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

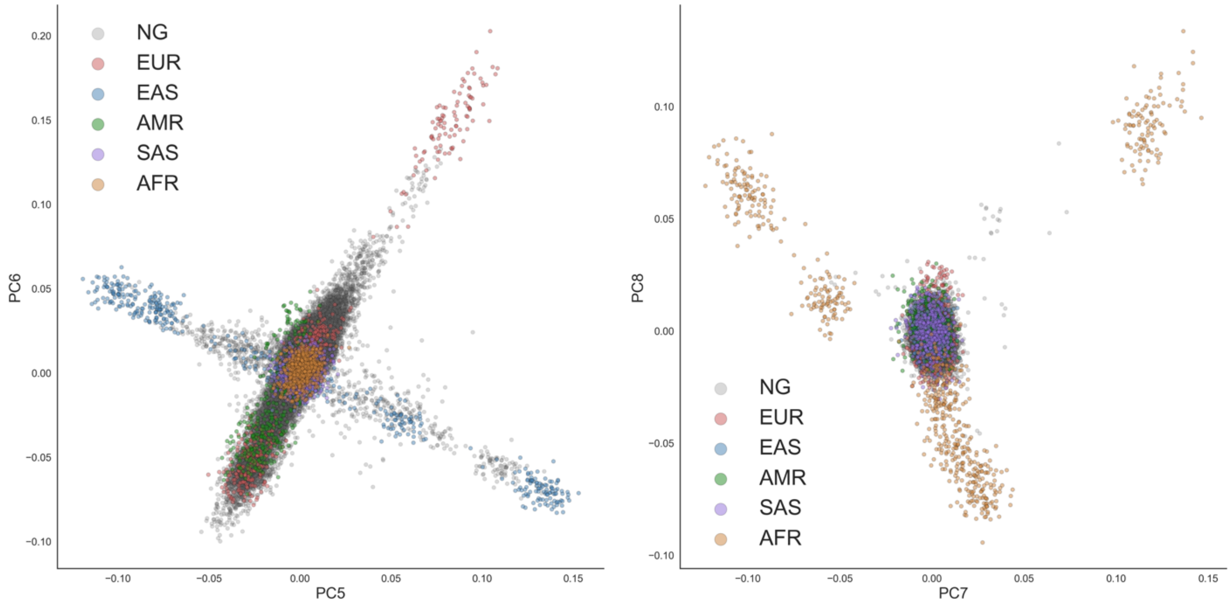
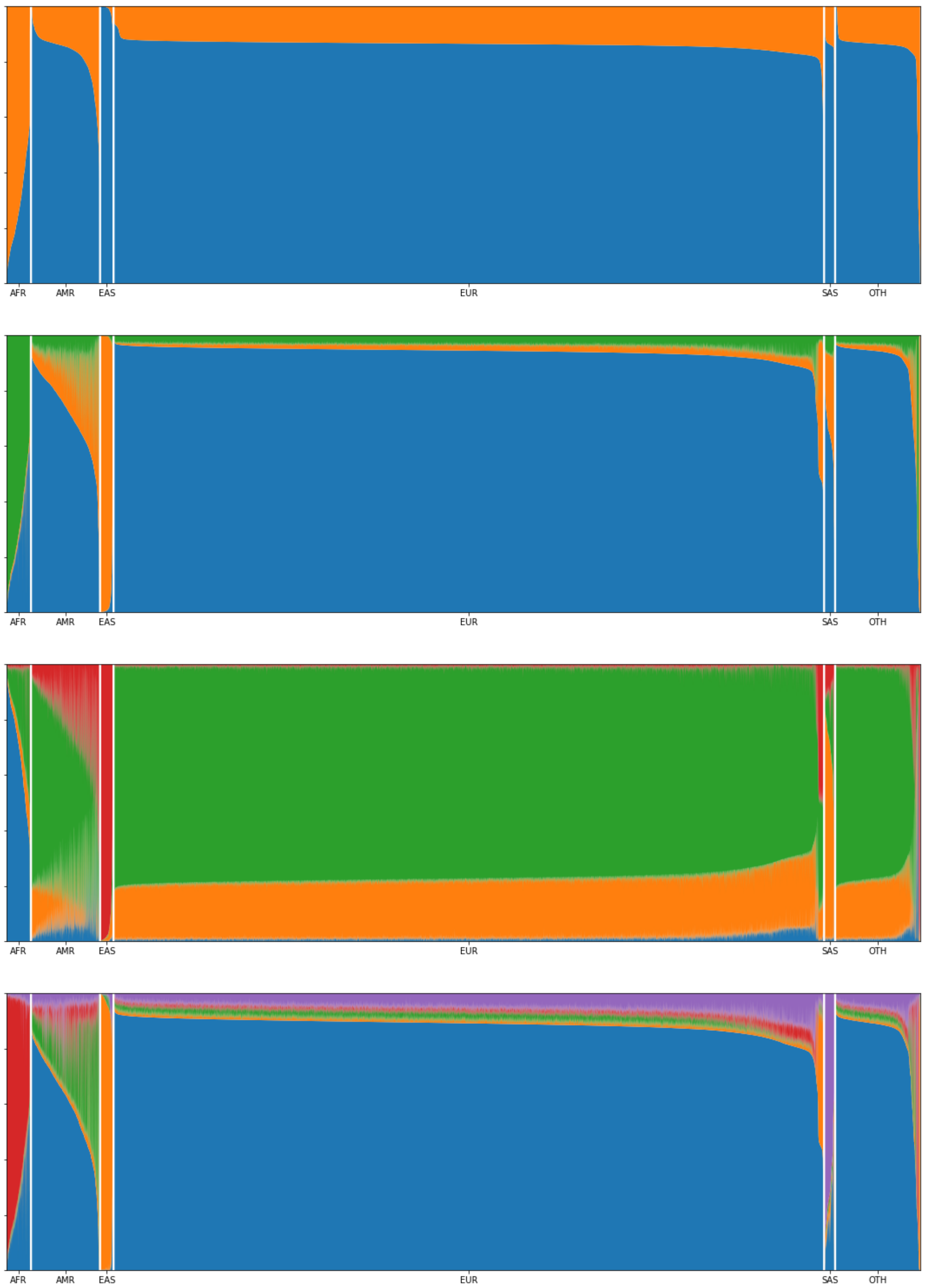


