 are shown.

(B) Similar to (A), except principal components (PC) 3 and PC 4 are shown.

(C) ADMIXTURE analysis at K = 5 of individuals in this study. Each individual was assigned a continent-level ancestry label using a random forest model trained on the super population labels and the first 10 PCs of the 1000 Genomes Project dataset. OTH, individuals who did not meet the 90% confidence threshold for classification.

(D and E) UMAP projection of the first 20 PCs. Each dot represents one individual. In (D), individuals in the 1000 Genomes Project are colored by population while US individuals from the National Geographic Genographic Project are in gray. In (E), 1000 Genomes Project individuals are colored in gray while US individuals from the National Geographic Genographic Project are colored based on their admixture proportions from ADMIXTURE. The color for each dot was calculated as a linear combination of each individual's admixture proportion and the RGB values for the colors assigned to each continental ancestry (EUR, red; AFR, yellow; NAM, green; EAS, blue; SAS, purple). See Material and Methods for specific population labels.

To uncover population substructure, we performed dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP) on the first 20 PCs of a combined Genographic and 1000 Genomes Project dataset.[37,38] By leveraging multiple PCs at once, UMAP can disentangle subcontinental structure (Figures 1D, 1E, S4, and S5). Similar to a previous analysis,[38] populations in the 1000 Genomes Project form distinct clusters corresponding to ancestry and geography. The Genographic Project individuals project into several clusters, overlapping with the 1000 Genomes Project clusters. Consistent with the PCA and ADMIXTURE analysis, the largest clusters correspond to European ancestry and cluster closely with the 1000 Genomes CEU and GBR populations (CEU = Utah Residents with Northern and Western European Ancestry, GBR = British in England and Scotland).

While UMAP is a visualization tool with no direct interpretation of genetic distance, the continuum of points connecting UMAP clusters reflects the varying degrees of estimated admixture between different continental ancestries. In particular, the complex population structure of Hispanics/Latinos is shown by the points spanning between the clusters of European, Native American, and African ancestry. Coloring of these points based on ancestry proportions affirms the relationship between the degree of admixture and their relative position between reference clusters. Interestingly, African American individuals from both datasets form a single continuum from the European cluster to the Yoruba (YRI) and Esan (ESN) populations of Nigeria in the 1000 Genomes Project, indicative of the West African origins of most African Americans. This observation is consistent with and further expands the previous finding that the African tracts in the admixed 1000 Genomes Project populations of ACB and ASW are similar to the Nigerian YRI and ESN populations.[2,47]

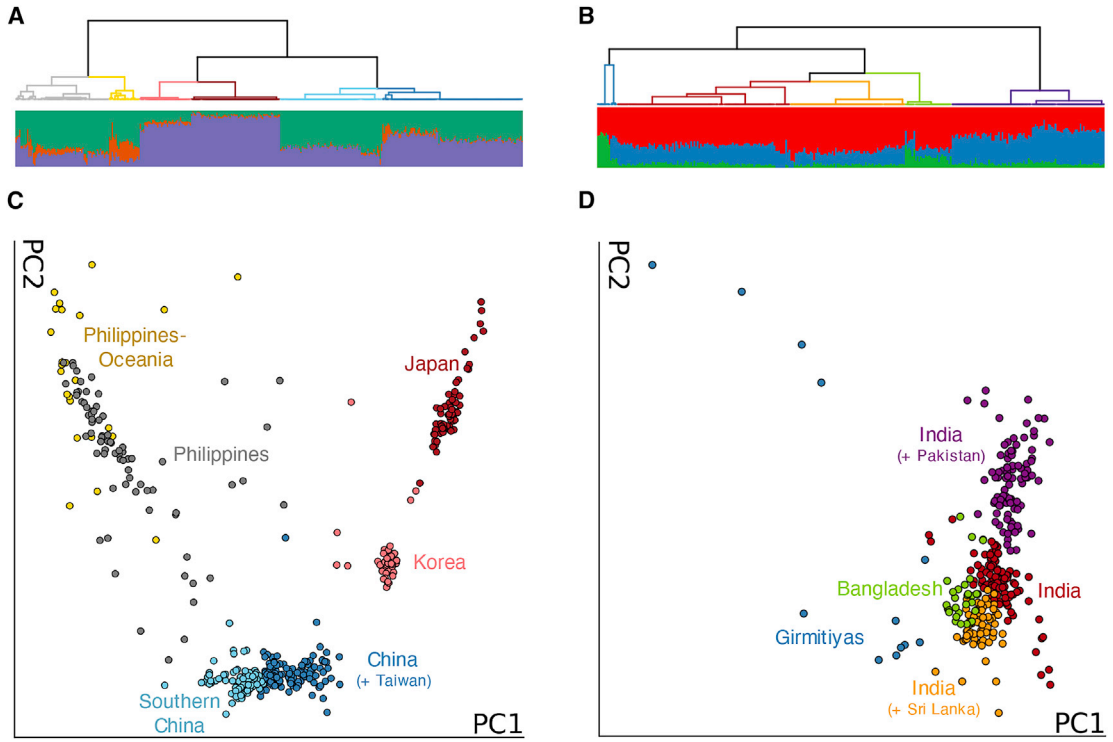


Figure 2. Population Structure of East Asian and South Asian Individuals in the US
(A and B) fineSTRUCTURE dendrogram showing the hierarchical relationship between clusters inferred using the genotypes of classified East Asian individuals (A) and South Asian individuals (B). Branch colors represent clades with shared ancestral origins. The admixture proportion of each individual is displayed as a bar plot in the corresponding position below the dendrogram. The number of ancestral populations, K, is four for East Asians (A) and three for South Asians (B).
(C and D) Principal component analysis (PCA) of the fineSTRUCTURE co-ancestry matrix. Each individual (point) corresponds to a Genographic Project individual classified as either East Asian (C) or South Asian (D). The color of each point corresponds to a clade in the fineSTRUCTURE dendrogram shown in (A) and (B).

Fine-Scale Structure among US Individuals of Asian Ancestry

Existing genetic studies of the US population have largely overlooked East Asian and South Asian populations, likely due to their underrepresentation in datasets. We therefore explored the population structure of Genographic Project individuals classified as East Asians and South Asians. We used fineSTRUCTURE to first estimate patterns of haplotype similarities between individuals, taking into consideration linkage disequilibrium, and then hierarchically clustered individuals based on these patterns of shared ancestry to identify clusters of populations and their relationships.[29] We applied fineSTRUCTURE to unrelated individuals in each population and inferred a total of 40 East Asian clusters (Figure 2A) and 26 South Asian clusters (Figure 2B). These clusters further organized into clades on the tree to reveal broader genetic structure. To visualize these structures, we performed PCA on the fineSTRUCTURE coancestry matrix. Compared to traditional PCA, distinctions between groups of individuals were clearer with fineSTRUCTURE PCA, particularly at the broader levels of genetic differentiation (Figures 2C and S6A; Figures 2D and S7A; Material and Methods). We also estimated subcontinental admixture proportions with ADMIXTURE using the East Asian and South Asian populations in the 1000 Genomes Project and the Human Genome Diversity Project (HGDP) as reference populations (Figures S6B, S6C, S7B, and S7C). Finally, we leveraged data from individuals who provided grandparental birth origin to help annotate and interpret the clusters and clades.

The patterns of shared ancestry among these US individuals capture the genetic diversity of East Asia and South Asia. The East Asian clusters broadly organize into six major clades, reflecting the different countries of ancestral origin (Figure 2A). At the highest level of genetic differentiation (top level of the hierarchical tree), individuals from Southeast Asia separate from East Asians. This Southeast Asian clade is predominantly represented by Filipinos with a branch of individuals with more Oceanic origins (shown in gray and yellow, respectively). Admixture proportions vary among the Southeast Asian individuals, likely due to the large number of ethnolinguistic groups that are found in the Philippines and neighboring islands. The East Asian clade further separates into individuals of Chinese descent (light blue and dark blue) and those from Japan (dark red) and Korea (light red). While the two Chinese-related groups share a branch on the tree, Taiwanese ancestral origins are more prevalent in one of the groups (dark blue), the "China (+ Taiwan)" group,

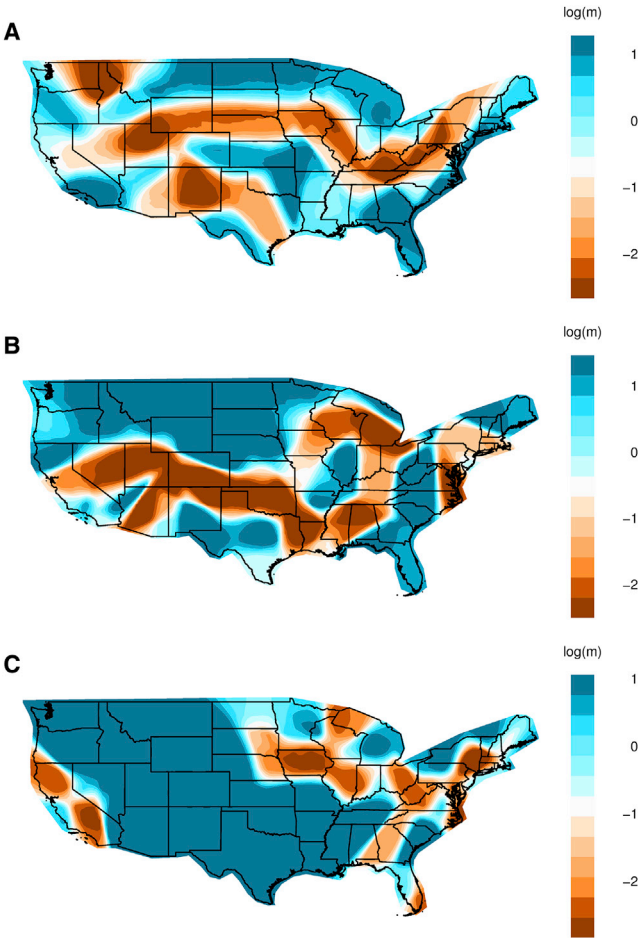


Figure 3. Effective Migration Rates of African Americans, Hispanics/Latinos, and Europeans within the United States.
Migration rates inferred with EEMS for African Americans (A), Hispanics/Latinos (B), and Europeans (C). EEMS models the relationship between genetics and geography by assessing the decay of genetic similarity with respect to geographic distance. Colors and values correspond to inferred rates, *m*, relative to the overall migration rate across the country. Shades of blue indicate higher migration (i.e., log(m) = 1 represents effective migration that is 10-fold faster than the average) and higher levels of genetic similarity while shades of orange indicate migration barriers and lower levels of genetic similarity.

while the other group (light blue), labeled "Southern China," also contains some individuals from Laos and Vietnam. Lower levels of hierarchy did not differentiate these ancestral origins into separate groups. PCA and ADMIXTURE analysis for these two groups show that the China (+ Taiwan) cluster resembles the Han Chinese (CHB) population in the 1000 Genomes Project while the Southern China group resembles the Southern Han Chinese (CHS) population (Figure S6). Among the South Asian individuals, we observed genetic differentiation between individuals with ancestral origins from India, reflecting the diverse population structure previously observed in India.[1,48] Of the three clades with majority Indian ancestral origin, ancestral origins from Pakistan was observed in the "India (+ Pakistan)" clade, while Sri Lankan ancestral origins were present in the "India (+ Sri Lanka)" clade. Individuals in these two clades resemble the Punjabi from Lahore, Pakistan (PJL) and Sri Lankan Tamil (STU) populations in the 1000 Genomes Project, respectively (Figure S7). Similarly, we also find a clade of individuals with Bangladesh ancestral origins that is similar to the 1000 Genomes Project Bengali from Bangladesh (BEB). Interestingly, we also inferred a small, but genetically distinct "Girmitiyas" clade (N = 12; blue branch in Figure 2B). While the small sample size makes it difficult to accurately assess this clade, we note that many former British colonies (e.g., Trinidad and Tobago, Fiji, Barbados, Guyana) are represented in the ancestral origins of these individuals. We therefore hypothesize that these individuals may potentially be descendants of Girmitiyas, indentured Indian laborers brought to those former colonies.[49]

Population Differentiation and Migration Rate Inference across the United States

Understanding the relationship between genetics and geography can provide insights into demographic history. Previous analyses of this relationship in the US population have primarily compared data aggregated at the state or regional level.[7,9] Such approaches, however, do not capture the fine-scale patterns of genetic similarity that are not influenced by discrete political boundaries. We therefore sought to infer population structure across continuous space with the estimating effective migration surfaces (EEMS) method.[41] EEMS statistically measures effective migration rates by overlaying a dense grid of evenly spaced demes and calculating genetic differentiation (i.e., resistance distance) between neighboring demes. Higher rates of migration are inferred in locations where genetic similarity is high (colored in blue in Figure 3) while lower rates of migration are inferred in locations where genetic similarity is low (colored in dark orange). Areas with low effective migration are also referred to in EEMS as "barriers," which can be intuitively interpreted as regions in which neighboring populations are more genetically dissimilar than expected. In more homogeneous populations, these barriers tend to indicate isolation by distance, while in more heterogenous populations, they may reflect differences in population structure. We applied EEMS to genetically classified Europeans, African Americans, and Hispanic/Latinos across the contiguous 48 states. We excluded East Asians and South Asians due to low sample density.

The inferred migration rates for African Americans reveal genetic signatures of historical demographic events (Figures 3A, S8, and S9). Along the Atlantic coast from the Florida Panhandle to southern Maine, genetic similarity and effective migration rates are relatively high, indicating the constant migration and similar effective population sizes of African Americans in these states. However, we also observe a strong north-south barrier to migration starting along the Appalachian Mountain Range, continuing north up the Mississippi River, and extending

Table 1. Summary of Haplotype Clusters

Cluster	Samples (Count)	Median Sum of ROH (Mb)	Median Cumulative IBD (cM)
Northwest Europe 1	11,725	2.88	15.23
Northwest Europe 2	1,571	2.80	15.15
Ireland	2,137	2.85	15.42
Central Europe	3,116	2.83	15.06
Eastern Europe	2,471	3.16	15.37
Southern Europe	1,626	2.73	14.98
Italy	697	6.91	14.64
Greece-Italy	238	7.28	15.02
Scandinavia	717	3.02	15.54
Finland	314	3.67	17.50
Acadia	249	3.89	19.48
French Canadian	314	2.89	16.60
Ashkenazi Jewish	1,475	11.26	31.75
Admixed Jewish	445	2.75	15.50
Hispanics/Latinos	810	3.53	16.38
Hispanics/Latinos in California	573	4.10	17.11
Hispanics/Latinos in New Mexico	163	5.52	21.92
Hispanics/Latinos in Texas	177	6.27	23.65
Puerto Rico	350	8.01	26.23
African Americans South	761	3.34	19.56
African Americans North	420	2.94	15.90
East Asia	561	3.65	19.63
Southeast Asia	325	8.44	17.90
South Asia	389	10.42	14.82
Greater Middle East	93	9.01	17.16

Sum of runs of homozygosity (sROH) was calculated by summing the lengths of continuous homozygous segments $\geq$1 Mb. Cumulative IBD was determined by summing IBD segments of $\geq$3 cM and filtering for only pairs $\geq$12 cM and $\leq$72 cM. Statistics were determined within haplotype clusters, rather than across the ancestrally heterogeneous and imbalanced full network.

west across the rest of the country. This migration barrier, along with the migration barrier spanning Texas and New Mexico, reveals a pattern of genetic relatedness across geography that is consistent with the Great Migration from the 1910s to the 1960s in which an estimated 6 million African Americans migrated out of the South to cities across the Northeast, Midwest, and West.[7,50] To understand whether this migration barrier is influenced by sampling bias, we subsampled individuals and simulated three different sampling schemes (Material and Methods). We found that the north-south migration barrier was consistently present in all three sampling schemes, confirming that the inferred migration results of EEMS are robust to irregular sampling (Figure S10).[41]

A highly complex pattern of genetic similarity exists among present-day Hispanics/Latinos across the country, capturing regional genetic structure. Across the southwestern states, two regions bordering Mexico—one in California and another extending from New Mexico to Texas—exhibit high levels of genetic similarity and effective migration rates (Figures 3B, S8, and S9). Separated by a migration barrier in Arizona, these two distinct regions likely reflect known differences in the northward migration from east versus west Mexico.[8,51] High genetic similarity and relative rates of effective migration are also observed in Florida and continue northward. However, barriers to migration are observed in states immediately east of the Mississippi River, likely resulting from varying degrees of admixture.

The patterns of genetic similarity for Europeans capture subcontinental structure. With the exception of the states in the Midwest and along the Atlantic coast, elevated levels of genetic similarity and relative migration rates are observed across most of the country. We find low effective migration rates surrounding Minnesota and Michigan, likely due to the genetic dissimilarity of Finnish and Scandinavian ancestry that is abundant in the region (Figures 3C, S8, and S9).[8] We also find reduced migration rates across Ohio, West Virginia, and Virginia, suggesting the existence of genetic differentiation along the Appalachian Mountains. Many of the major cities, such as Washington, DC, Philadelphia, and Miami, also exhibit low genetic similarity, perhaps due to greater genetic diversity and admixture within cities.

Coupling Fine-Scale Haplotype Clusters and Multigenerational Birth Records Uncovers Distinct Subcontinental Structure

To disentangle more recent and subtle population structure, we performed identity-by-descent (IBD) clustering on the Genographic cohort and annotated clusters using multigenerational self-reported birth origin data. We first built an IBD network from pairwise IBD sharing among 31,783 unrelated individuals, where vertices represent individuals and edges represent the cumulative IBD (in centimorgans, cM) between pairs of individuals. We employed the Louvain method, a greedy heuristic algorithm, to recursively partition vertices in the graph into clusters that maximize modularity for each iteration.[8,43] The clusters of individuals resulting from each iteration can be interpreted as having greater amounts of cumulative IBD shared between individuals within the cluster than with those outside of the cluster. To aid in the interpretation of the clusters, we merged clusters with low genetic differentiation ($F_{ST} < 0.0001$), resulting in a final set of 25 clusters (Table 1). We annotated each cluster based on ancestral birth origin and ethnicity data and constructed a neighbor-joining tree based on the F_{ST} values (Figure S11). 98% of the 3,028 individuals that were not classified by our Random Forest model were assigned to a haplotype cluster. No single cluster was overrepresented by unclassified

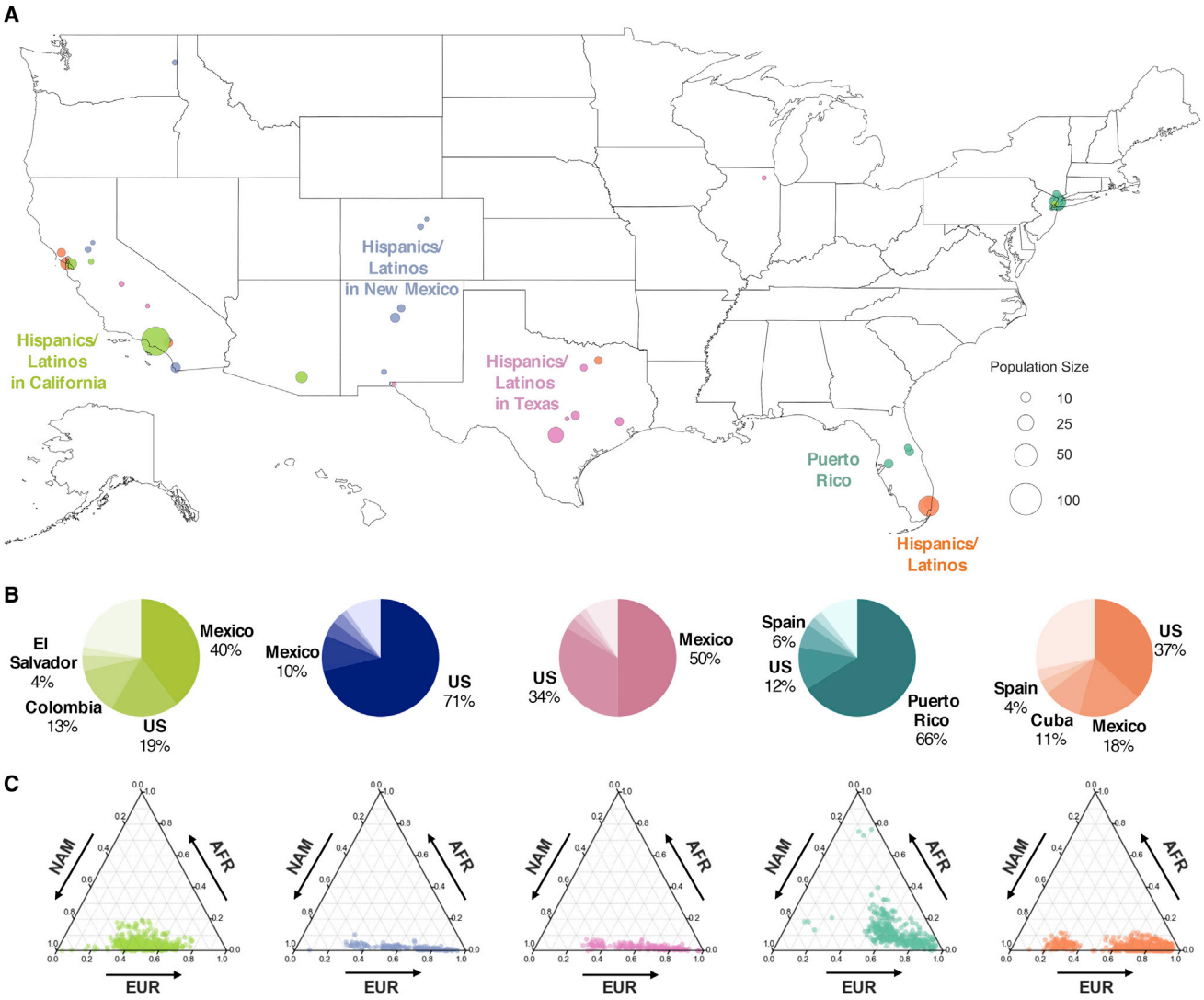


Figure 4. Geographical Distribution of Hispanic/Latino Haplotype Clusters

(A) Each dot corresponds to a county containing present-day individuals and the size of the dot signifies the number of samples of the particular cluster in that county. Only the Hispanic/Latino cluster with the highest odds ratio is shown for each county, and for clarity, only the top ten locations with the highest odds ratio are shown for each cluster. Maps showing the full distribution for each haplotype cluster can be found in the supplement (Figure S15).

(B) Ancestral birth origin proportions of each cluster for individuals with complete pedigree annotations, up to grandparent level. Proportions were calculated from aggregating the birth locations of all grandparents corresponding to members of each haplotype cluster. For each chart, only the top five birth origins are shown as individual proportions; the remaining birth origins are aggregated into one slice (lightest color).

(C) Ternary plots of ancestry proportions based on local ancestry inference for each haplotype cluster. Each dot represents one individual.

individuals, as unclassified individuals comprised of 8%–11% of each cluster.

Genetic and geographic differences are greatest among Hispanic/Latino haplotype clusters. We identified a total of five Hispanic-related clusters (Figure 4). The largest of these cluster ($n = 810$; orange in Figure 4A) is strongly associated with south Florida (OR = 10.4; $p = 2.5 \times 10^{-25}$; Table S5) but is also found in California and Texas (OR ≥ 2; $p < 0.05$; Table S5). No single ancestral birthplace characterizes this cluster, as the US, Mexico, and Cuba each make up more than 10% of the birth origin labels (Figure 4B). Proportions of European ancestry tracts inferred with RFMix[44] are higher in this cluster (mean = 72.7%, SD = 20.4%; Figure 4C) than in the other Hispanic/Latino clusters (mean = 48.0%–67.4%; Figure 4C). Puerto Ricans characterize a substantial proportion of another Hispanic/Latino cluster associated with Florida (OR > 4) as well as New York City (OR > 5). Unlike the other Hispanic clusters, the Puerto Rican cluster shares the same branch on the F_{ST} tree as the African American clusters (Figure S11), likely due to relatively high proportions of African ancestry (mean = 11.2%, SD = 9.0%) among Puerto Ricans. Median lengths of sROH and cumulative IBD in Puerto Ricans are also the highest among the Hispanic clusters (8.01 Mb

and 26.23 cM, respectively; Table 1). Consistent with other studies,[45,52] we found evidence of a strong bottleneck in Puerto Ricans approximately 9–14 generations ago (Figure S12), coinciding with the colonization of America and likely explaining the elevated levels of IBD and sROH.

Three distinct clusters of Hispanics/Latinos were found in the Southwest (Figure 4A): one strongly associated with New Mexico (OR > 4; $p < 0.05$), another primarily in Texas (OR > 3; $p < 0.05$), and the third associated with Southern California (OR > 2; $p < 0.05$). Combined with the EEMS analysis, these clusters confirm our observation of parallel migration routes from east and west Mexico into Southwestern United States. While genetic differentiation between these three clusters are subtle ($F_{ST} = 0.001–0.003$), comparison of the ancestral birth origin patterns and local ancestry proportions of these clusters reveal meaningful differences in their population history. Whereas the majority of Hispanics/Latinos in New Mexico report US ancestral origins, the recent ancestors of Hispanics/Latinos in Texas are predominantly from Mexico. Nonetheless, these two clusters share similar local ancestry proportions with only slight genetic dissimilarity that result in a moderate decrease in migration rate (from darker blue to light blue in Figure 3B). Unlike the Hispanic/Latino clusters associated with New Mexico and Texas, the Hispanics/Latinos in California cluster contain greater proportions of ancestors from Central and South America (e.g., Colombia and El Salvador). Proportions of Native American ancestry (Figure 4C) and effective population size (Figure S12) are also higher in this cluster, but median cumulative IBD and sROH length are shorter (Table 1), similar to Central/South Americans found in New York City.[52] Taken together, these two differences further explain the presence of the migration barrier in Arizona between Hispanic/Latino individuals in California and those in New Mexico.

Historical immigration of Europeans into the US occurred in successive waves, with northern and western Europeans making up one wave from the 1840s to 1880s and another wave comprised of southern and eastern Europeans occurring from the 1880s to 1910s.[53] Consistent with this immigration pattern, haplotype clusters with ancestries from northwest and central Europe have higher proportions of US ancestral birth origins than haplotype clusters from southern and eastern Europe, suggesting earlier immigration (Figures 5A and 5B). The two clusters with the highest proportion (>75%) of US ancestral birth origin ("Northwest Europe 1" and "Northwest Europe 2") have ~4.5% of UK ancestral origins. The central European cluster and the Irish cluster both have 66.1% and 68.5% of US ancestral origins, respectively (Figure 5B). In contrast, the US makes up only 62.2% and 34.5% of ancestral birth origin for the clusters of southern Europeans and eastern Europeans, respectively.

Unlike the larger European clusters, the smaller European clusters reflect the structure of recent immigrants and genetically isolated populations, recapitulating earlier findings.[8] The geographic distributions of these subpopulations are more concentrated, and their ancestral birth origin proportions are overrepresented by specific countries and ethnicities (Figures 6A and 6B). Specifically, Finns and Scandinavians are abundant in the Upper Midwest and Washington; French Canadians are found in the Northeast; Acadians are present in the Northeast and Louisiana; and Italians, Greeks, and Jews are mostly located in the metropolitan area of New York City (Figure 6A). Of the European clusters, median cumulative IBD sharing and sROH lengths are highest among Ashkenazi Jews (31.8 cM and 11.3 Mb, respectively; Table 1), reflective of past founding events and endogamy.[21,54] The two Jewish-related clusters were identified using self-reported ancestral ethnicity data rather than birth origin data, since Jewish ancestry is not specific to any single location. Jewish ancestry, particularly Ashkenazi Jewish ancestry, is more consistently reported on both sides of the family in the larger cluster, while individuals in the smaller cluster more commonly reported Jewish ancestry on only one side of the family, suggesting the presence of admixture with non-Jewish ancestries. Therefore, the larger cluster is labeled "Ashkenazi Jewish" and the smaller cluster is labeled "Admixed Jewish."

We inferred two haplotype clusters of African Americans separated along a north-south cline, recapitulating the EEMS migration barrier inference. One cluster is primarily distributed among the northern and western states ("African Americans North"), while the other is distributed among the states southeast of the Appalachian Mountains ("African Americans South") (Figure S13). The proportion of US birth origin is higher in the northern cluster than the southern cluster, providing further evidence of isolation-by-distance among African Americans in the north.[7] These two clusters share similar sROH lengths but differ in admixture proportions and median IBD sharing (Table 1), pointing to a cluster with consistent African American ancestors and a cluster with more admixed ancestors. Median cumulative IBD sharing is higher among African Americans in the south (median cumulative IBD = 19.6 cM, median sROH = 3.3 Mb) than in the north (median = 15.9 cM; Table 1), resulting in different patterns of effective population size over antecedent generations (Figure S12),[45,46] while the average proportion of African ancestry is higher in the northern cluster than the southern cluster.

Four of the clusters reflect recent immigrants from Asia (Figure S14), which grew rapidly in the mid-20th century after the elimination of national origin quotas.[55] The recency of immigration among these clusters is supported by the observation that fewer than 30% of grandparents were born in the US. Geographically, individuals in these clusters primarily reside in major cities. East Asians predominantly inhabit the metropolitan areas of the west and northeast (OR > 2), Southeast Asians are enriched in the west (OR > 2.5), and South Asians are strongly associated with the northeast (OR > 2.5). Despite its small size, the cluster of Greater Middle East individuals reflects many of the known

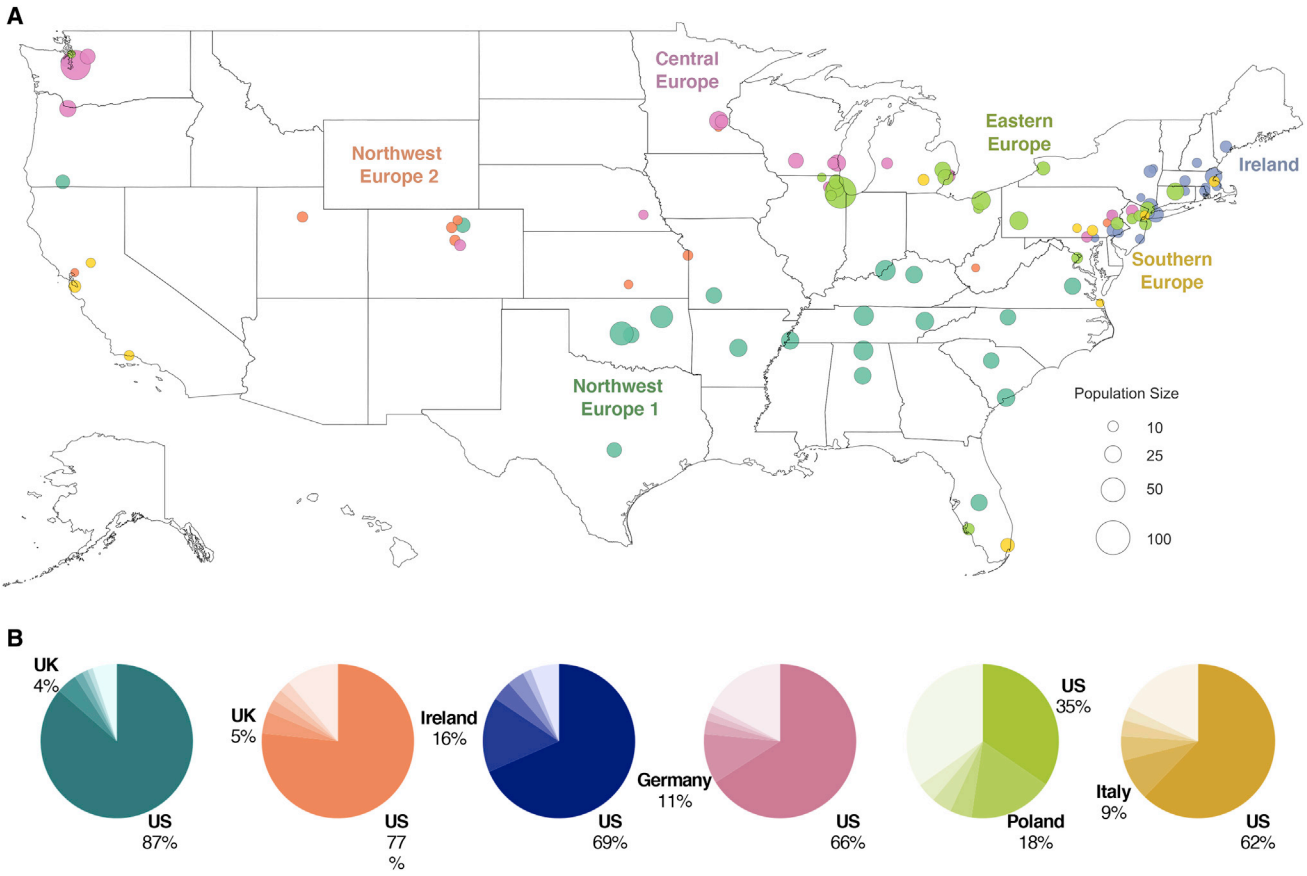


Figure 5. Geographical Distribution of European American Haplotype Clusters
(A) Each dot represents a county containing present-day individuals. The size of the dot represents the number of individuals of the particular cluster in that county. For each cluster, the top 20 locations with the highest odds ratio are shown. Maps showing the full distribution for each cluster can be found in the supplement (Figure S16).
(B) Ancestral birth origin proportions for each cluster in (A). Only individuals with complete pedigree annotations, up to grandparent level, are included. For each chart, only the top five birth origins are shown as individual proportions; the remaining birth origins are aggregated into one slice (lightest color).

demographic patterns of Arab Americans, as individuals in this cluster are primarily of Lebanese origin and are distributed in the northeast as well as metropolitan Detroit. sROH lengths are particularly long for South Asians (median sROH = 10.3 cM; Table 1), Southeast Asians (median sROH = 7.8 cM), and Middle Easterners (median sROH = 8.2 cM), potentially reflecting patterns of consanguinity and inbreeding in their ancestral regions.[56] In particular, the median sROH length in the South Asia cluster is the second highest among all clusters, but the median cumulative IBD length is similar to most clusters (Table 1). The population of South Asia is large and diverse, with many endogamous groups making up the 1.5 billion people living in the region.[57,58] The pattern of IBD and sROH among individuals in South Asian cluster thus may reflect the result of recent consanguinity in a large population.[48,59]

Discussion

As the US population is becoming increasingly diverse, genomic studies are simultaneously growing in scale and relevance; to increase scientific and ethical parity, these studies must move beyond the current practice of evaluating genetically homogeneous groups in isolation.[47,60] Here, we provide an integrative framework for analyzing population structure in ancestrally heterogeneous individuals. Our comprehensive approach has allowed us to capture spatial patterns of gene flow within and between subpopulations that are difficult to infer from a single method alone. For example, while EEMS enabled us to examine genetic similarity at a finer scale than previous studies and identify genetic differentiation within a state, EEMS can only compare neighboring demes and does not directly evaluate the genetic similarity of geographically distant individuals. Haplotype clustering, on the other hand, can identify population structures over long distances, but it does not measure genetic similarity with respect to geography. Since individuals are exclusively assigned to a single cluster, information regarding admixture, especially between neighboring clusters, are lost during haplotype clustering. An integrative approach can thus enable greater insights into populations with complex histories, as

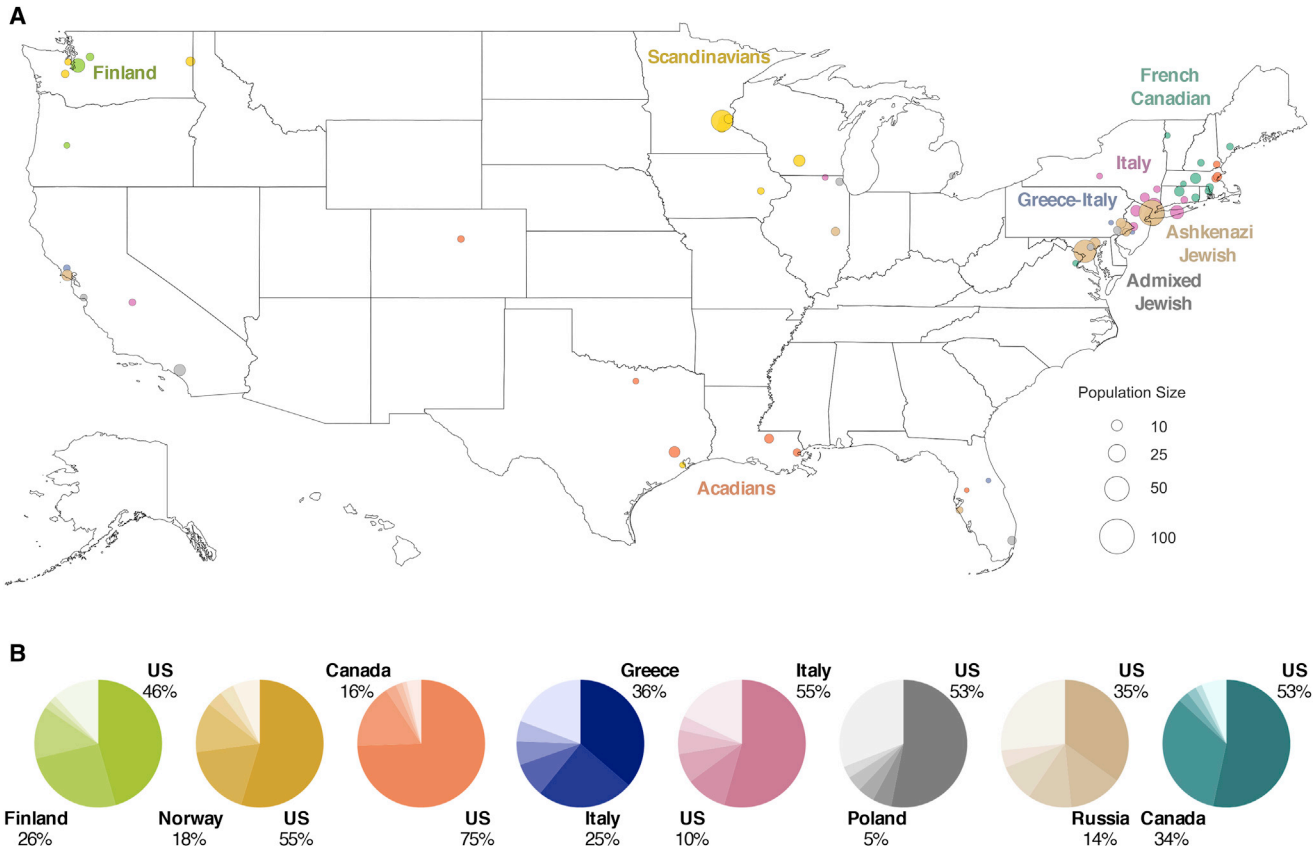


Figure 6. Geographical Distribution of Genetically Differentiated European American Haplotype Clusters
(A) Similar to Figure 5A but corresponding to European populations that are more genetically isolated. For clarity, the top ten locations with the highest odds ratio are shown for each cluster. Full distributions for each cluster can be found in the supplement (Figure S17). (B) Ancestral birth origin proportions for each cluster in (A). Only individuals with complete pedigree annotations, up to grandparent level, are shown. For each chart, only the top five birth origins are shown as individual proportions; the remaining birth origins are aggregated into one slice (lightest color).

well as populations typically overlooked in previous studies such as Asian Americans.

The genetic structure and history of Hispanic/Latino populations is particularly complex due to many historical migration and admixture events.[4,9] This complexity is reflected in the variable migration rates across the country and the large variations in admixture proportions within and between subpopulations. While prior analysis of Hispanics/Latinos in the US found differences in ancestry proportions aggregated at the state level,[9] we demonstrate that considerable differences in genetic ancestry also exist within a state. For example, two distinct clusters—Puerto Rico and Hispanics/Latinos—are found in Florida, with the Puerto Rico cluster having higher average African ancestry proportions than the Hispanics/Latinos cluster (9.0% versus 2.5%, respectively). EEMS also enabled direct measures of genetic similarity within states and between subpopulations. While the mean ancestry proportions are similar between the New Mexican cluster and the Texan cluster, individuals in northern New Mexico are more genetically differentiated than individuals in southern New Mexico, as indicated by the migration barrier. The individuals in northern New Mexico are likely *Nuevomexicanos*, descendants of Spanish colonial settlers, while those in the south are more genetically similar to Hispanic/Latino individuals in central Texas, likely because they share a common ancestral origin (i.e., Mexico). We also built upon the use of pedigree annotation[8] by quantifying ancestral origins to better understand the differences in genetic ancestry between subpopulations. For example, in the Hispanics/Latinos in California cluster, the mean proportion of European ancestry is smaller when compared to the New Mexican and Texan clusters, reflecting the lower proportion of US ancestral origins. Comparison of sROH and IBD lengths of these clusters further reveal evidence of founder effects. Puerto Ricans, Hispanics/Latinos in Texas, and Hispanics/Latinos in New Mexico had the highest median IBD lengths and showed evidence of recent bottleneck (Figure S12), consistent with prior studies.[45,52] In general, median sROH and IBD lengths were higher in Hispanic-related clusters than European clusters, reflecting the patterns found in reference populations[56,61] and in line with recent findings in New York City.[52]

The demographic history of African Americans is characterized by large-scale migration and admixture,

primarily due to the transatlantic slave trade and racial segregation.[50,62] The patterns of genetic ancestry and relatedness between states and regions of the US reflect these events.[3,7,9] Our results show, at a finer scale, the barriers to migration and gene flow, particularly along the Appalachian Mountains. This migration barrier overlaps with the boundary between slave states and free states, as well as the boundary between states that enacted laws enforcing racial segregation and states that forbade segregation. The north-south separation of two African American clusters further emphasize this divide. The African Americans South cluster contains more recent ancestors from outside the US, particularly from the Caribbean, than the African Americans North cluster. These insights further emphasize the impact of historical migration and socioeconomic divide on the present-day patterns of genetic relatedness among African Americans.