Figure S1. Principal Component Analysis of 1000 Genome Project and Genographic samples

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

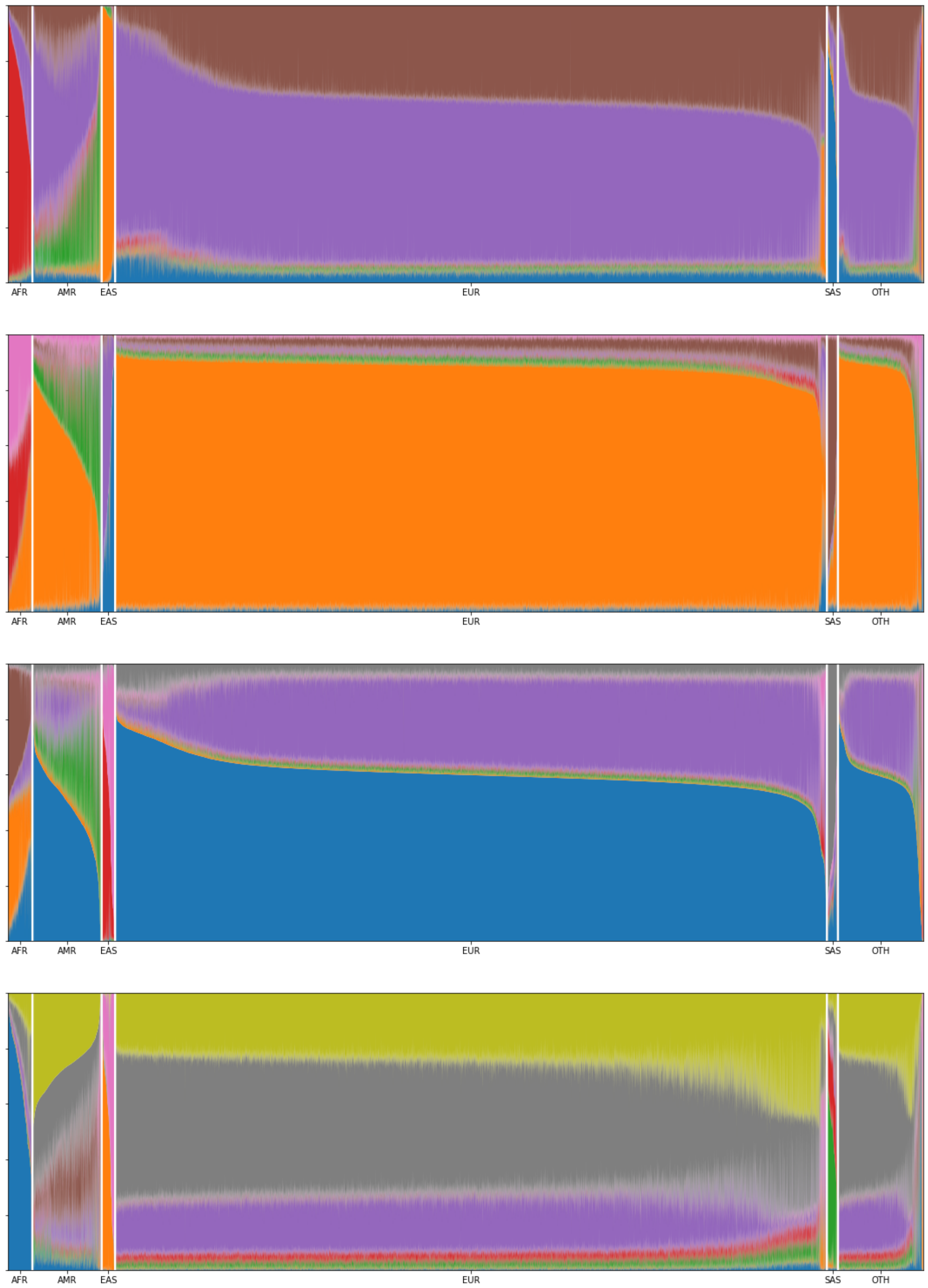


Figure S2. ADMIXTURE from K = 2 to 9.

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.