Despite accounting for more than 5% of the US population, individuals with Asian ancestries are underrepresented in US population genetics studies, hindering the ability of prior studies to investigate of their ancestry.[8] Our analyses of these individuals therefore provide new insights into their genetic structure. Many of these individuals are descendants of recent immigrants, as indicated by the high proportions of non-US grandparental ancestral origin; therefore, they likely reflect the population of their ancestral region. The genetic structure of these individuals is particularly diverse. Using fineSTRUCTURE, genetic differentiation was found between East Asian and South Asian individuals of different ancestral origin as well as between individuals with the same ancestral origin. At the same time, longer sROH was observed in the Southeast Asia, South Asia, and Greater Middle East haplotype clusters, likely reflecting consanguinity or endogamy patterns in their ancestral countries. For example, the long sROH in South Asians may reflect endogamy related to the caste system in India, while similar patterns among the Middle Eastern and Southeast Asian clusters may be capturing consanguineous marriage practices in those regions.[48,63,64] Understanding population genetic structure and patterns of homozygosity are important in determining the genetic profile of diseases within subpopulations, especially since these recent immigrants are becoming less similar to those in their ancestral countries due to outbreeding, admixture, and population growth.[65,66] As populations mix, heterozygosity increases and allele frequencies change. This, in turn, can alter the prevalence of certain diseases, particularly rare recessive disorders that are often more prevalent in populations with increased homozygosity.[67] At the same time, changes in allele frequencies can also reduce the accuracy of genetic predictors of complex traits (i.e., polygenic risk scores), especially if the prediction model was built using a homogeneous cohort of individuals from a divergent ancestry.[60]

Population history in the US is best characterized among individuals of European descent. Genetic diversity tends to be highest in more densely populated regions, likely due to multiple populations living in the same place. Many of the European subpopulations we identified are similar to those previously found—e.g., French Canadians, Acadians, Scandinavians, and Ashkenazi Jews.[8] The geographic distribution of these subpopulations, particularly those that are more genetically diverged, overlap in the metropolitan areas of the Northeast, Midwest, and California. These overlaps may explain the presence of certain EEMS-inferred migration barriers. For example, the migration barrier and lower genetic similarity encompassing metropolitan New York City may be explained in part by the large presence of Greeks, Italians, and Ashkenazi Jews in that area.

The precision of population labels assigned to clusters of individuals is a function of demographic complexity and sample size. For example, Finnish ancestry is clearly European but genetically distinct from several other European populations due to historical bottlenecks, making this ancestry cluster relatively easily separable. By contrast, most Americans of European descent have heterogeneous ancestors from several northwestern European countries who have admixed over time, resulting in relatively evenly distributed ancestry overlapping that of present-day Europeans from multiple primarily northwestern countries. Additionally, while we identify and describe some substantial structure among Hispanic/Latino populations, considerably more is likely to exist and remains to be learned from larger and more diverse future studies. Similarly, sub-regional resolution into the ancestry of recent Asian immigrants has been relatively limited in population genetics studies, and the structure of this immigration will be learned from larger future studies. Interestingly, we found that fineSTRUCTURE was able to disentangle Asian subpopulations at a finer resolution than haplotype clustering, demonstrating the tradeoff between resolution and scale of these two methods and further highlighting the value of an integrated approach. The accuracy of self-reported birth records and variable granularity of geopolitical boundaries also provide additional considerations regarding the precision of population labels.

In addition to being of anthropological interest, understanding fine-scale human history and its role in shaping genetic variation is also important for interpreting the genetic basis of biomedical traits. The emergence of biobank-scale genomic data is enabling the imputation of pedigree structure regardless of whether some relatives have contributed DNA,[68] greater insights into the impact of fine-scale population structure on genetic associations with disease,[14,17,27,60,69] and population-based screening for individuals with serious genetic and health-related associations.[70] Standard practice in genetic studies to date has involved identifying the largest genetically homogeneous population in a study (typically European ancestry) and conducting genetic analysis excluding other populations.[71,72] However, as genetic studies become increasingly promising for clinical translation, this practice has led to concerns about genetic tools exacerbating health

disparities, particularly for populations underrepresented in genetic studies.[47,60] Participation in genetic programs is increasing in the US, for example with the All of Us Research Program (Web Resources) or with direct-to-consumer genetic tests that an estimated 26 million people have taken, and many of these participants are of diverse non-European ancestry.[73,74] As a result, the need for including more diverse populations in genetic studies and for inferring more granular demographic histories in diverse study cohorts is becoming greater. Understanding such structure is important to account for stratification in association studies, prevent the overgeneralization of potentially confounded results, and avoid exacerbating existing Eurocentric study biases.[60,71,75,76] This study demonstrates how genetic data can be coupled with geographic and birth origin data to reconstruct such demographic histories, particularly in a large and heterogeneous population.

Data and Code Availability

Genotype data and associated metadata are available to researchers through an application process and data usage agreement. We encourage qualified researchers to email the Genographic team at National Geographic Society (genographic@ngs.org) for information on and access to the Genographic database. For more information, please visit the Genographic Project website (https://genographic.nationalgeographic.com/for-scientists/).

Custom scripts generated to analyze the data in this paper are available through GitHub (https://github.com/chengdai/genographic_ancestry).

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.02.002.

Acknowledgments

We thank the National Geographic Genographic Project participants who consented to research for making this study possible. We also thank Gregory Vilshansky for helping organize and manage the data for the Genographic Project. This work was supported by funding from the National Institutes of Health (K99MH117229 to A.R.M.). C.L.D., M.M.V., R.T., and C.R. would also like to thank all the members of the MIT Senseable City Lab Consortium for supporting this research. M.G.V. acknowledges support from the National Geographic Society.

Declaration of Interests

M.G.V. is the Senior Program Officer for the National Geographic Society and lead scientist for the Genographic Project. R.S.W. was the former Director of the Genographic Project and is a cofounder for Insitome. M.J.D. is a member of the Scientific Advisory Board at Ancestry.com LLC.

Received: September 22, 2019
Accepted: February 5, 2020
Published: March 5, 2020

Web Resources

1000 Genomes Project, https://www.internationalgenome.org
All of Us Research Hub, https://www.researchallofus.org/
Ancestry pipeline, https://github.com/armartin/ancestry_pipeline
Human Genome Diversity Project, https://www.hagsc.org/hgdp/
Interactive Tree of Life, https://itol.embl.de/
POPRES: Population Reference Sample, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v4.p2
python-ternary, https://github.com/marcharper/python-ternary
scikit-bio, http://scikit-bio.org/

References

1. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
2. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.
3. Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S.A., and Bustamante, C.D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA *107*, 786–791.
4. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., and Ostrer, H. (2010). Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc. Natl. Acad. Sci. USA *107* (*Suppl 2*), 8954–8961.
5. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. Nature *488*, 370–374.
6. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.
7. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J., Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., and Gravel, S. (2016). The Great Migration and African-American Genomic Diversity. PLoS Genet. *12*, e1006059.
8. Han, E., Carbonetto, P., Curtis, R.E., Wang, Y., Granka, J.M., Byrnes, J., Noto, K., Kermany, A.R., Myres, N.M., Barber, M.J., et al. (2017). Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nat. Commun. *8*, 14238.
9. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., and Mountain, J.L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet. *96*, 37–53.
10. Zakharia, F., Basu, A., Absher, D., Assimes, T.L., Go, A.S., Hlatky, M.A., Iribarren, C., Knowles, J.W., Li, J., Narasimhan,

B., et al. (2009). Characterizing the admixed African ancestry of African Americans. Genome Biol. *10*, R141.
11. Pfaff, C.L., Parra, E.J., Bonilla, C., Hiester, K., McKeigue, P.M., Kamboh, M.I., Hutchinson, R.G., Ferrell, R.E., Boerwinkle, E., and Shriver, M.D. (2001). Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am. J. Hum. Genet. *68*, 198–207.
12. Parra, E.J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E., and Shriver, M.D. (1998). Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. *63*, 1839–1851.
13. Price, A.L., Patterson, N., Yu, F., Cox, D.R., Waliszewska, A., McDonald, G.J., Tandon, A., Schirmer, C., Neubauer, J., Bedoya, G., et al. (2007). A genomewide admixture map for Latino populations. Am. J. Hum. Genet. *80*, 1024–1036.
14. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K., and Coop, G. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. eLife *8*, e39725.
15. Henn, B.M., Cavalli-Sforza, L.L., and Feldman, M.W. (2012). The great human expansion. Proc. Natl. Acad. Sci. USA *109*, 17758–17764.
16. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
17. Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W., Hirschhorn, J., Daly, M.J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife *8*, e39702.
18. Kayser, M., Brauer, S., Weiss, G., Schiefenhövel, W., Underhill, P., Shen, P., Oefner, P., Tommaseo-Ponzetta, M., and Stoneking, M. (2003). Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am. J. Hum. Genet. *72*, 281–302.
19. Montinaro, F., Busby, G.B.J., Pascali, V.L., Myers, S., Hellenthal, G., and Capelli, C. (2015). Unravelling the hidden ancestry of American admixed populations. Nat. Commun. *6*, 6596.
20. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013). Reconstructing the population genetic history of the Caribbean. PLoS Genet. *9*, e1003925.
21. Bray, S.M., Mulle, J.G., Dodd, A.F., Pulver, A.E., Wooding, S., and Warren, S.T. (2010). Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. Proc. Natl. Acad. Sci. USA *107*, 16222–16227.
22. Peltonen, L., Palotie, A., and Lange, K. (2000). Use of population isolates for mapping complex traits. Nat. Rev. Genet. *1*, 182–190.
23. Scriver, C.R. (2001). Human genetics: lessons from Quebec populations. Annu. Rev. Genomics Hum. Genet. *2*, 69–101.
24. Martin, A.R., Karczewski, K.J., Kerminen, S., Kurki, M.I., Sarin, A.-P., Artomov, M., Eriksson, J.G., Esko, T., Genovese, G., Havulinna, A.S., et al. (2018). Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. Am. J. Hum. Genet. *102*, 760–775.
25. Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A., et al. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science *356*, 543–546.
26. Mooney, J.A., Huber, C.D., Service, S., Sul, J.H., Marsden, C.D., Zhang, Z., Sabatti, C., Ruiz-Linares, A., Bedoya, G., Freimer, N., Lohmueller, K.E.; and Costa Rica/Colombia Consortium for Genetic Investigation of Bipolar Endophenotypes (2018). Understanding the Hidden Complexity of Latin American Population Isolates. Am. J. Hum. Genet. *103*, 707–726.
27. Belbin, G.M., Odgis, J., Sorokin, E.P., Yee, M.-C., Kohli, S., Glicksberg, B.S., Gignoux, C.R., Wojcik, G.L., Van Vleck, T., Jeff, J.M., et al. (2017). Genetic identification of a common collagen disease in puerto ricans via identity-by-descent mapping in a health system. eLife *6*, e25060.
28. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science *343*, 747–751.
29. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, e1002453.
30. Elhaik, E., Greenspan, E., Staats, S., Krahn, T., Tyler-Smith, C., Xue, Y., Tofanelli, S., Francalacci, P., Cucca, F., Pagani, L., et al.; Genographic Consortium (2013). The GenoChip: a new tool for genetic anthropology. Genome Biol. Evol. *5*, 1021–1031.
31. Royal, C.D., Novembre, J., Fullerton, S.M., Goldstein, D.B., Long, J.C., Bamshad, M.J., and Clark, A.G. (2010). Inferring genetic ancestry: opportunities, challenges, and implications. Am. J. Hum. Genet. *86*, 661–673.
32. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
33. Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., Briley, L.P., Maruyama, Y., Waterworth, D.M., Waeber, G., et al. (2008). The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am. J. Hum. Genet. *83*, 347–358.
34. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2019). Insights into human genetic variation and population history from 929 diverse genomes. bioRxiv. https://doi.org/10.1101/674986.
35. Abraham, G., Qiu, Y., and Inouye, M. (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics *33*, 2776–2778.
36. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
37. McInnes, L., and Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv, 1802.03426.
38. Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., and Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS Genet. *15*, e1008432.
39. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.

40. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
41. Petkova, D., Novembre, J., and Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. Nat. Genet. *48*, 94–100.
42. Browning, B.L., and Browning, S.R. (2013). Detecting identity by descent and estimating genotype error rates in sequence data. Am. J. Hum. Genet. *93*, 840–851.
43. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. *2008*, P10008.
44. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
45. Browning, S.R., Browning, B.L., Daviglus, M.L., Durazo-Arvizu, R.A., Schneiderman, N., Kaplan, R.C., and Laurie, C.C. (2018). Ancestry-specific recent effective population size in the Americas. PLoS Genet. *14*, e1007385.
46. Browning, S.R., and Browning, B.L. (2015). Accurate Nonparametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am. J. Hum. Genet. *97*, 404–418.
47. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.
48. Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P.-R., Govindaraj, P., Berger, B., Reich, D., and Singh, L. (2013). Genetic evidence for recent population mixture in India. Am. J. Hum. Genet. *93*, 422–438.
49. Lal, B.V. (1983). Girmitiyas: The Origins of the Fiji Indians (J. Pac. Hist.).
50. US Census Bureau. The Great Migration, 1910 to 1970, Retrieved February 21, 2019. https://www.census.gov/dataviz/visualizations/020/.
51. Massey, D.S., Rugh, J.S., and Pren, K.A. (2010). The Geography of Undocumented Mexican Migration. Mex. Stud. *26*, 129–152.
52. Belbin, G.M., Wenric, S., Cullina, S., Glicksberg, B.S., Moscati, A., Wojcik, G.L., Shemirani, R., Beckmann, N.D., Cohain, A., Sorokin, E.P., et al. (2019). Towards a fine-scale population health monitoring system. bioRxiv. https://doi.org/10.1101/780668.
53. Passel, J.S., and Fix, M. (1994). U.S. Immigration in a Global Context: Past, Present, and Future. Indiana J. Glob. Leg. Stud. *2*, 5–19.
54. Ostrer, H., and Skorecki, K. (2013). The population genetics of the Jewish people. Hum. Genet. *132*, 119–127.
55. Grieco, E.M., Trevelyan, E., Larsen, L., Acosta, Y.D., and Gambino, C. (2012). The Size, Place of Birth, and Geographic Distribution of the Foreign-Born Population in the United States: 1960 to 2010 (US Census Bureau). https://www.census.gov/content/dam/Census/library/working-papers/2012/demo/POP-twps0096.pdf.
56. Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., and Li, J.Z. (2012). Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum. Genet. *91*, 275–292.
57. Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., Patterson, N., Bhavani, G.S., Girisha, K.M., Mustak, M.S., Srinivasan, S., et al. (2017). The promise of discovering population-specific disease-associated genes in South Asia. Nat. Genet. *49*, 1403–1407.
58. Wall, J.D., Stawiski, E.W., Ratan, A., Kim, H.L., Kim, C., Gupta, R., Suryamohan, K., Gusareva, E.S., Purbojati, R.W., Bhangale, T., et al.; GenomeAsia100K Consortium (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature *576*, 106–111.
59. Auton, A., Bryc, K., Boyko, A.R., Lohmueller, K.E., Novembre, J., Reynolds, A., Indap, A., Wright, M.H., Degenhardt, J.D., Gutenkunst, R.N., et al. (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res. *19*, 795–803.
60. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
61. Ceballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M., and Wilson, J.F. (2018). Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet. *19*, 220–234.
62. National Archives (2016). The Slave Trade. https://www.archives.gov/education/lessons/slave-trade.html.
63. Tadmouri, G.O., Nair, P., Obeid, T., Al Ali, M.T., Al Khaja, N., and Hamamy, H.A. (2009). Consanguinity and reproductive health among Arabs. Reprod. Health *6*, 17.
64. Hussain, R., and Bittles, A.H. (2004). Assessment of association between consanguinity and fertility in Asian populations. J. Health Popul. Nutr. *22*, 1–12.
65. López, G., Ruiz, N., and Patten, E. (2017). Key facts about Asian Americans, a diverse and growing population. Pew Research Center Fact Tank, September 8, 2017. https://www.pewresearch.org/fact-tank/2017/09/08/key-facts-about-asian-americans/.
66. Bialik, K. (2017). Key facts about race and marriage, 50 years after Loving v. Virginia. Pew Research Center Fact Tank, June 12, 2017. https://www.pewresearch.org/fact-tank/2017/06/12/key-facts-about-race-and-marriage-50-years-after-loving-v-virginia/.
67. Nalls, M.A., Simon-Sanchez, J., Gibbs, J.R., Paisan-Ruiz, C., Bras, J.T., Tanaka, T., Matarin, M., Scholz, S., Weitz, C., Harris, T.B., et al. (2009). Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet. *5*, e1000415.
68. Erlich, Y., Shor, T., Pe'er, I., and Carmi, S. (2018). Identity inference of genomic data using long-range familial searches. Science *362*, 690–694.
69. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.
70. Bastarache, L., Hughey, J.J., Hebbring, S., Marlo, J., Zhao, W., Ho, W.T., Van Driest, S.L., McGregor, T.L., Mosley, J.D., Wells, Q.S., et al. (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science *359*, 1233–1239.
71. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
72. Need, A.C., and Goldstein, D.B. (2009). Next generation disparities in human genomics: concerns and remedies. Trends Genet. *25*, 489–494.

73. Regalado, A. (2019). More than 26 million people have taken an at-home ancestry test. MIT Technology Review, February 11, 2019. https://www.technologyreview.com/s/612880/more-than-26-million-people-have-taken-an-at-home-ancestry-test/.
74. Parker, K., Horowitz, J.M., Mortin, R., and Lopez, M.H. (2015). Multiracial in America: Proud, Diverse and Growing in Numbers. Pew Research Center, June 11, 2015. https://www.pewsocialtrends.org/2015/06/11/multiracial-in-america/.
75. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic Misdiagnoses and the Potential for Health Disparities. N. Engl. J. Med. *375*, 655–665.
76. Caswell-Jin, J.L., Gupta, T., Hall, E., Petrovchich, I.M., Mills, M.A., Kingham, K.E., Koff, R., Chun, N.M., Levonian, P., Lebensohn, A.P., et al. (2018). Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. Genet. Med. *20*, 234–239.