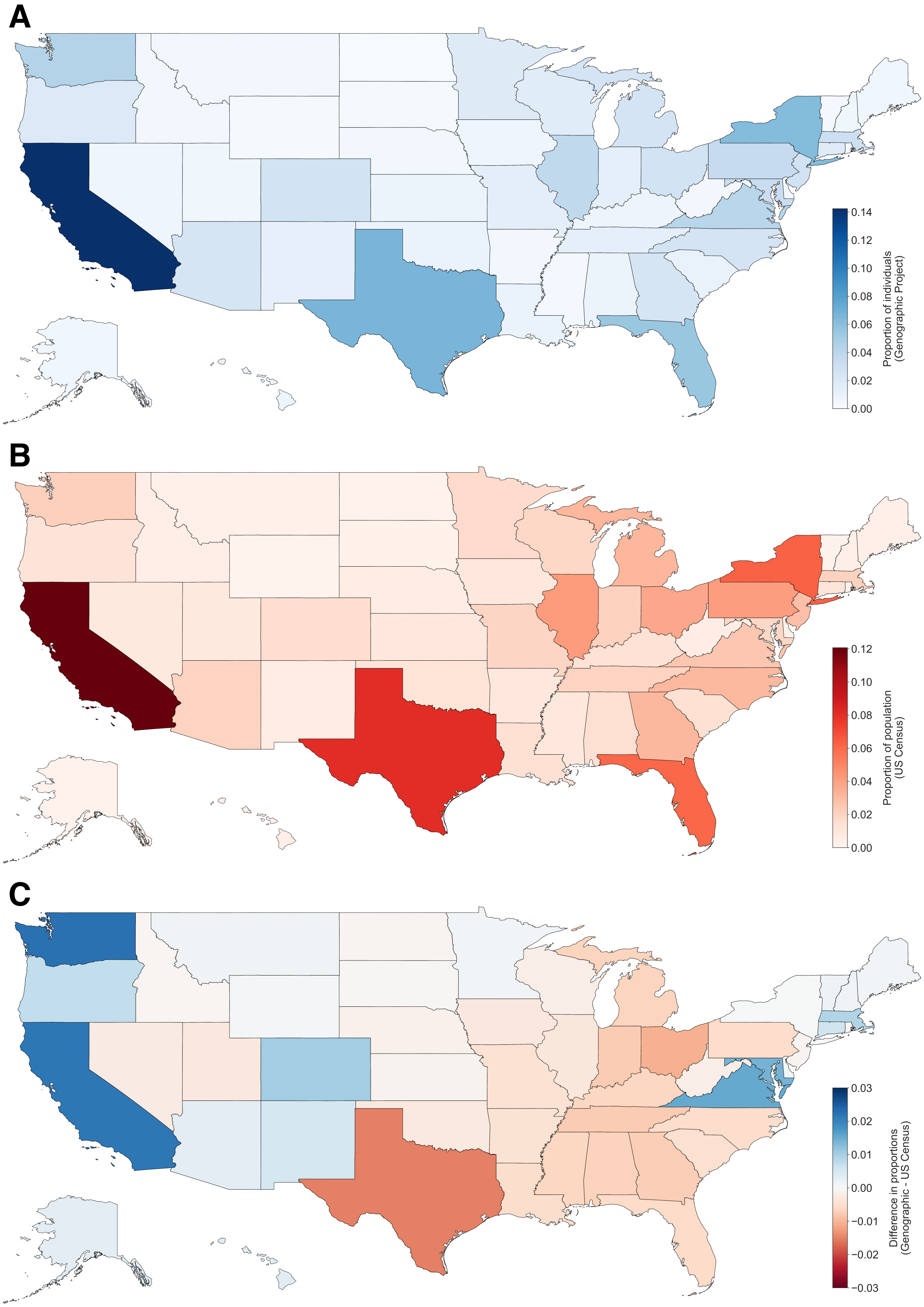

A
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Proportion of individuals
(Genographic Project)
B
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Proportion of population
(US Census)
C
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
Difference in proportions
(Genographic - US Census)

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.

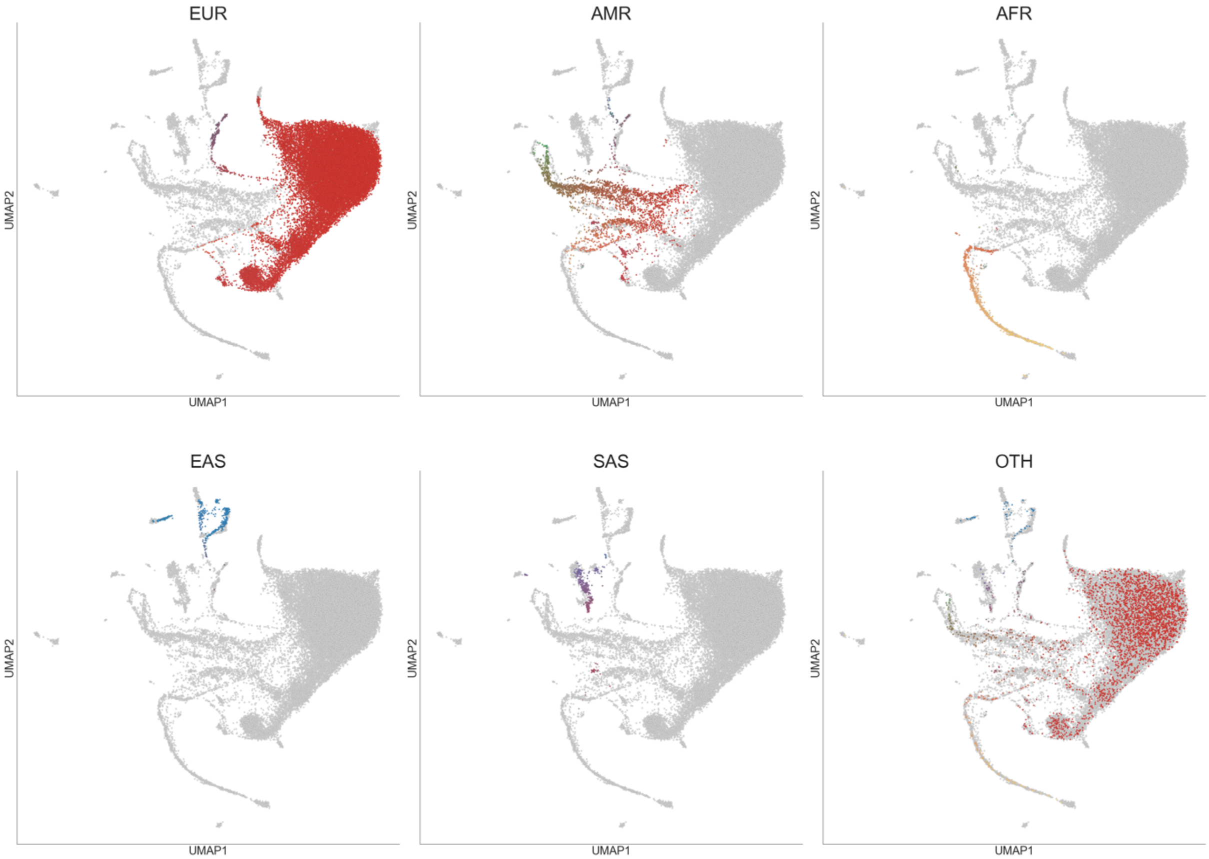


Figure S4. Uniform Manifold Approximation and Projection (UMAP) of Classified Genographic Individuals

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.

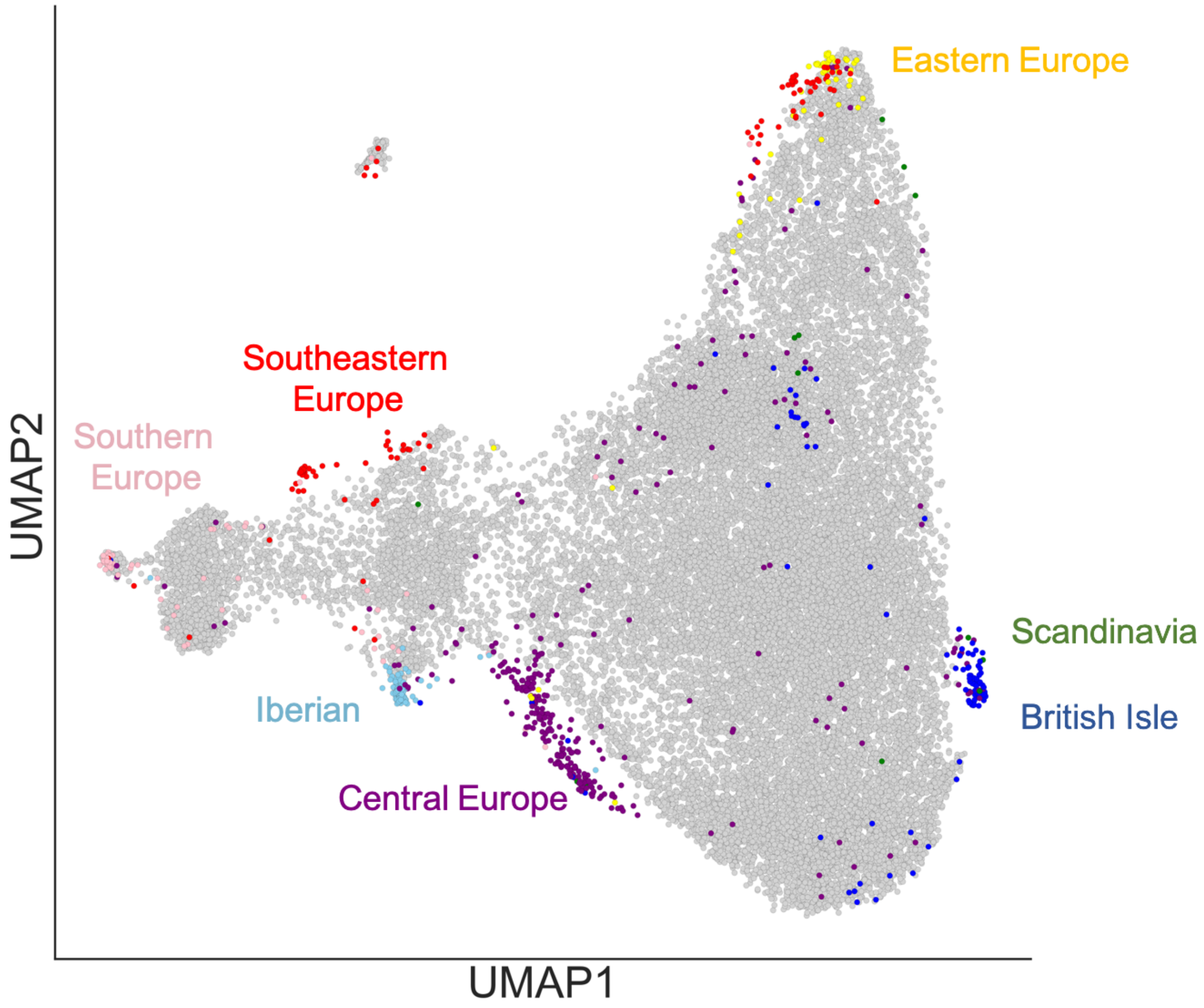


Figure S5. UMAP of Classified Genographic European Americans and POPRES Reference Samples.

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.