The American Journal of Human Genetics, Volume *106*

Supplemental Data

Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis

Chengzhen L. Dai, Mohammad M. Vazifeh, Chen-Hsiang Yeang, Remi Tachet, R. Spencer Wells, Miguel G. Vilar, Mark J. Daly, Carlo Ratti, and Alicia R. Martin

Supplementary Figures

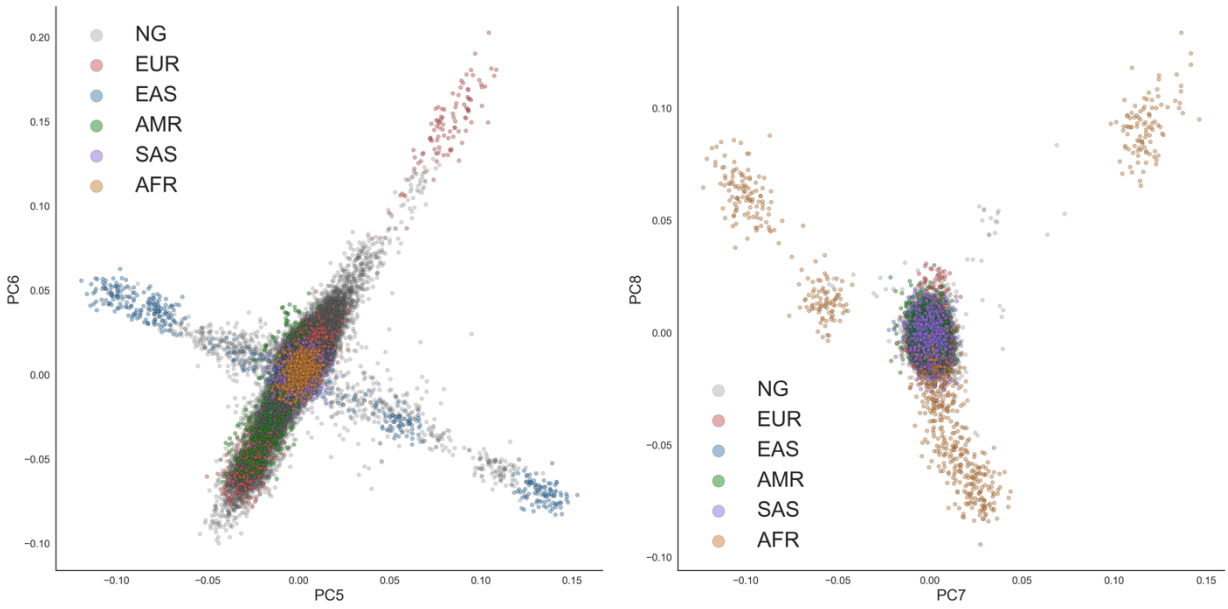
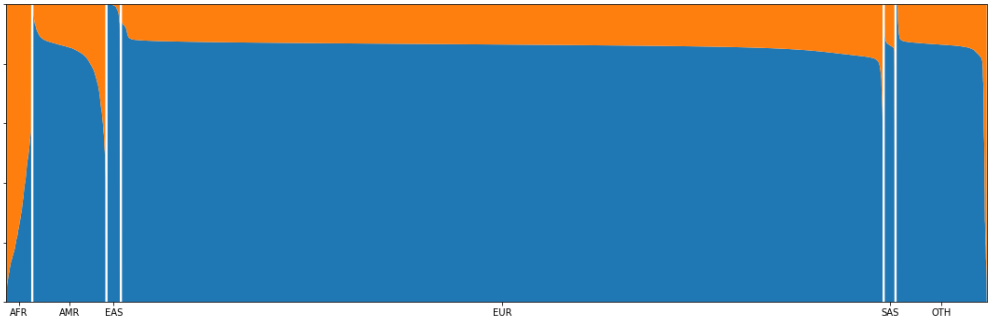
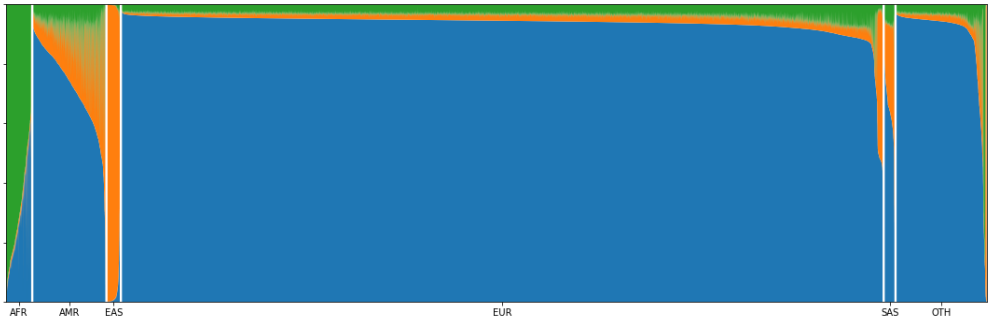
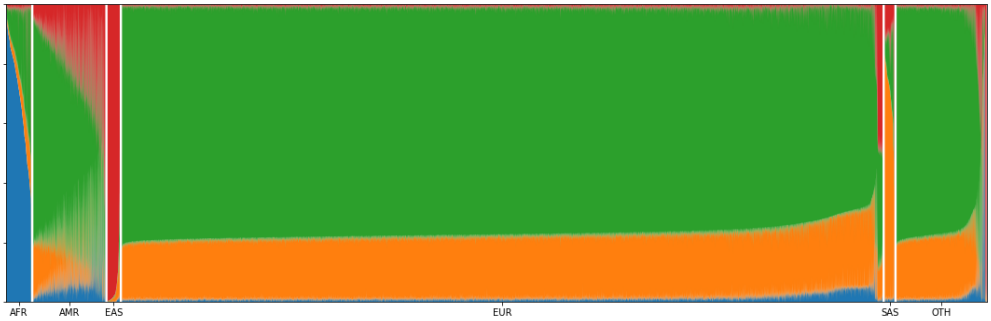


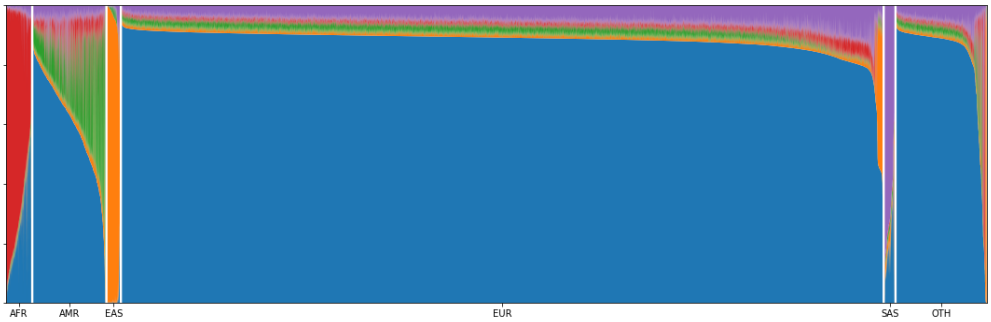
Figure S1. Principal Component Analysis of 1000 Genome Project and Genographic samples

PCA projects at for PC 5 and PC 6 (left); and for PC 7 and PC 8 (right).


AFR
AMR
EAS
EUR
SAS
OTH


AFR
AMR
EAS
EUR
SAS
OTH


AFR
AMR
EAS
EUR
SAS
OTH


AFR
AMR
EAS
EUR
SAS
OTH

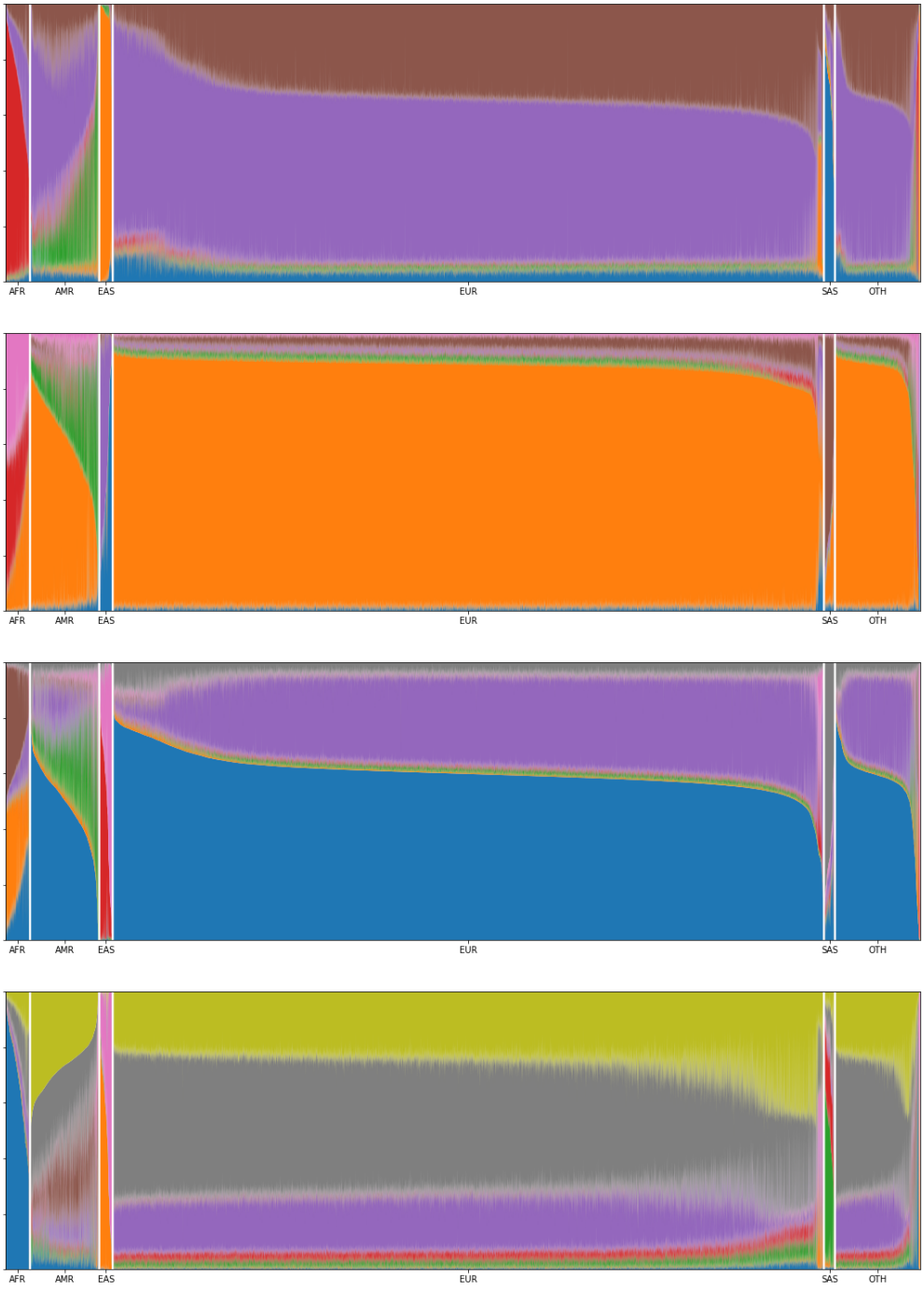


Figure S2. ADMIXTURE from K = 2 to 9.

ADMIXTURE analysis results for each K between 2 and 9 of U.S. individuals. Individuals were classified into continent level ancestry groups with a Random Forest model trained on the PCs from the 1000 Genome Project dataset.

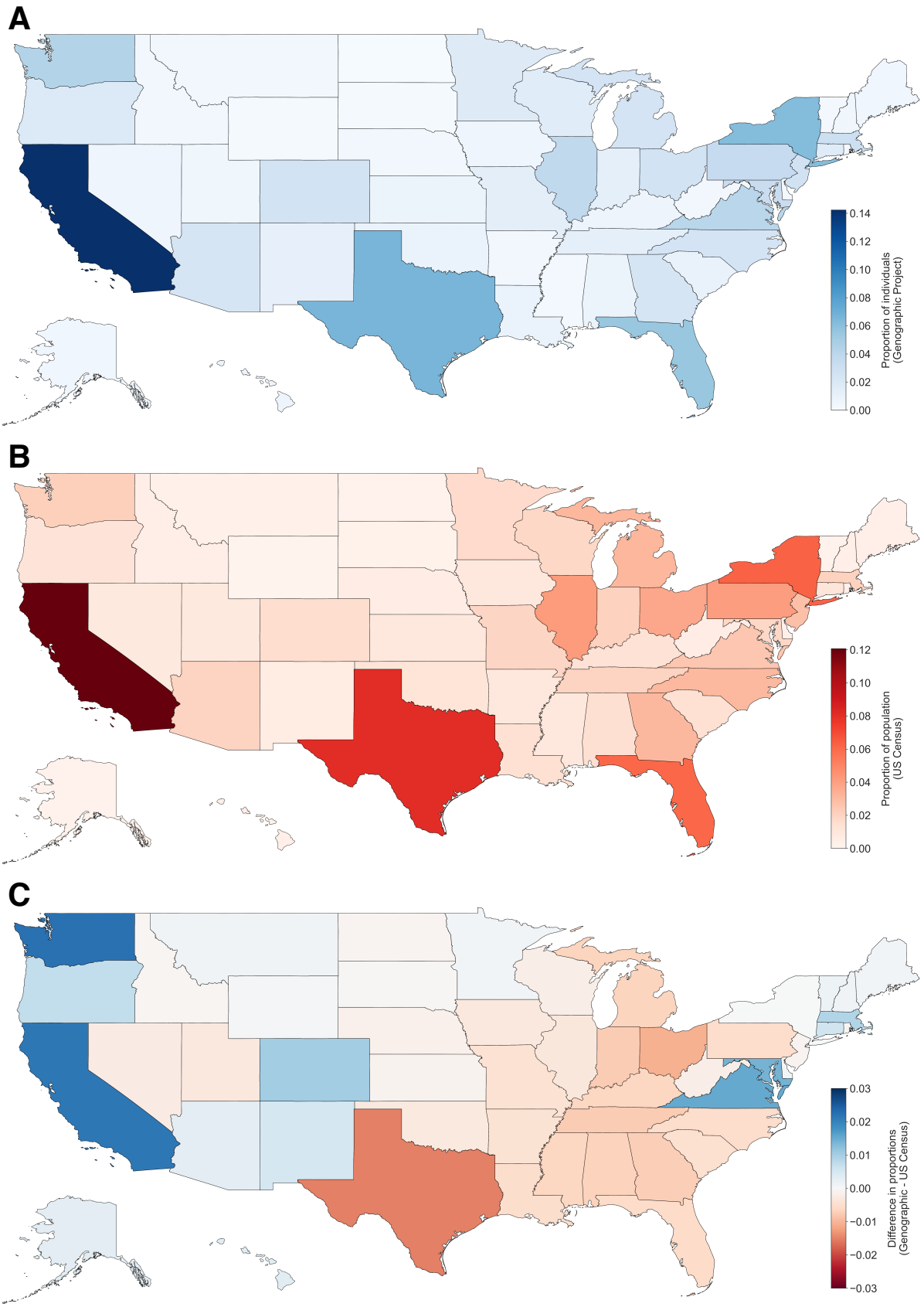

A
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Proportion of individuals
(Genographic Project)
B
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Proportion of population
(US Census)
C
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
Difference in proportions
(Genographic - US Census)

Figure S3. Comparison of Population Distribution by State of Genographic Participants and US Census.

(A) Distribution of Genographic participants by state. Darker shades of blue represent higher proportion of Genographic samples in a state. The five most represented states are: California, Texas, New York, Florida, and Washington.

(B) Distribution of US population according to the 2010 US Census. Darker shades of green present higher priorition of the population. The five most populous states are: California, Texas, New York, Florida, and Illinois.

(C) Difference in the distribution of Genographic participants and US Census population distribution. Positive values represent higher proportions in the Genographic cohort while lower values represent higher proportions reported in the US Census.

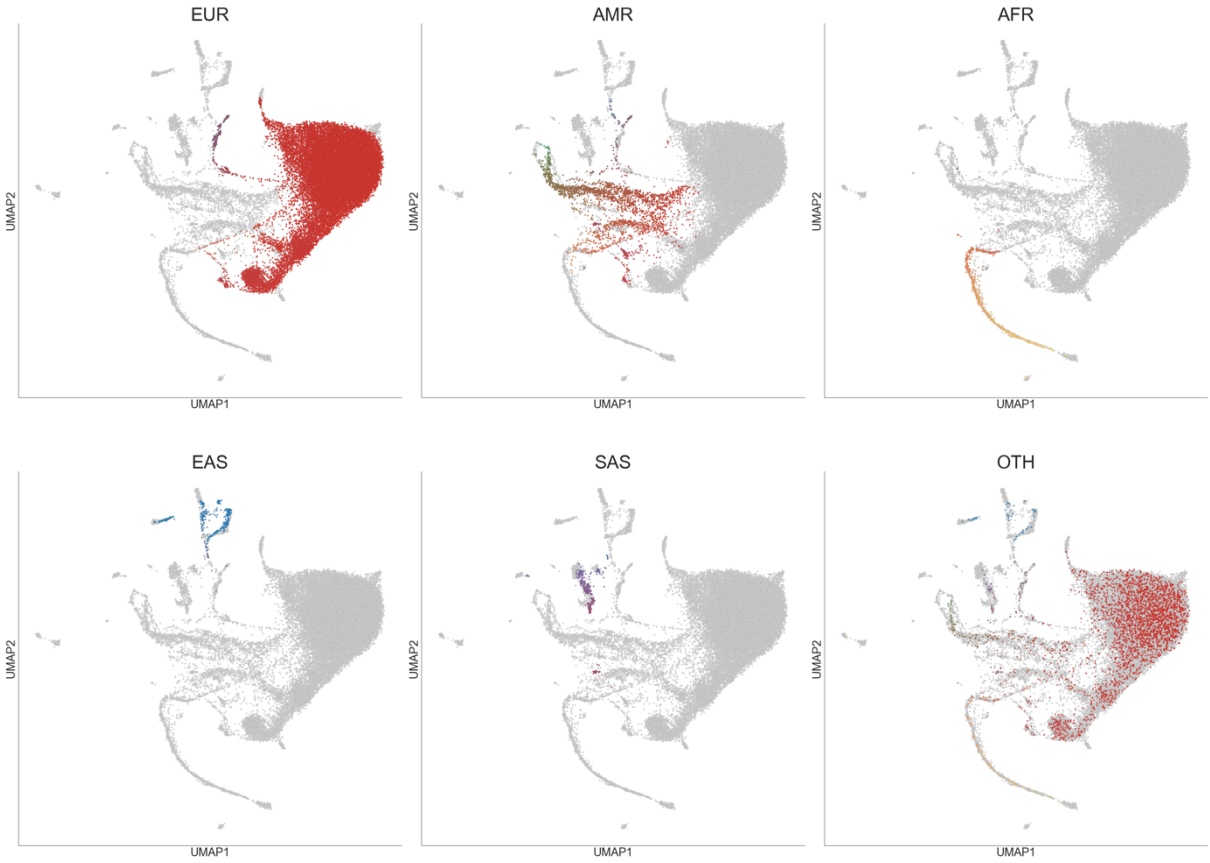


Figure S4. Uniform Manifold Approximation and Projection (UMAP) of Classified Genographic Individuals

UMAP projection of the first 20 PCs. Each dot represents one individual. Each plot represents the set of individuals classified at continental-level ancestry with the Random Forest model trained on the 1000 Genomes Project data. 1000 Genome Project individuals are colored in grey while U.S. individuals are colored based on their admixture proportions from ADMIXTURE. The color for each dot was calculated as a linear combination of each individual's admixture proportion and the RGB values for the colors assigned to each continental ancestry (EUR = red, AFR = yellow, NAM = green, EAS = blue, SAS = purple). Continental level ancestries are: EUR = European, AFR = African, NAM = Native American, EAS = East Asian, SAS = South Asian.

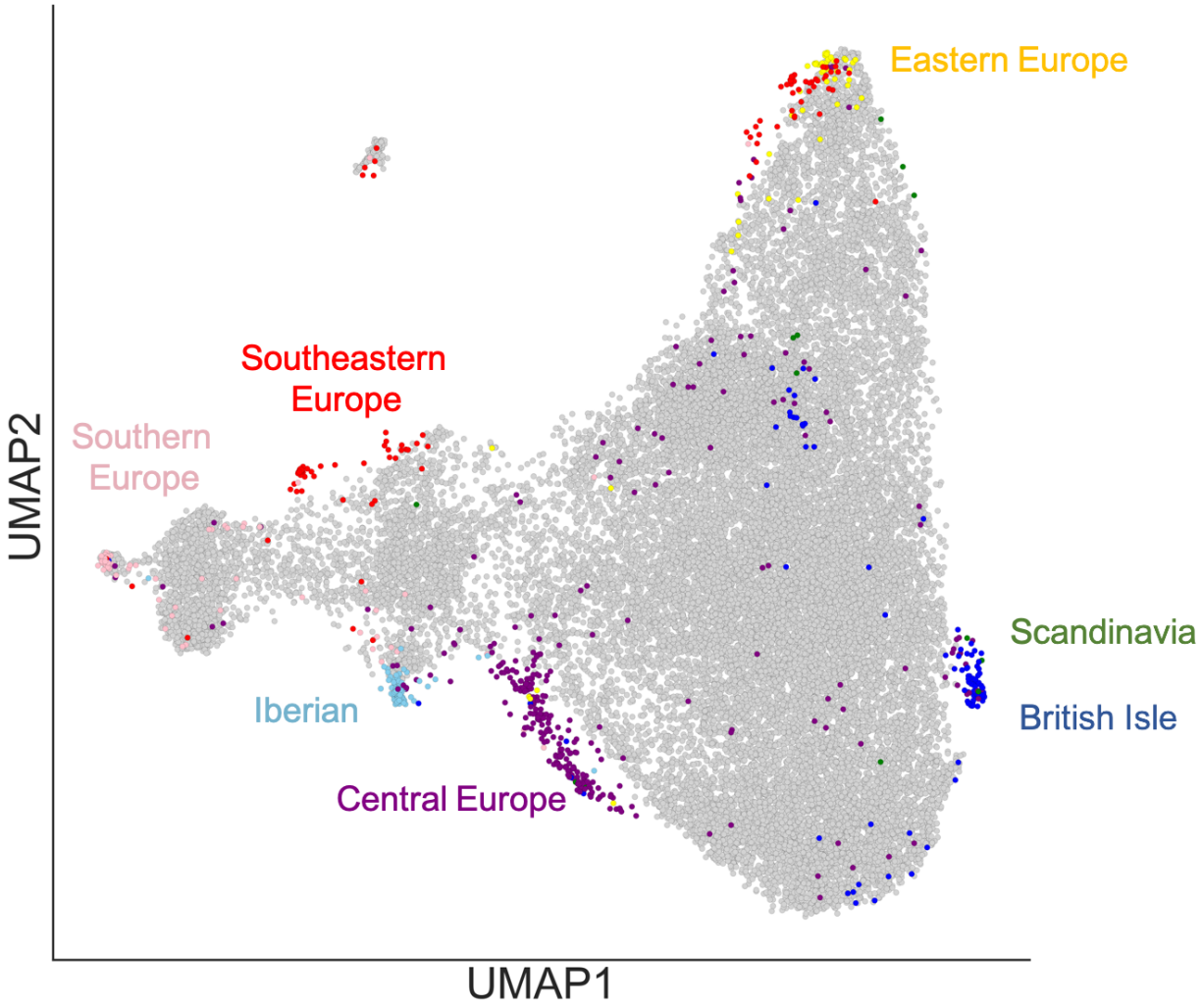


Figure S5. UMAP of Classified Genographic European Americans and POPRES Reference Samples.

UMAP projection of the first 20 PCs. PCs were calculated by first finding the PCs of the POPRES reference samples and then projecting the Random Forest classified Europeans in the Genographic cohort. Each dot represents one individual. Southeast Europeans = Croatia, Yugoslavia, Bosnia-Herzegovina, Serbia, Romania, Hungary, Albania, Macedonia; Central Europe = Switzerland, France, Germany, Germany, Swiss-Italian, Belgium, Swiss-French, Netherlands, Swiss-German; British Isle = Scotland, Ireland, United Kingdom; South Europe = Italy, Cyprus, Turkey, Greece; Iberian = Portugal, Spain; Eastern Europe = Austria, Czech Republic, Poland, Russia; Scandinavia = Sweden, Norway.

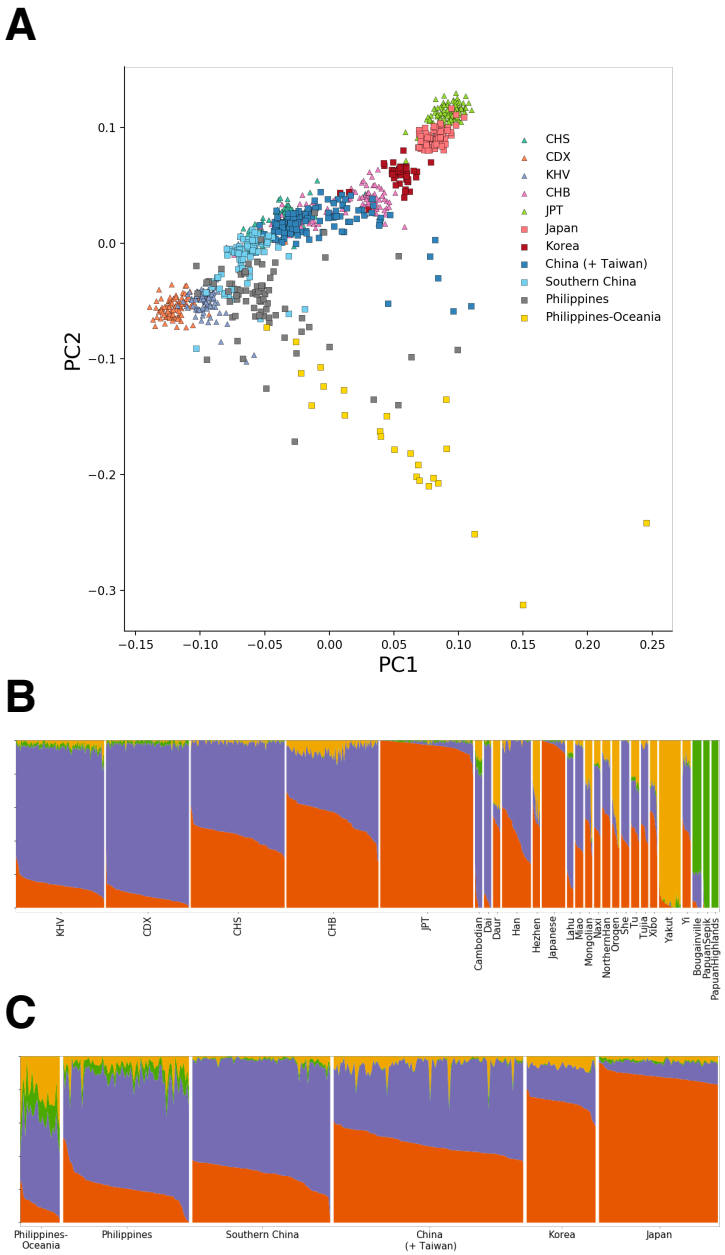


Figure S6. PCA and ADMIXTURE Analysis of East Asians

(A) PCA analysis of classified unrelated East Asian Genographic individuals (plotted in squares) with East Asian samples from 1000 Genome Project (plotted in triangles). Genographic individuals are colored based on fineSTRUCTURE grouping (clade-level) while 1000 Genome Project Samples are colored based on super population.

(B) ADMIXTURE analysis of East Asian 1000 Genome Project samples (left five sections) and East Asia and Oceania HGDP samples (right 21 sections)

(C) ADMIXTURE analysis of classified East Asian Genographic individuals, grouped by fineSTRUCTURE clades.

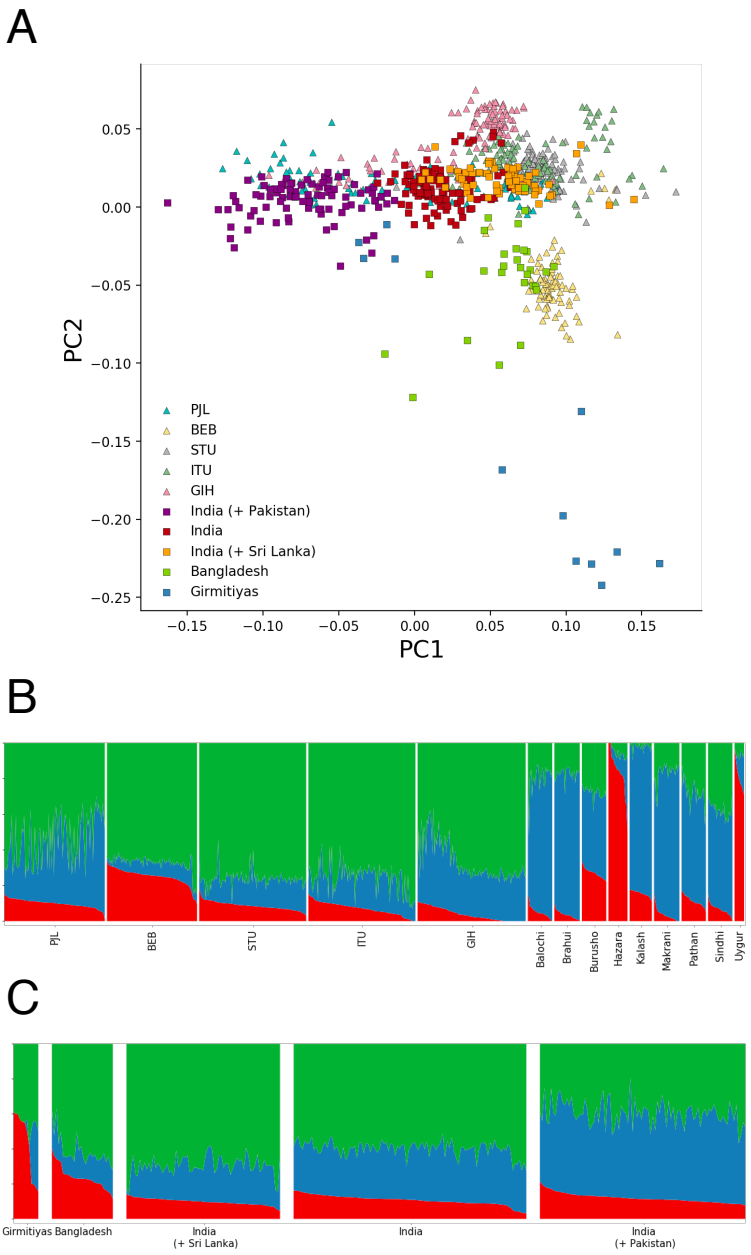


Figure S7. PCA and ADMIXTURE Analysis of South Asians

(A) PCA analysis of classified unrelated South Asian Genographic individuals (plotted in squares) with South Asian samples from 1000 Genome Project (plotted in triangles). Genographic individuals are colored based on fineSTRUCTURE grouping (clade-level) while 1000 Genome Project Samples are colored based on super population.

(B) ADMIXTURE analysis of South Asian 1000 Genome Project samples (left five sections) and Central & South Asia HGDP samples (right nine sections)

(C) ADMIXTURE analysis of classified South Asian Genographic individuals, grouped by fineSTRUCTURE clades.

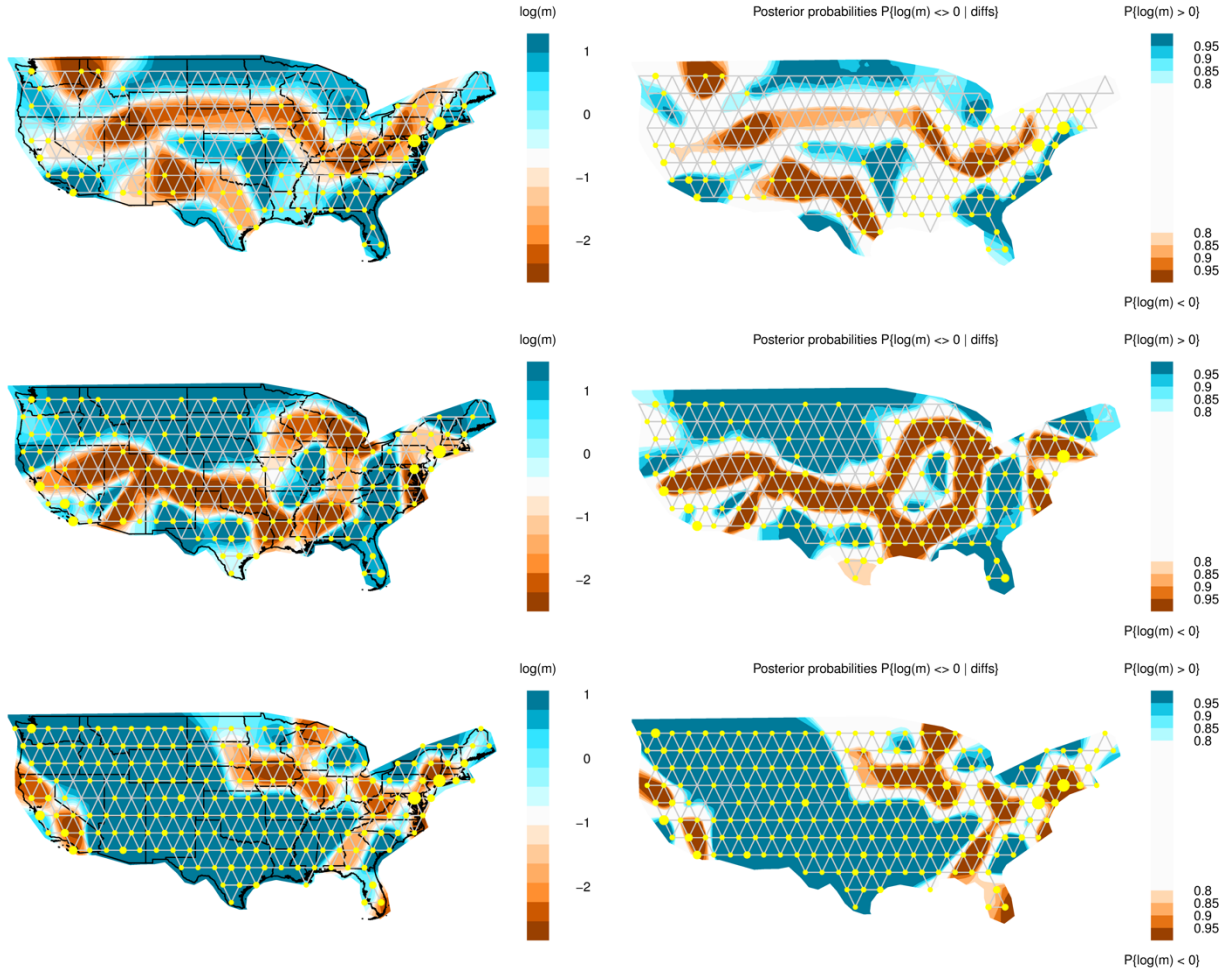


Figure S8. Estimated Effective Migration Surfaces with 250 Demes and Posterior Probabilities.

Figures on the left represent migration rates inferred using EEMS: African Americans (top left), Hispanics/Latinos (middle left), and Europeans (bottom left). Colors and values correspond to inferred rates, m, relative to the overall migration rate across the country. Shades of blue indicate logarithmically higher migration (i.e. log(m) = 1 represents effective migration that is tenfold faster than the average) while shades of orange indicate migration barriers. Figures on the right represent the inferred posterior probabilities (>80%) of relative effective migration: African Americans (top right), Hispanics/Latinos (middle right), and Europeans (bottom right). Darker shades of blue represent greater probability that the relative migration is greater than average while darker shades of orange represent greater probability that the relative migration is lower than average (i.e. migration barrier). Each individual is snapped to a vertex, which is represented by yellow points. The size of points corresponds to the size of the subpopulation at the vertex.

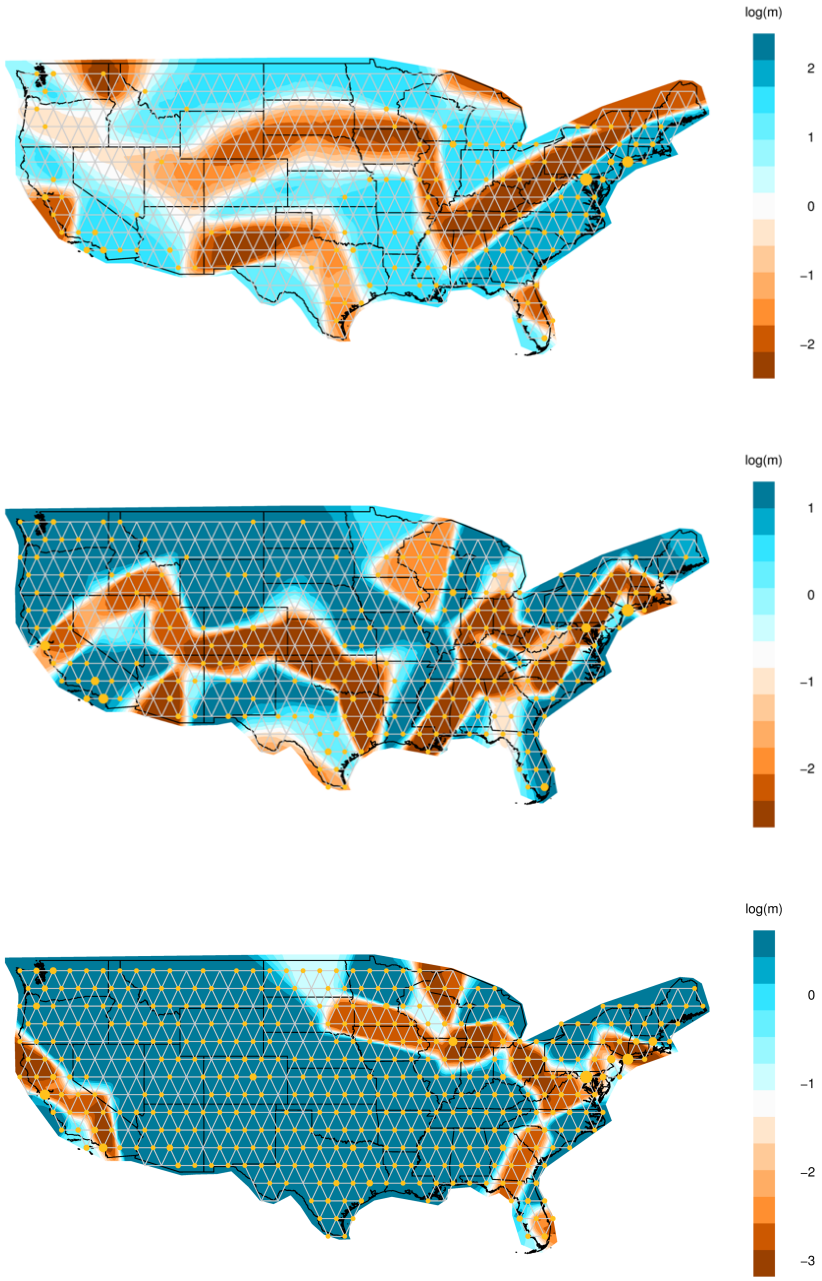


Figure S9. Estimated Effective Migration Surfaces with 500 Demes.

Inferred effective migration rates of African Americans (top), Hispanics/Latinos (middle), and Europeans (bottom) using 500 demes reveal similar patterns to 250 demes. Similar to above, colors and values correspond to inferred rates, m, relative to the overall migration rate across the country.

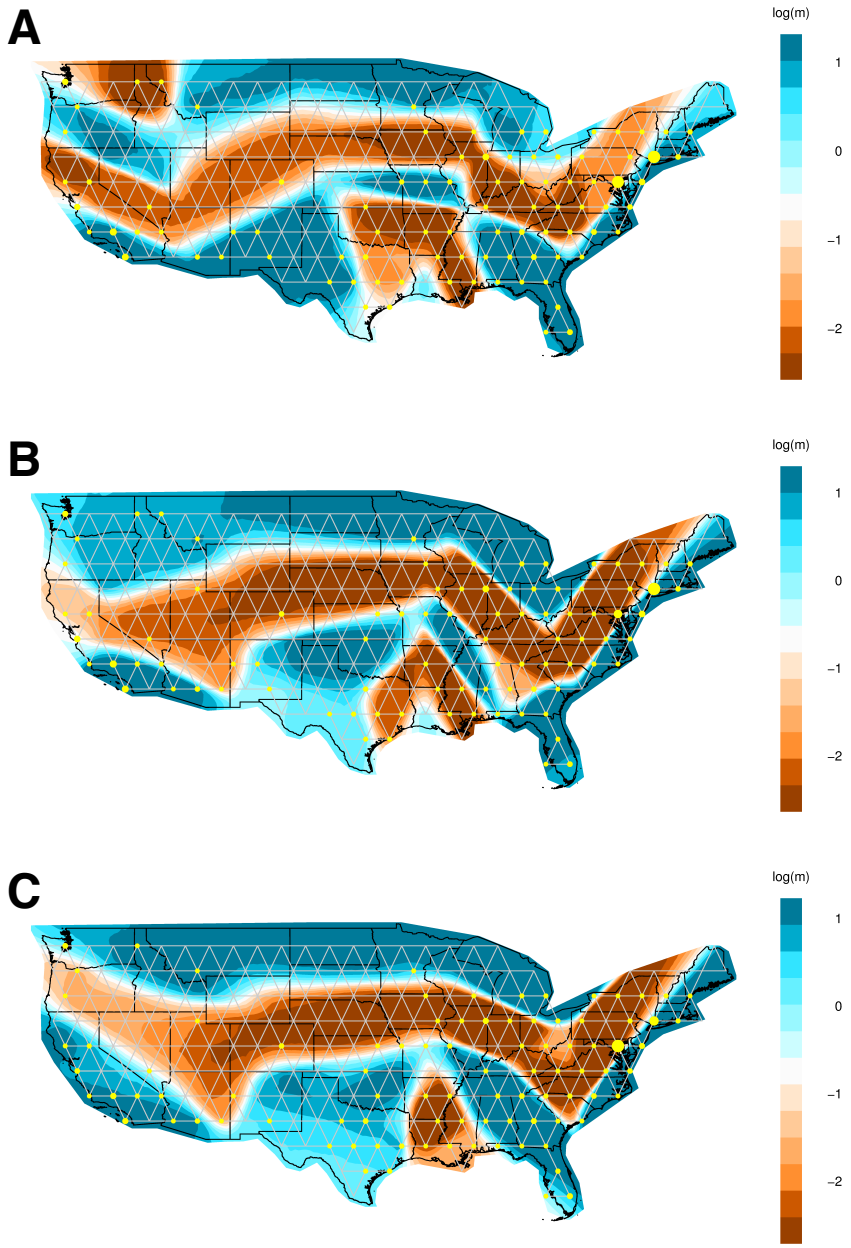


Figure S10. Comparison of Inferred Migration Surfaces with Different Sampling Schemes. **(A)** Random subsampling of classified African American individuals to 80% of the original size. **(B)** Even sampling across the four major US Census Regions. African American individuals were subsampled to 80% of original sample size by selecting evenly across all Census Region so that each was represented in equal proportions in the final sample set. **(C)** Oversampling of the South. Since African Americans are populous in the South, we subsampled African American individuals to 80% of the original sample size by selecting half of the final samples from the south and the other half evenly from the remaining regions.