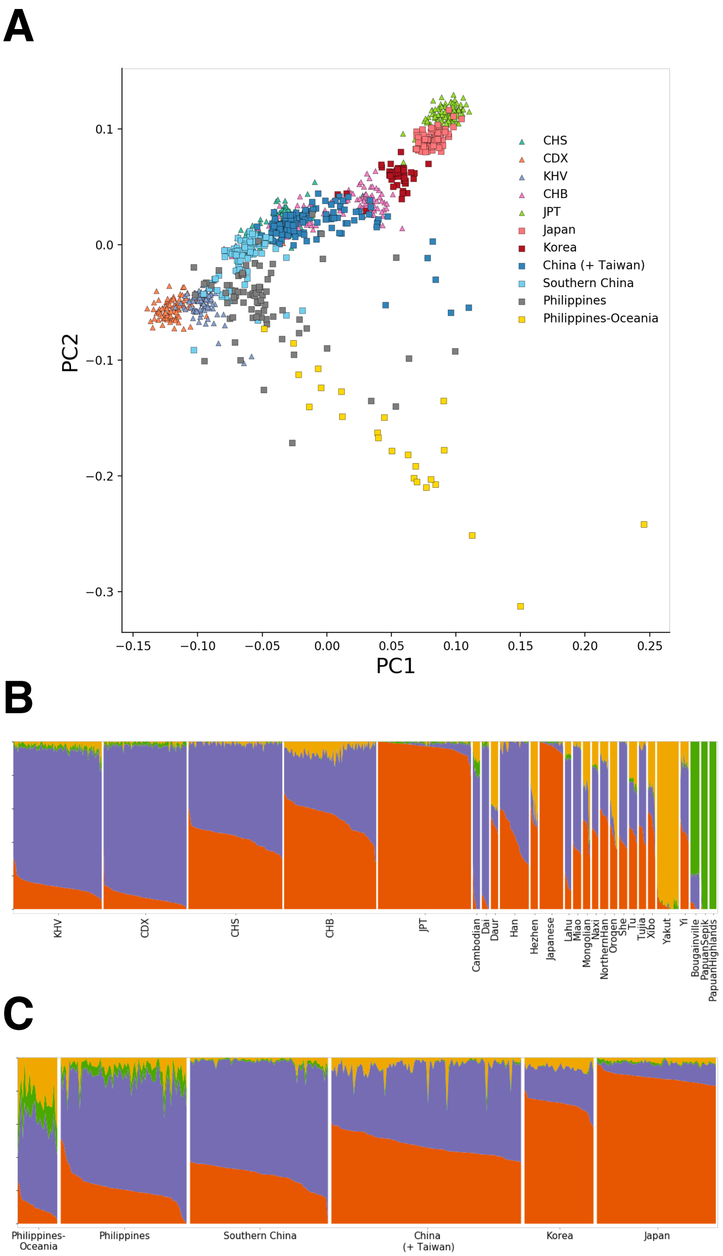


Figure S6. PCA and ADMIXTURE Analysis of East Asians

(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.

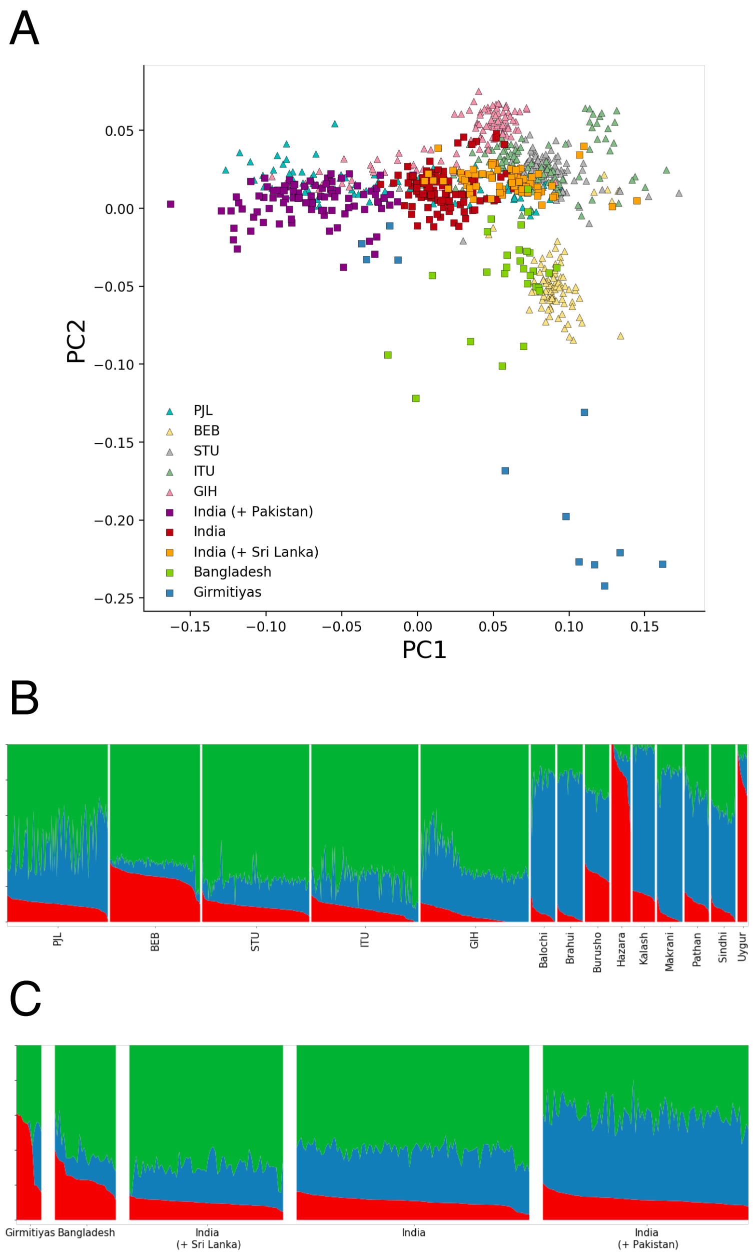


Figure S7. PCA and ADMIXTURE Analysis of South Asians

(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.