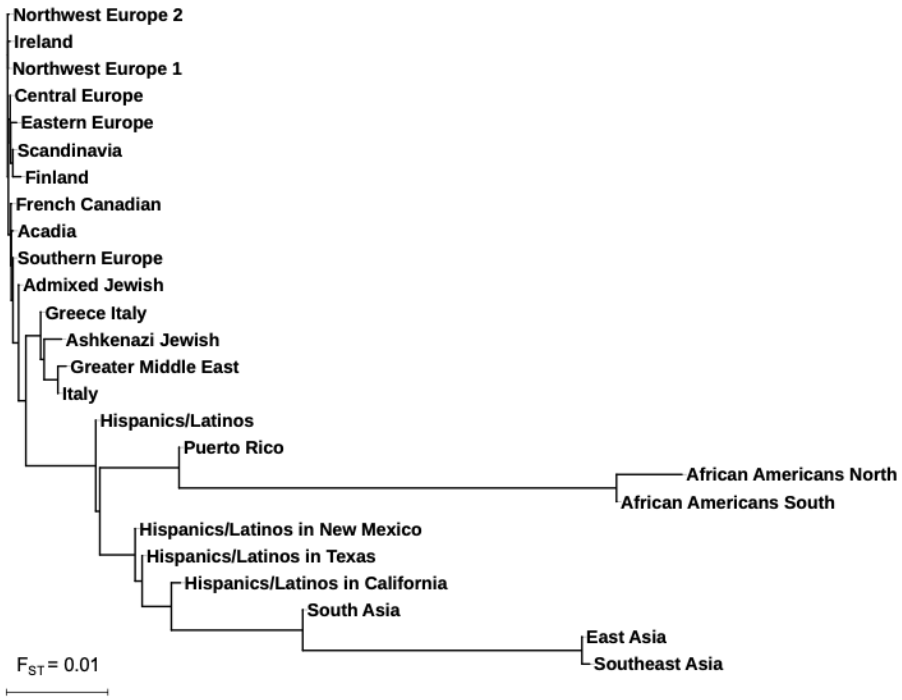


Figure S11. Genetic Differentiation of Haplotype Clusters

Unrooted phylogenetic tree of haplotype clusters was constructed using the neighbor joining method with F_{ST} as genetic distance. Negative branch lengths were converted to zero.

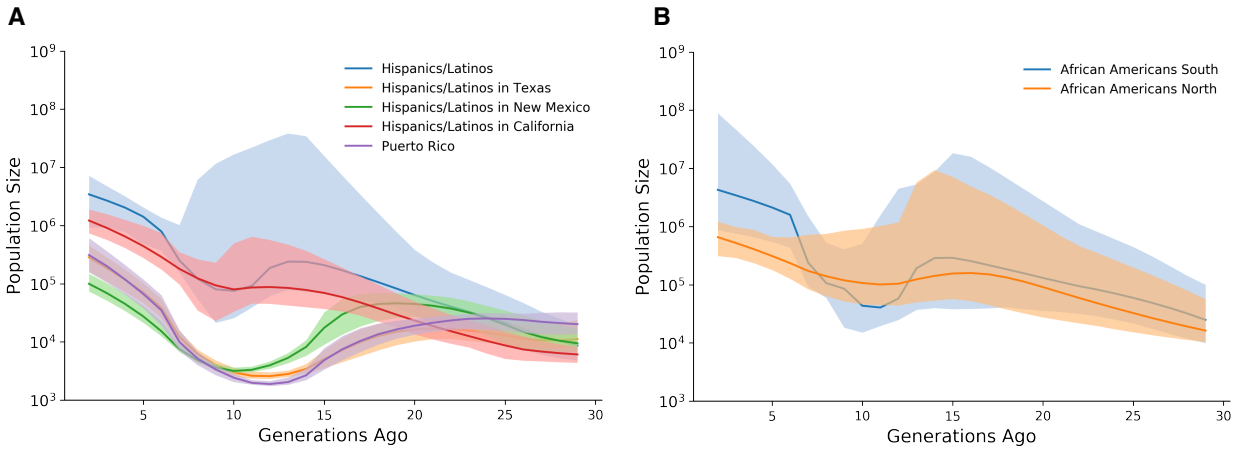


Figure S12. Effective Population Size over Antecedent Generations

(A) Evidence of populations bottlenecks are present in the Hispanics-related clusters, with many of them occurring 8-14 generations ago. Despite being more admixed with other ancestries, the Hispanics/Latinos and Hispanics/Latinos in California cluster still shows some signs of population bottleneck, but to a lesser degree than the other clusters. Inferred effective population size are shown in solid lines while 95% confidence intervals are displayed in lighter shades.

(B) Population bottleneck is evident in the African American South cluster, while the the African American North cluster does not show much of evidence of a bottleneck, potentially due to the lower sharing of IBD and relatedness between individuals in the cluster.

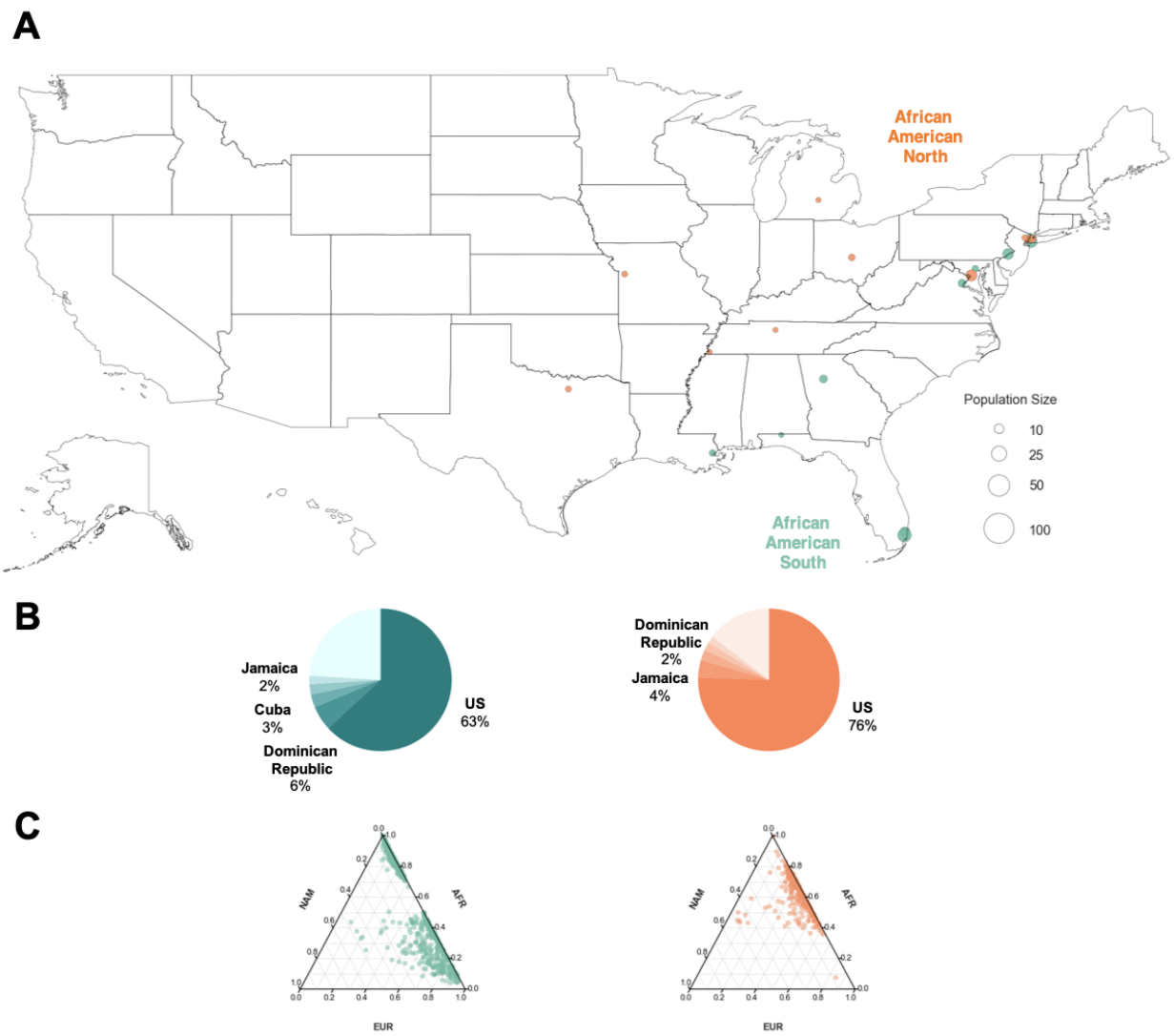


Figure S13. Distribution of African American Haplotype Clusters

(A) Map of haplotype clusters corresponding to Africans ancestries. Each county containing present-day individuals is represented by a dot. The top 10 locations with the highest odds ratio are shown for each cluster. Maps showing the full distribution for each cluster can be found in **Figure S18**.

(B) Ancestral birth origin proportions for each cluster in (A). Only individuals with complete pedigree annotations, up to grandparent level, are shown.

(C) Ternary plots of ancestry proportions based on local ancestry inference for each haplotype cluster. Each dot represents one individual. Variations in the proportion of African ancestry amongst African Americans in the Genographic Project are consistent with previous studies.[1,2] However, the mean proportion of African ancestry is slightly lower, potentially due to sampling bias.

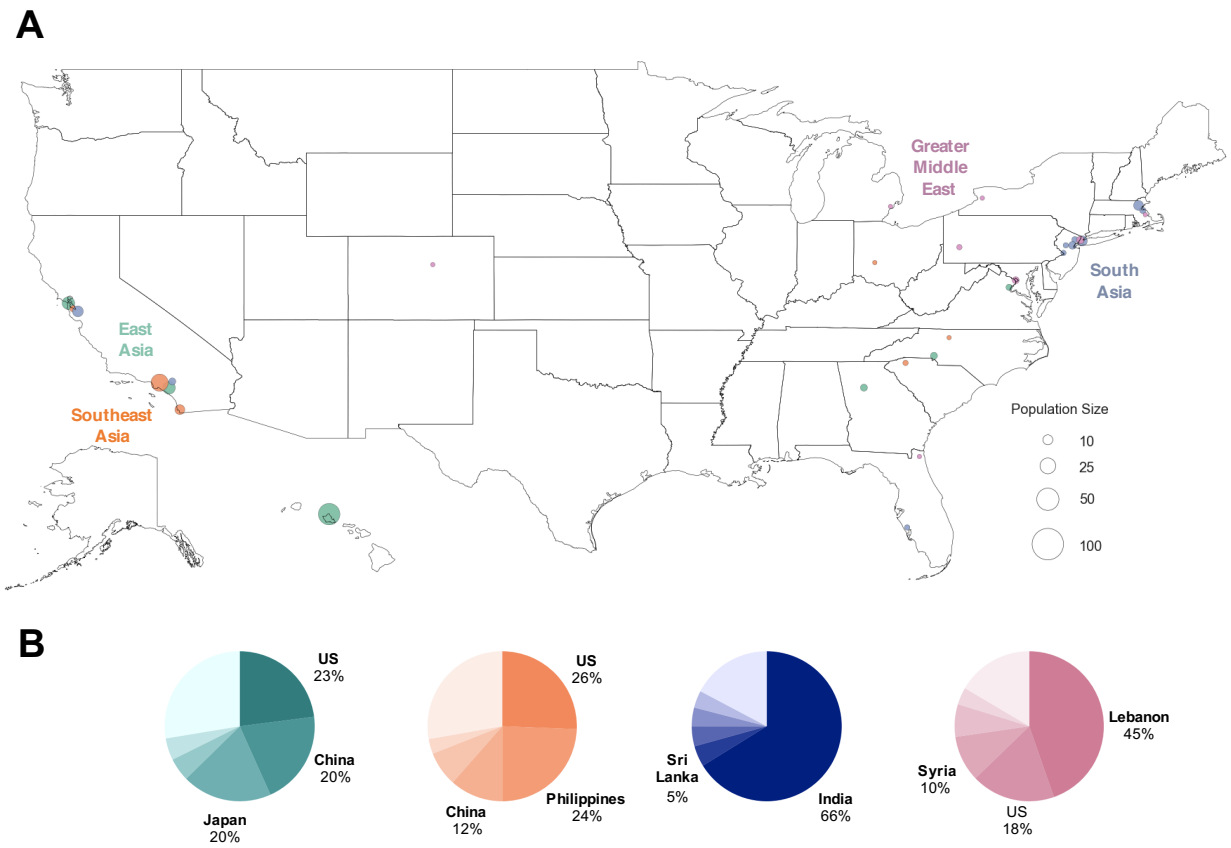
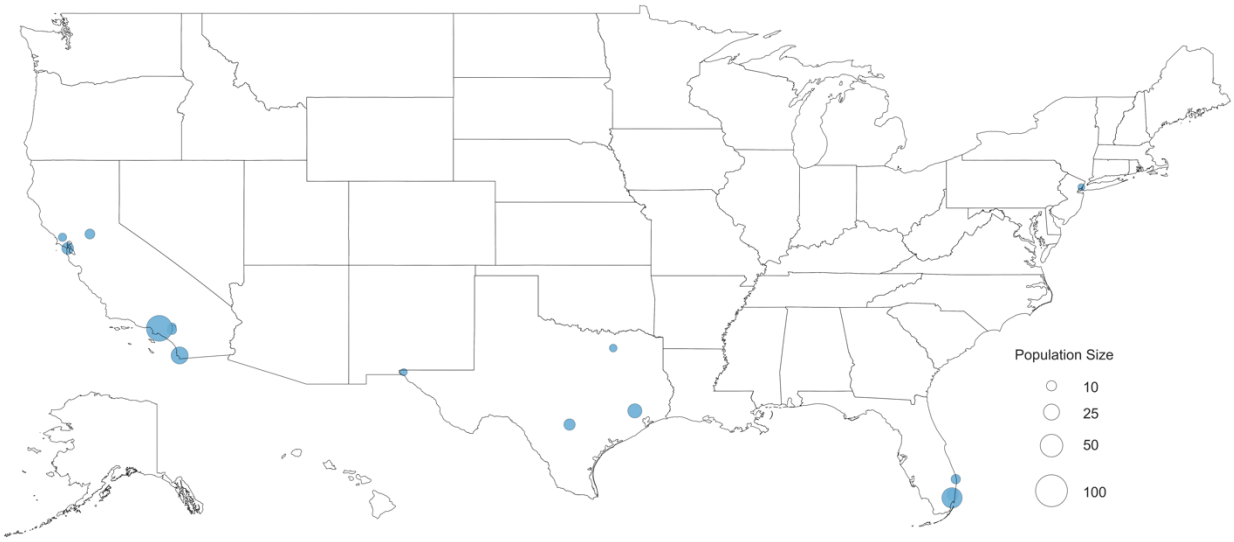


Figure S14. Distribution of Haplotype Clusters with Asian Ancestries

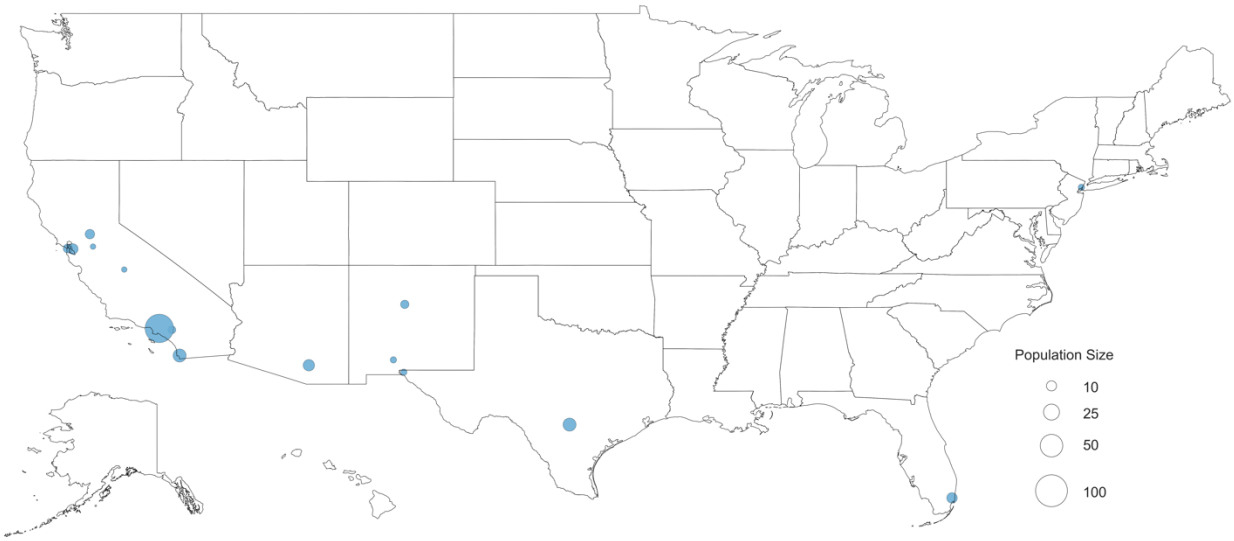
(A) Map of haplotype clusters corresponding to regional Asian ancestries. Each county containing present-day individuals is represented by a dot. The top 10 locations with the highest odds ratio are shown for each cluster. Maps showing the full distribution for each cluster can be found in **Figure S19**.

(B) Ancestral birth origin proportions for each cluster in (A). Only individuals with complete pedigree annotations, up to grandparent level, are shown.