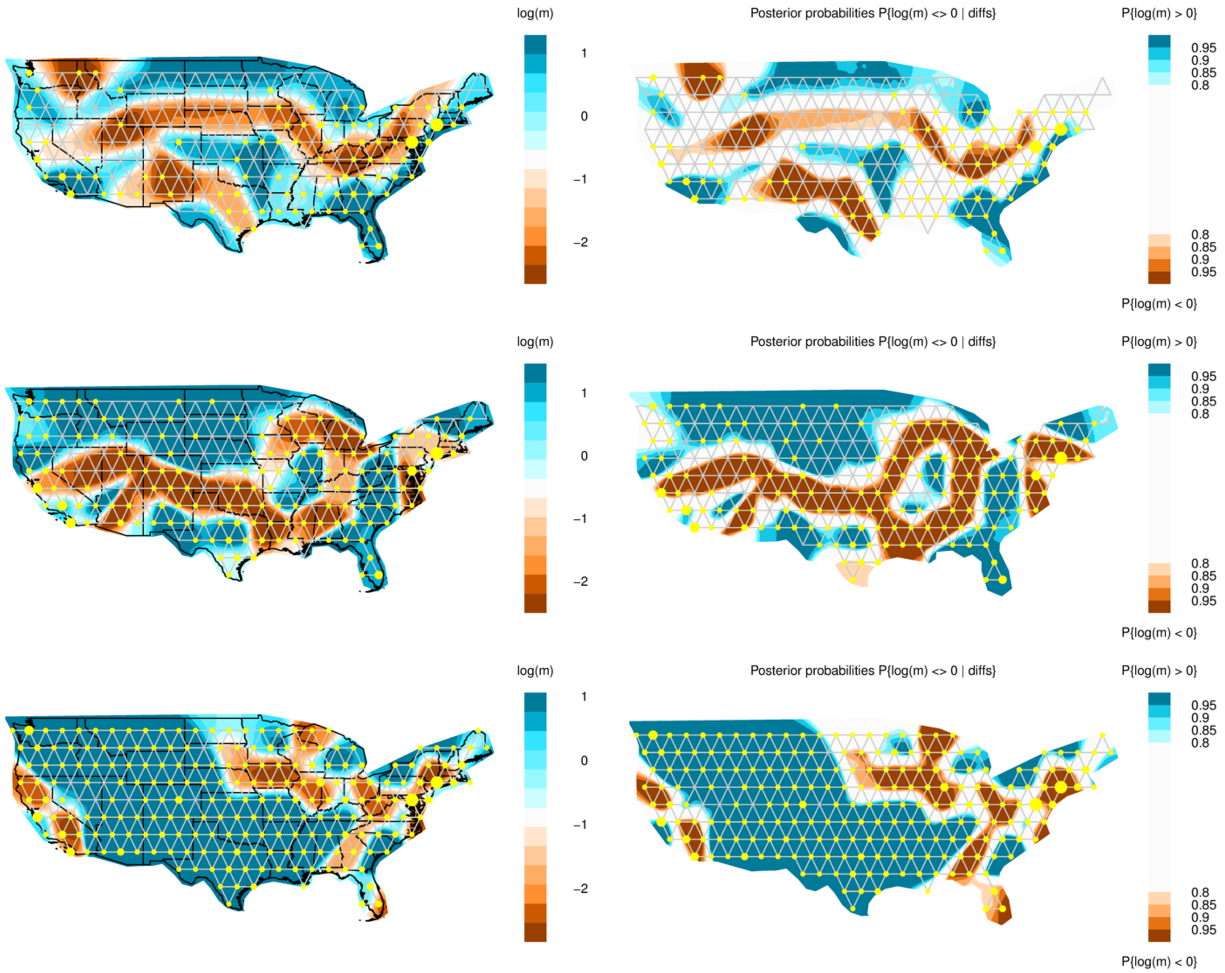


Figure S8. Estimated Effective Migration Surfaces with 250 Demes and Posterior Probabilities.

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.

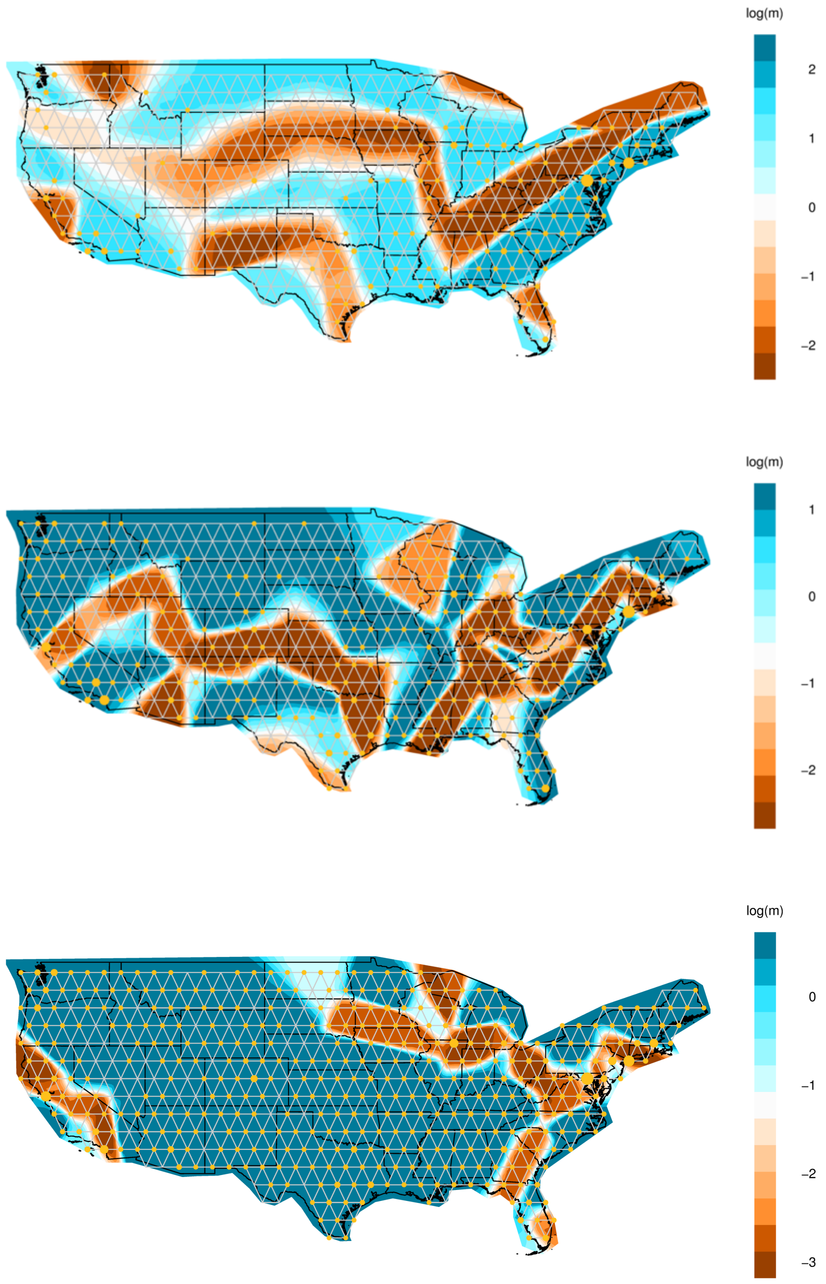


Figure S9. Estimated Effective Migration Surfaces with 500 Demes.

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.

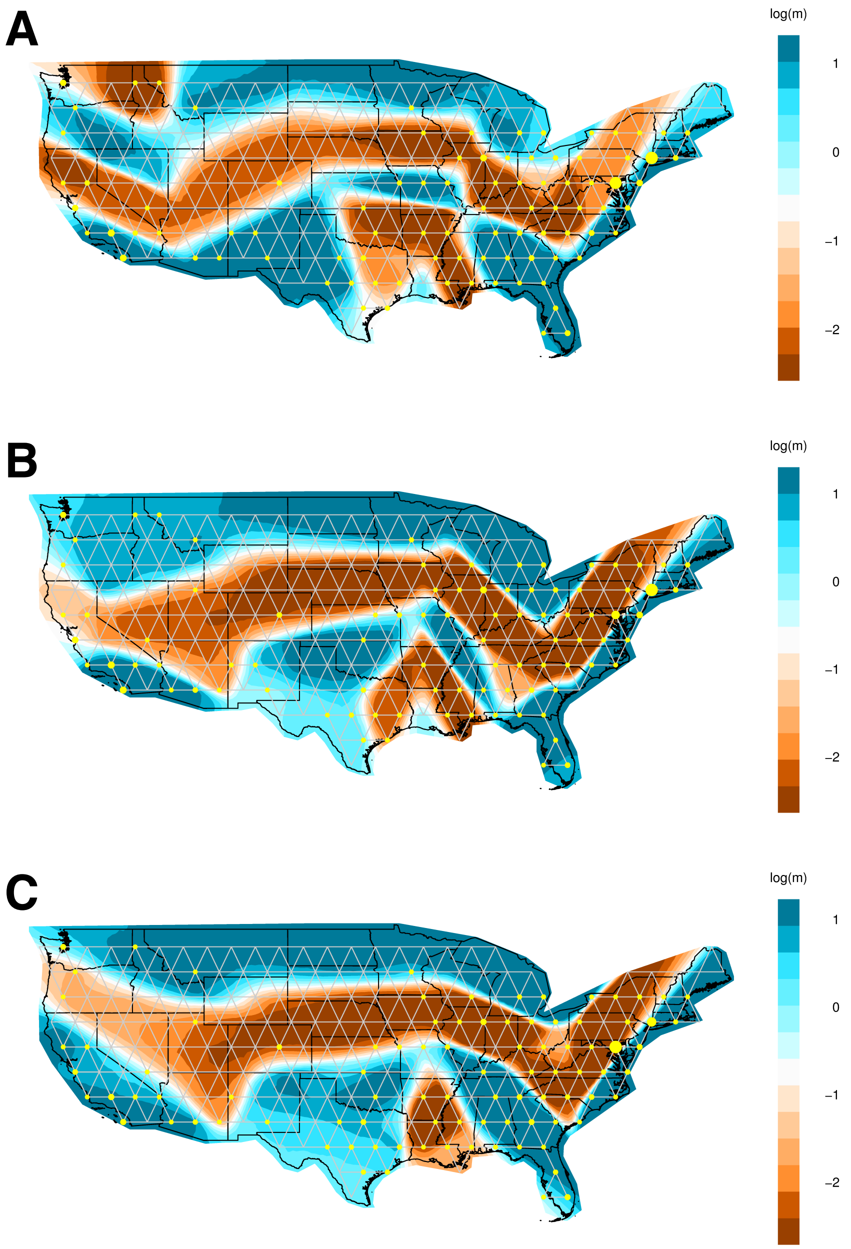


Figure S10. Comparison of Inferred Migration Surfaces with Different Sampling Schemes.
(A) Random subsampling of classified African American individuals to 80% of the original size.
(B) Even sampling across the four major US Census Regions. African American individuals were subsampled to 80% of original 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.

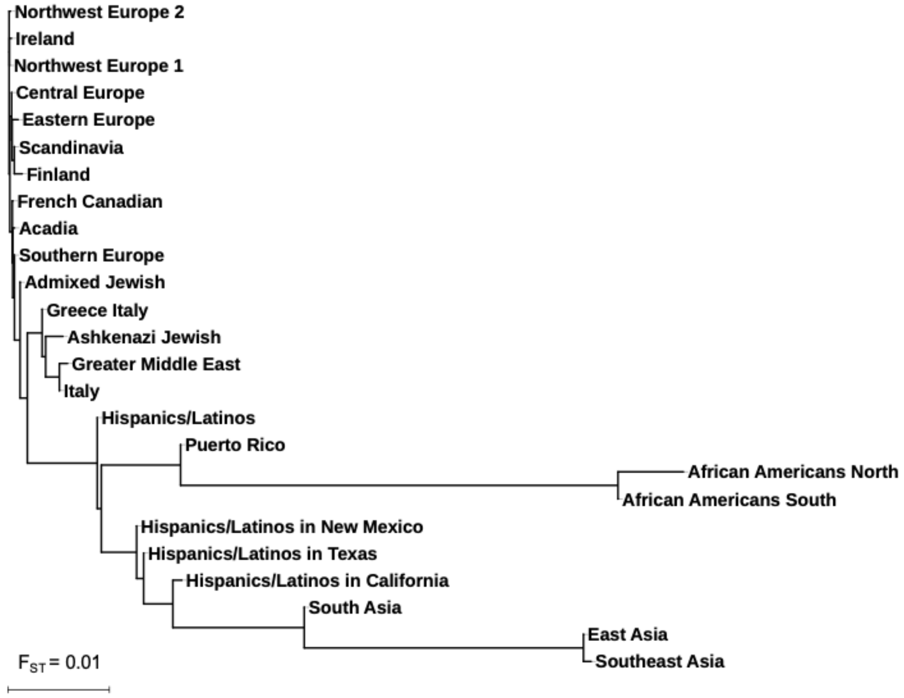


Figure S11. Genetic Differentiation of Haplotype Clusters

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.