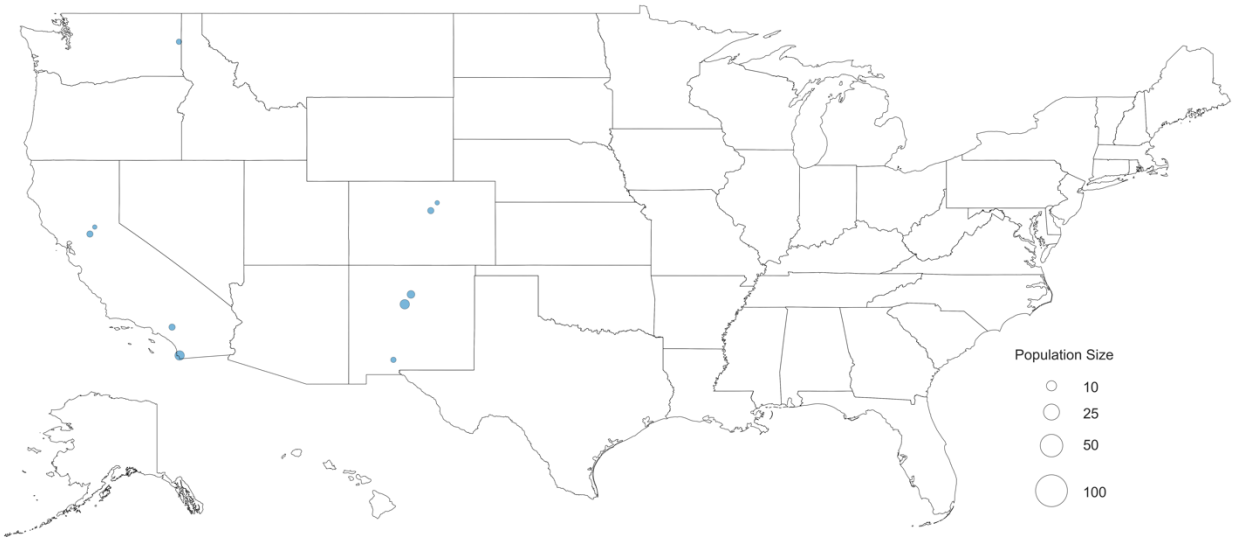
Hispanics/Latinos



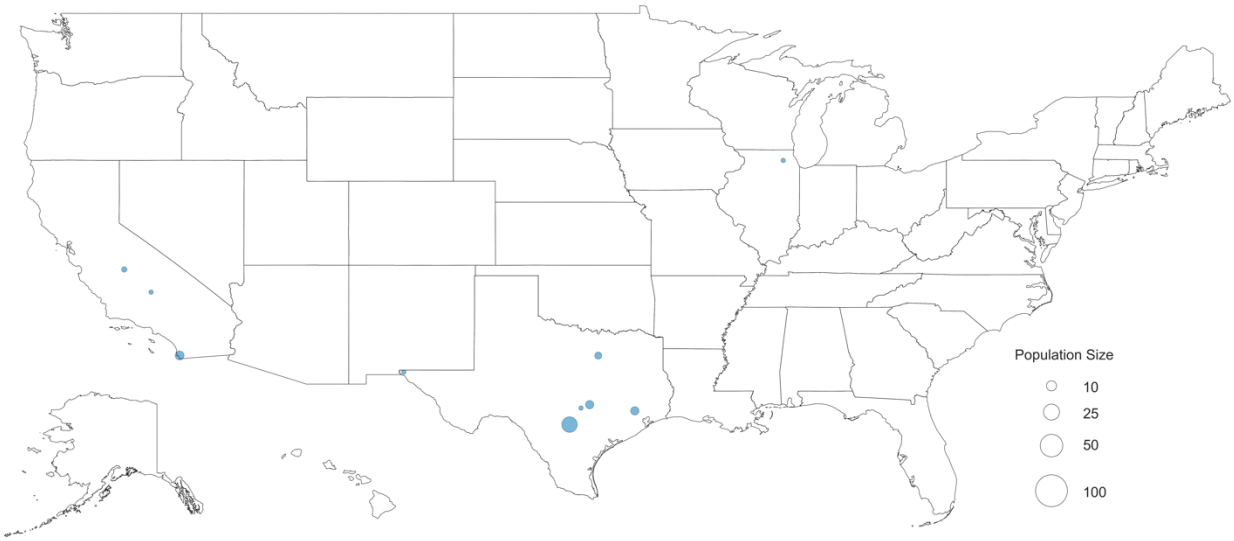
Hispanics/Latinos in California



Hispanics/Latinos in New Mexico



Hispanics/Latinos in Texas



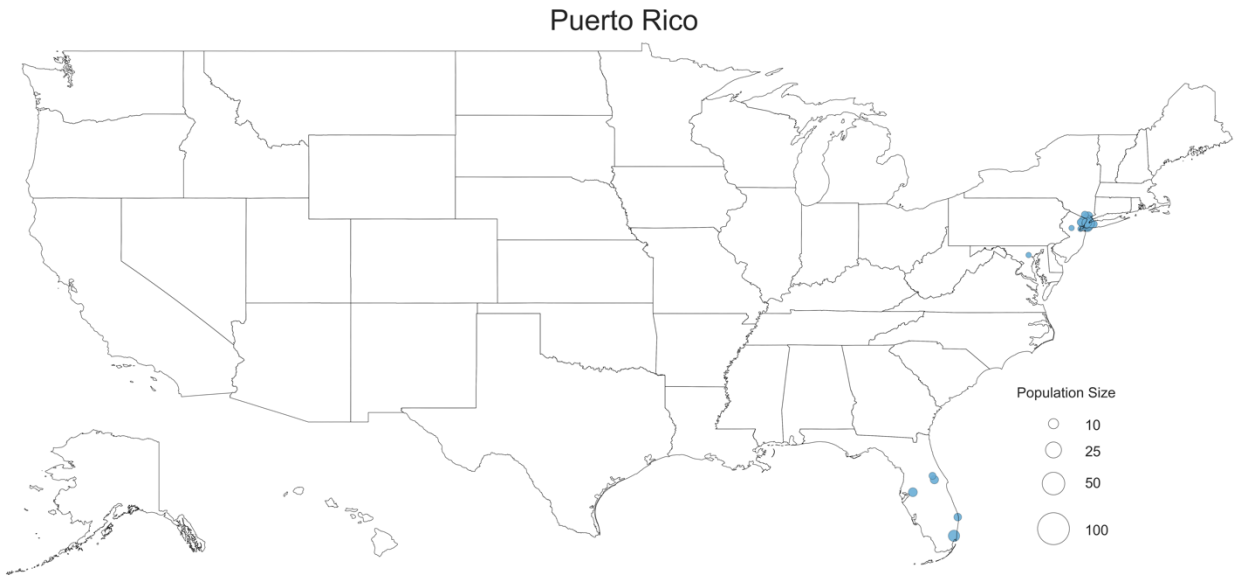
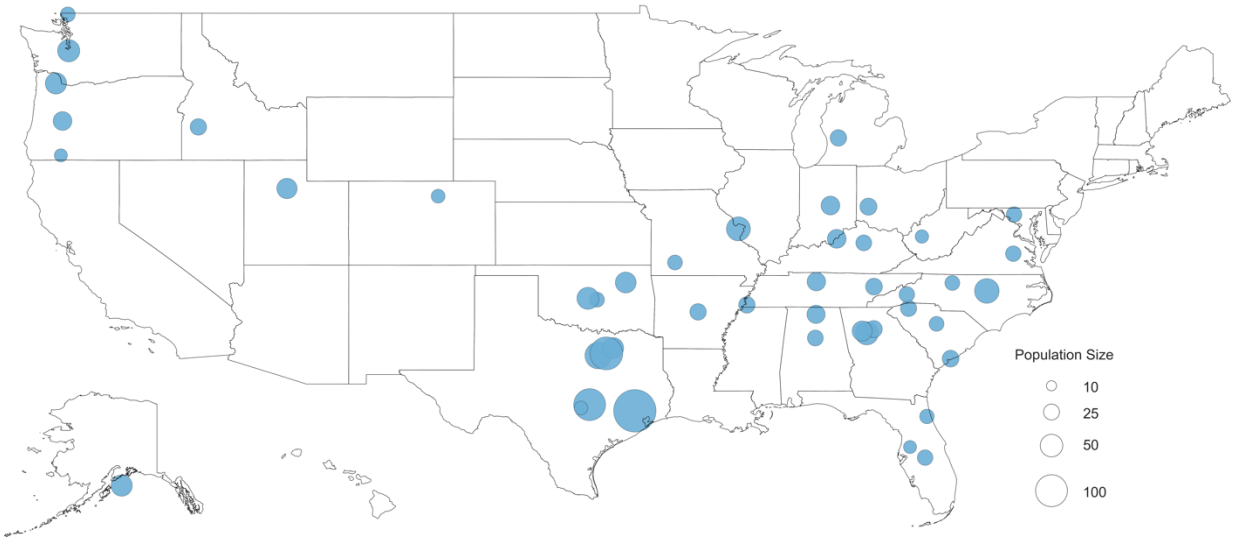


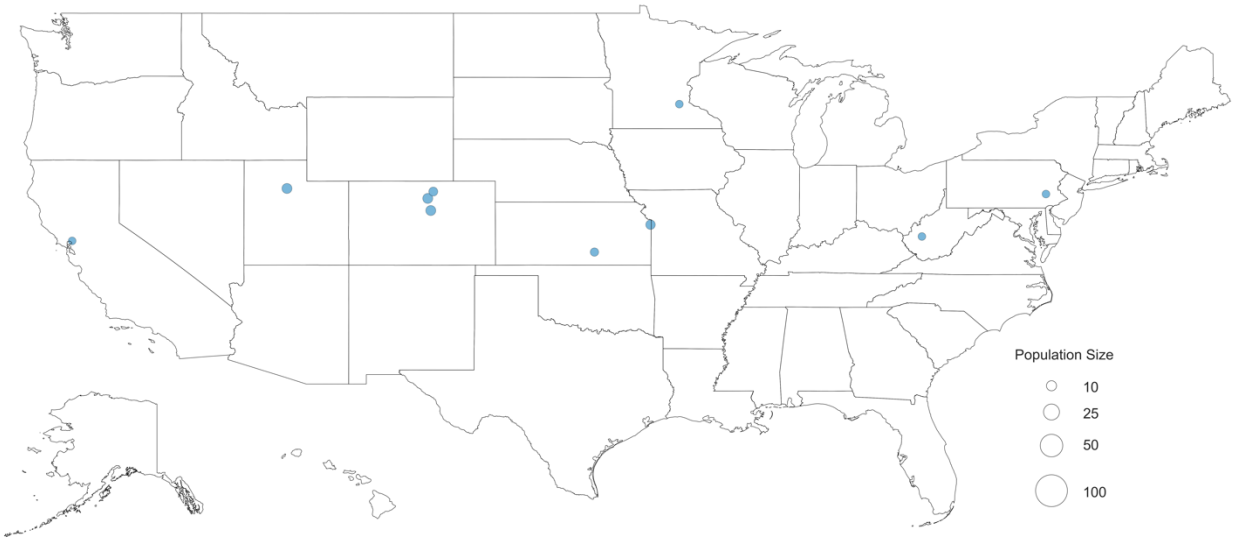
Figure S15. Geographical Distribution of Hispanic/Latino Haplotype Clusters

The five Hispanic/Latino-related clusters we identified recapitulate the state-by-state differences of the Hispanics population as reported in the US Census.[3] The presence of the Hispanics/Latinos and the Puerto Rican cluster in Florida are consistent with the large proportions of Hispanics in Florida reporting Puerto Rican (20%) and Cuban (29%) origin in US Census. Similarly, the distribution of the Puerto Rican cluster around New York City is in line with the high proportions (31%) of Hispanics/Latinos reporting Puerto Rican origin in New York state. In Southwestern states, smaller proportions of Hispanics/Latinos reporting Central and South America origins are found in Arizona than in neighboring California (3% in Arizona versus 10% in California) in the US Census, consistent with our ancestral birth origin data.

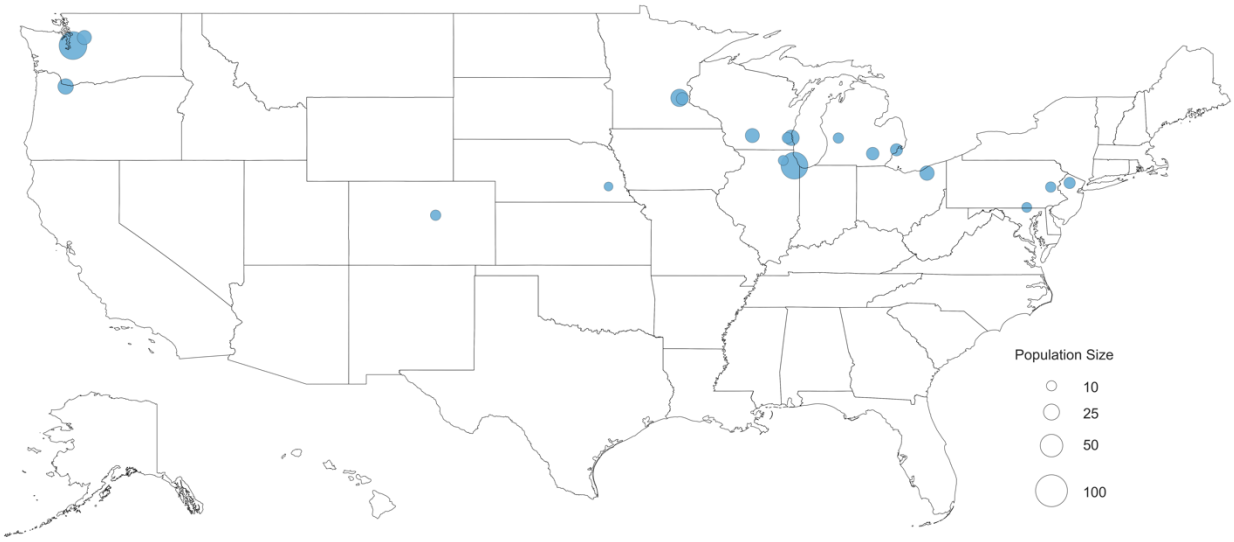
Northwest Europe 1



Northwest Europe 2



Central Europe



Ireland

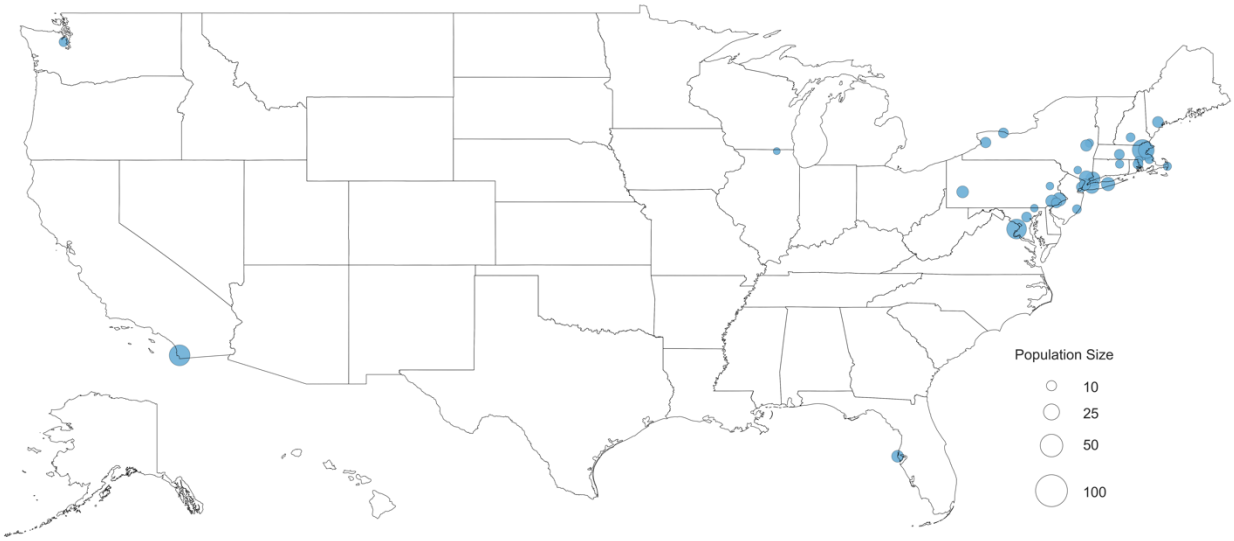
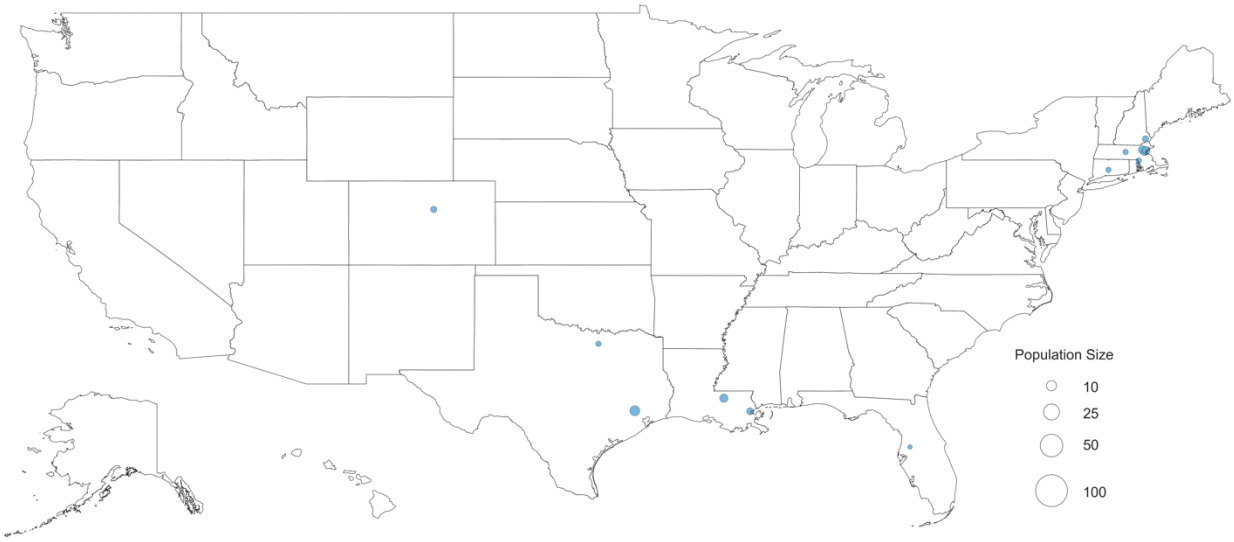


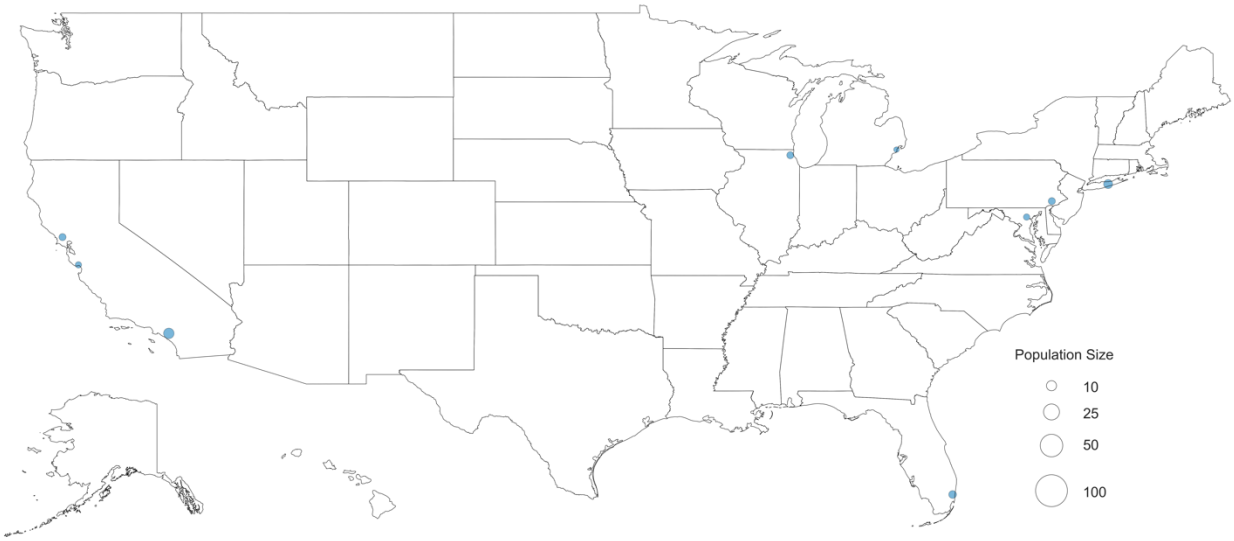
Figure S16. Geographical Distribution of European American Haplotype Clusters

Present-day location of individuals in each cluster. Each county is represented by a dot and only the counties with a significant odds ratio ($p < 0.05$) are shown for each cluster. These European haplotype clusters reflect broad regional ancestries, as corresponding birth origins are not clearly overrepresented in any particular country. The exception is the cluster of Irish individuals ("Ireland"). During the 19th and early 20th centuries, millions of Irish immigrants entered into the US, which experienced religious tensions and discrimination and resulting in high rates of in-group marriage amongst Irish individuals.[4] Nonetheless, present-day Irish Americans remain genetically similar to other Europeans from the central and northwestern parts of Europe.

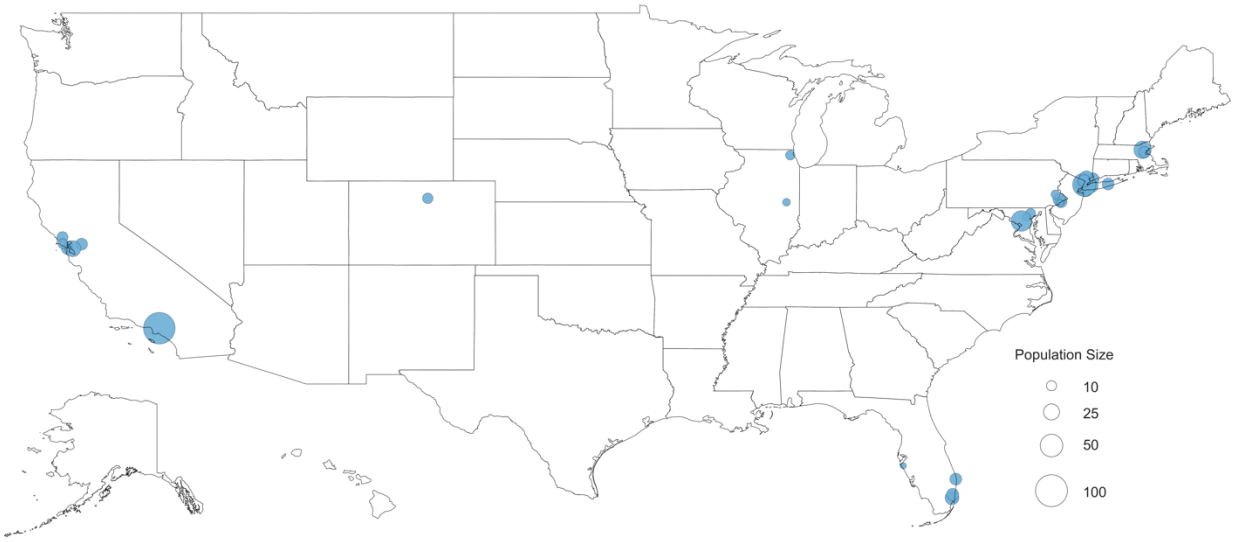
Acadia



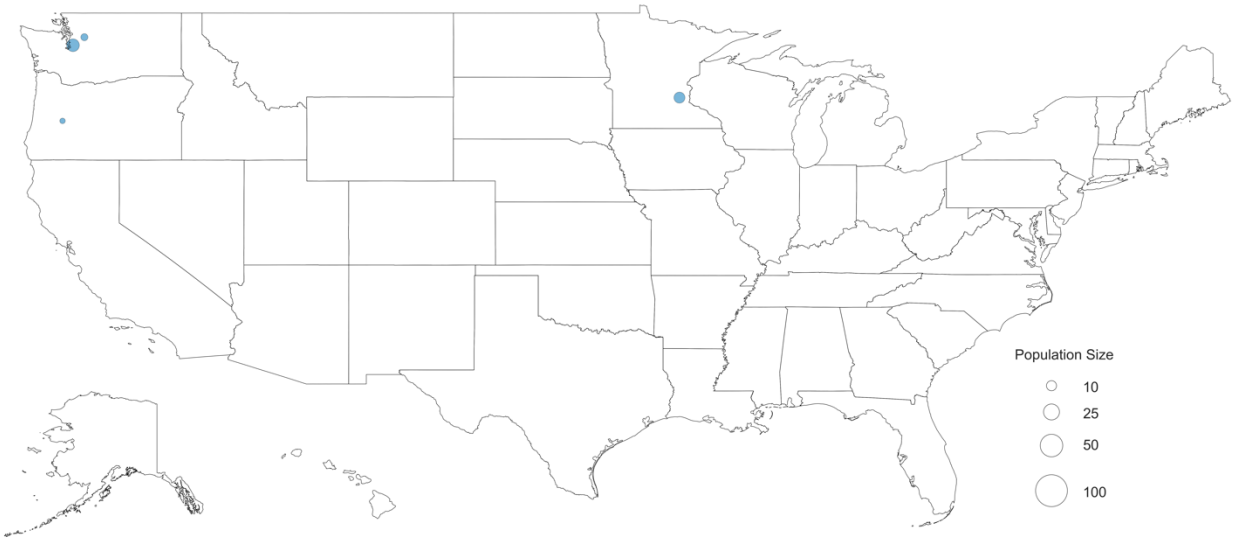
Admixed Jewish



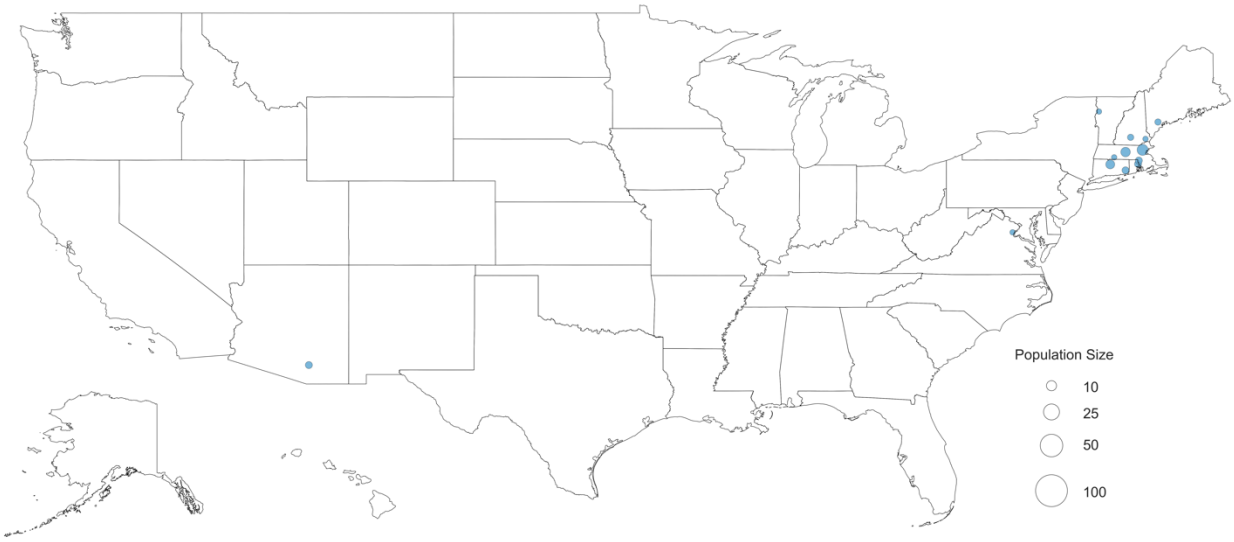
Ashkenazi Jewish



Finland



French Canadian



Greece-Italy

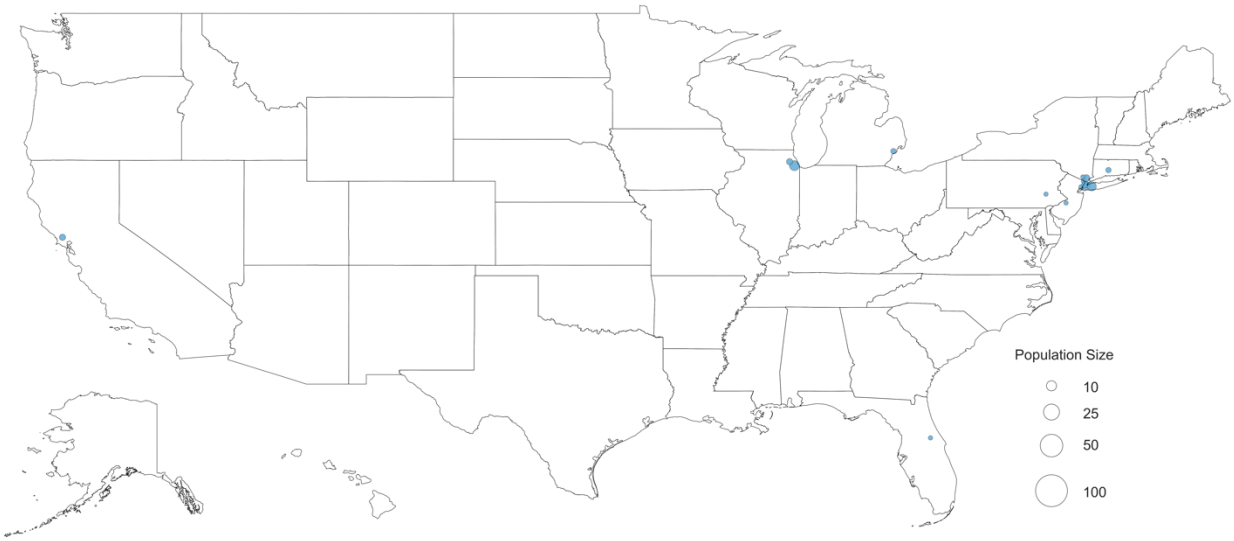


Figure S17. Geographical Distribution of Genetically-Differentiated European American Haplotype Clusters

Consistent with previous analysis,[5] we identify clusters of Scandinavians, Finns, French Canadians, Acadians, Ashkenazi Jews, Italians, and Greeks. We also identify a second cluster with Jewish ancestry ("Admixed Jews"). Unlike the Ashkenazi Jewish cluster, self-reported ethnicity suggests admixture between Jewish and non-Jewish ancestry individuals, as Jewish-ancestry is typically present only on one side of the family. Present-day location of individuals in each cluster. Each county is represented by a dot and only the cluster with a significant odds ratio ($p < 0.05$) are shown for each cluster.

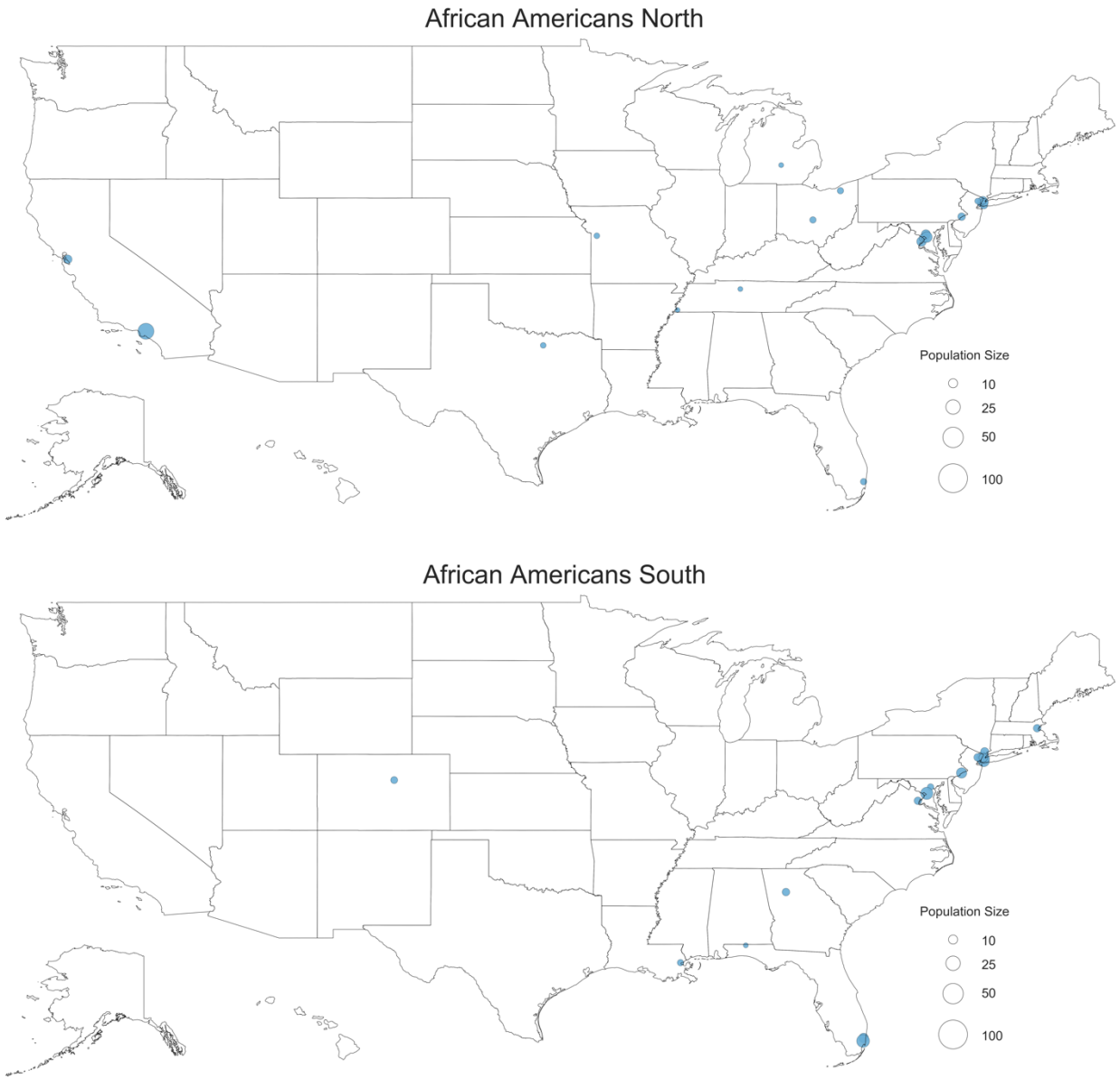
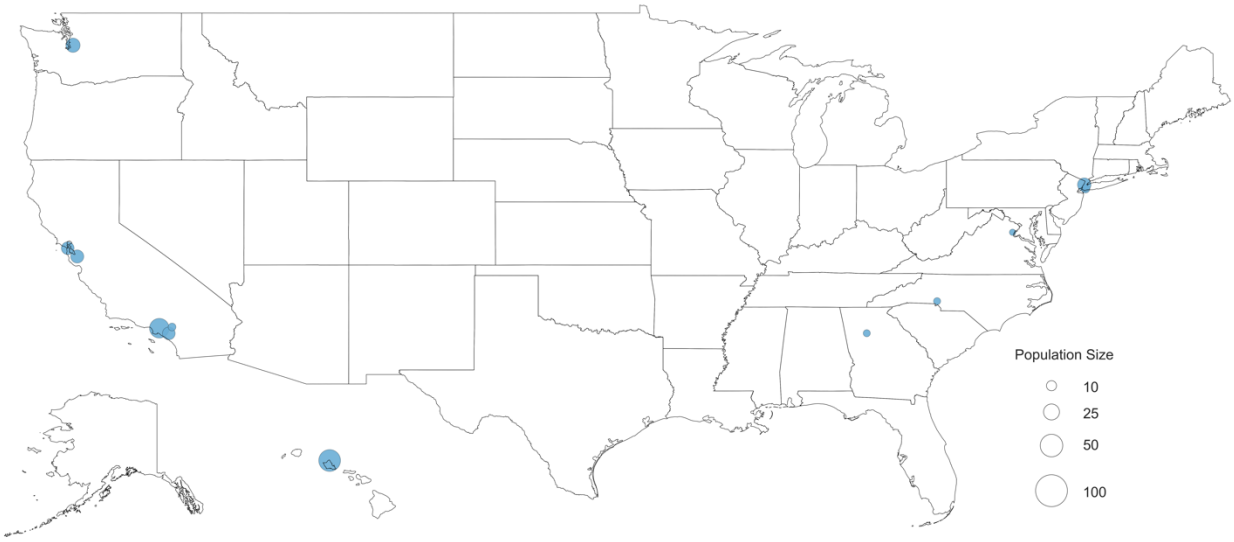


Figure S18. Geographical Distribution of African American Haplotype Clusters

Present-day location of individuals in each African American cluster. Each county is represented by a dot and only the counties with significant odds ratios ($p < 0.05$) are shown for each cluster.

East Asia



Greater Middle East

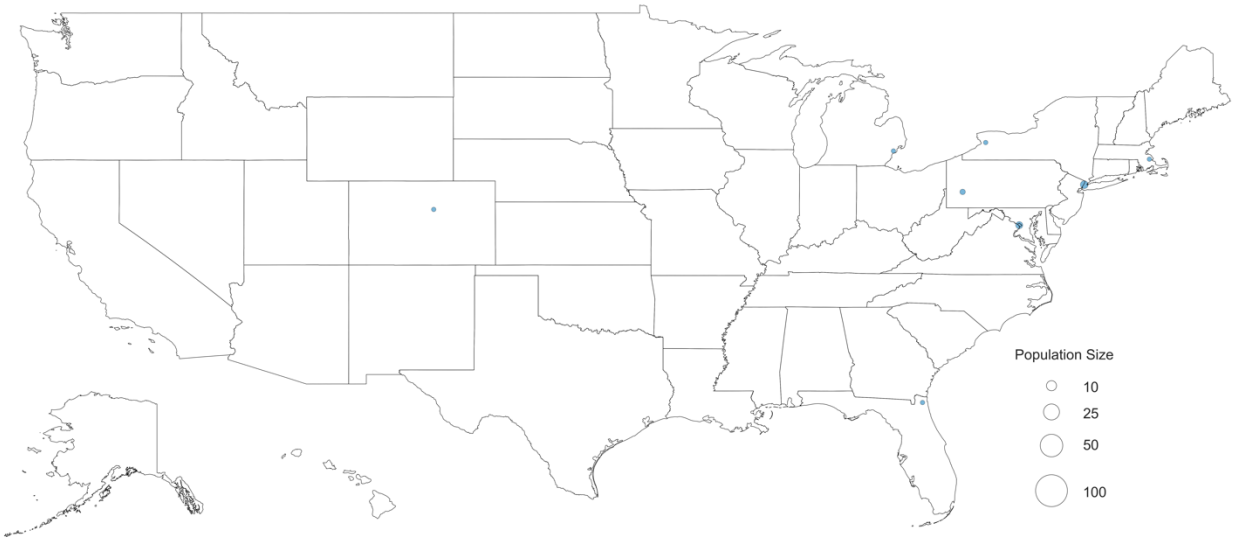


Figure S19. Geographical Distribution of Asian Haplotype Clusters

The ancestral origins and geographic distributions of these clusters are consistent with US Census reports. Since these populations descend from more recent immigrants, the observed patterns of homozygosity within several of these clusters likely reflect consanguinity patterns in some of their ancestral regions. Present-day location of individuals in each cluster. Each county is represented by a dot and only the counties with significant odds ratios ($p < 0.05$) are shown for each cluster.

Supplemental Materials and Methods

Self-reported Ancestral Birth Origin and Ethnicity

As part of the registration process to track and access the results of their DNA sample on Genographic Project website, Genographic participants were given the option to report birth origin data and ethnicity data on themselves, their parents, and their grandparents. A total of 24,566 individuals (75.4%) provided complete data (i.e. no missing data for any ancestors), resulting in 171,962 pedigree records. All analysis using ancestral birth origin and ethnicity data was performed using data at the grandparent level. Birth origin data was recorded at the country level, with the exception of certain territories and regions being listed separately. Participants provided ancestral birth origin data by selecting from a list of countries for each ancestor. Ethnicity data was provided in the form of free text and was therefore not standardized across participants, making aggregating and comparing self-reported ethnicity data challenging.

It is important to note that ancestry, ethnicity, and race are all complex terms that result from many factors, including appearance, culture, socioeconomics, geography, etc. The definition of these terms across individuals and populations depending on various social, cultural, religious, and economic factors. Therefore, ancestry, ethnicity, and race are not directly comparable, and there are limitations to comparing genetic ancestry with data on race and ethnicity from the US Census. For example, population genetic studies often analyze Hispanic/Latino, European American, and African American individuals separately.[1,5] The US Census, however, classifies race and ethnicity (specifically Hispanics) to be two separate and distinct concepts; Hispanics/Latinos may be of any race.[3] As such, comparing the proportion of genetically-classified Hispanic/Latino individuals in the US with the proportion of people declaring Hispanic/Latino origin in US Census is invalid as the percent of Hispanics in the US Census are not independent from the counts and percentages for racial categories.[3] We further note that the separation of race and ethnicity in the US Census has resulted in 43.5% of self-reported Hispanics not identifying with any of the race category in the US Census, approximately three times higher than the non-response rate for the total U.S. population.[6] This trend was observed independently in a separate survey study,[7] suggesting that while the US government separates Hispanic ethnicity from race, Hispanic individuals do not always self-identify with the current racial categories.

Family Relationship Inference

We used KING v2.0 to identify the set of unrelated individuals within the Genographic dataset separated by at least two degrees of relatedness.[8] 806 individuals had kinship coefficients greater than 0.0884 and were removed for downstream analysis using EEMS and haplotypes.

Coloring of UMAP plots

We colored the 1000 Genome Project samples in the UMAP plot based on their country level assignments (**Figure 1D**) and visualized the Genographic samples by coloring each sample based on their ancestry proportions from ADMIXTURE (**Figure 1E**). Specifically, the color (RGB value) of each Genographic sample is a linear combination of the sample's admixture proportions and the RGB values of each ancestry's color (EUR = red, AFR = yellow, NAM = green, EAS = blue, SAS = purple).

Comparison of filtered and unfiltered haplotype network

We evaluated two networks: one with filtering for minimum or maximum IBD sharing and one with pairs of individuals in which cumulative IBD sharing is ≥12 cM and ≤72 cM, similar to prior analysis.[5] Clustering of haplotype networks resulted in a total of 25 clusters for the filtered network (≥12 cM and ≤72 cM). For the unfiltered network, we arrived at 32 clusters, 4 of which had less than 10 individuals and were removed from subsequent analyses. Annotations for the 25 clusters from the filtered network were found to be more interpretable than annotations for the 28 clusters from the unfiltered networks. Specifically, many of the clusters from the unfiltered networks exhibited similar proportions of ancestral origins or ethnicities and were difficult to differentiate (**Table S6 and S7**). Certain populations (e.g. Finns, Middle Easterners) found from the filtered network were also not identified from the unfiltered network. We therefore used the 25 clusters from the filtered network in downstream analyses.

Supplemental References

1. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., and Mountain, J.L. (2015). The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet. *96*, 37–53.

2. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., Doumbo, O., et al. (2009). The Genetic Structure and History of Africans and African Americans. Science *324*, 1035–1044.

3. US Census Bureau About Hispanic Origin.

4. Funchion, M.F. (2010). Ties that Bind: Ethnic and Religious Factors in the Marriage Choices of Irish-American Catholics on the Dakota Frontier. New Hibernia Rev. Iris Éireannach Nua *14*, 121–142.

5. Han, E., Carbonetto, P., Curtis, R.E., Wang, Y., Granka, J.M., Byrnes, J., Noto, K., Kermany, A.R., Myres, N.M., Barber, M.J., et al. (2017). Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nat. Commun. *8*, 14238.

6. Ríos, M., Romero, F., and Ramírez, R. (2013). Race Reporting Among Hispanics: 2010 (US Census Bureau, Population Division).

7. NW, 1615 L. St, Suite 800Washington, and Inquiries, D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. (2012). When Labels Don’t Fit: Hispanics and Their Views of Identity.

8. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.