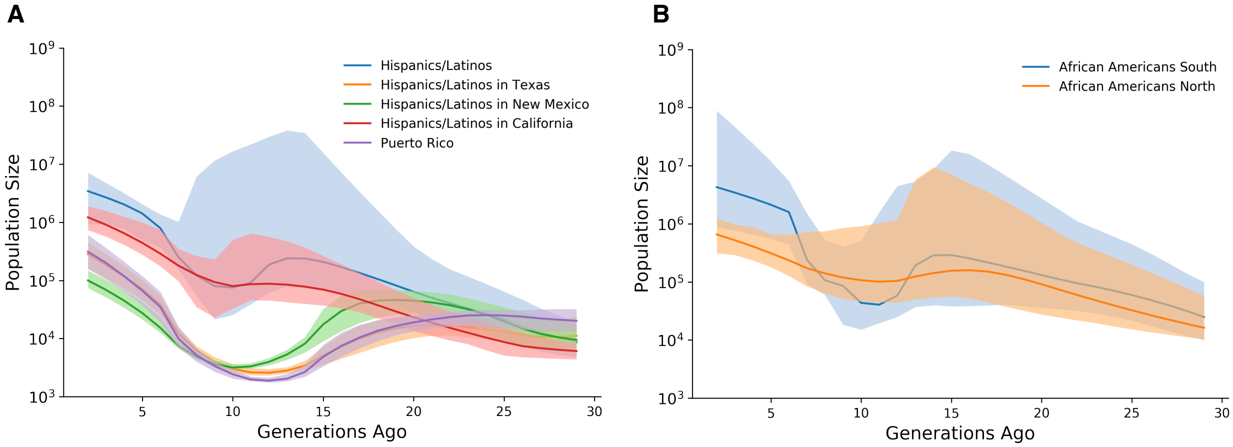


Figure S12. Effective Population Size over Antecedent Generations

(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.

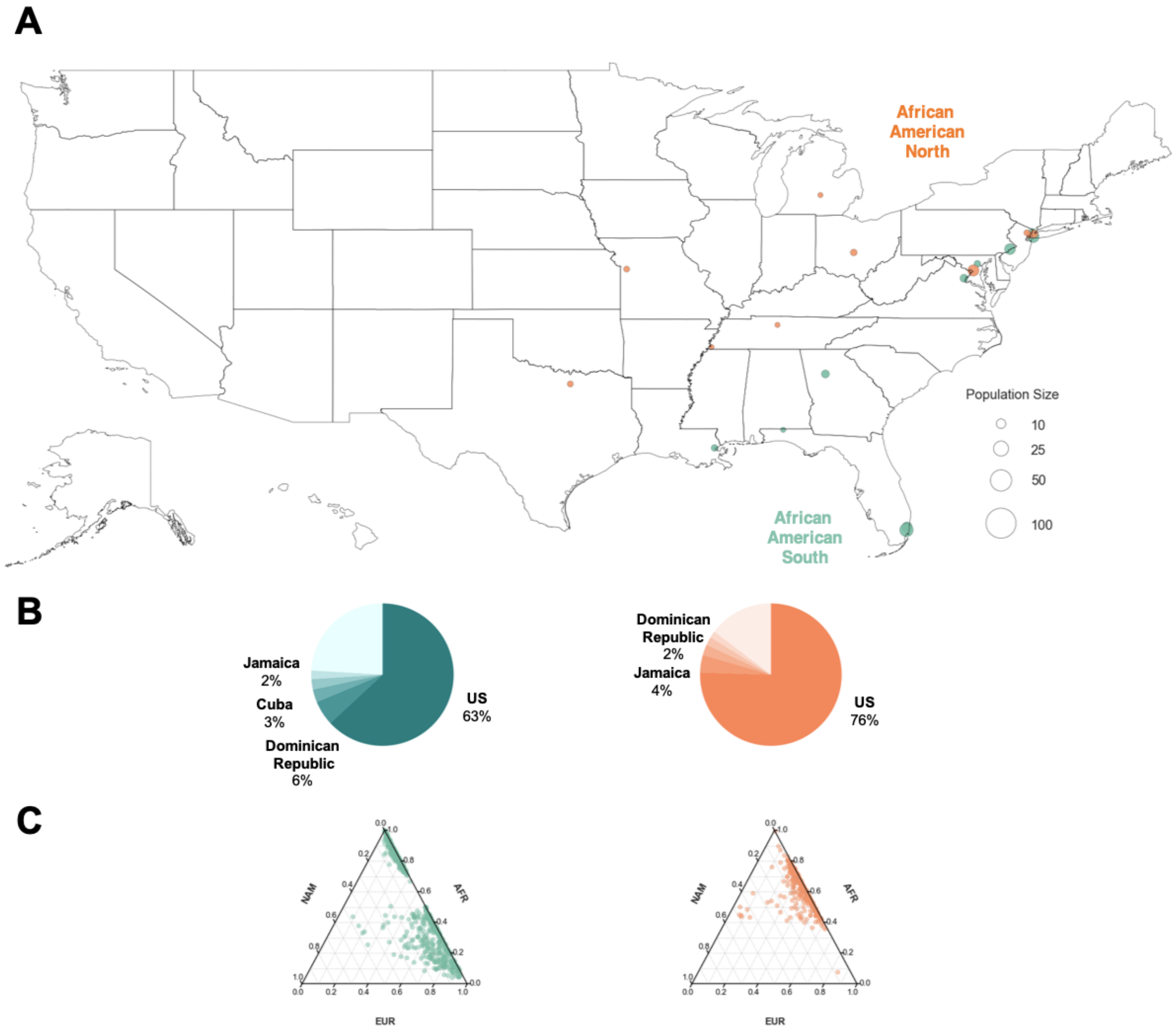


Figure S13. Distribution of African American Haplotype Clusters

(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.

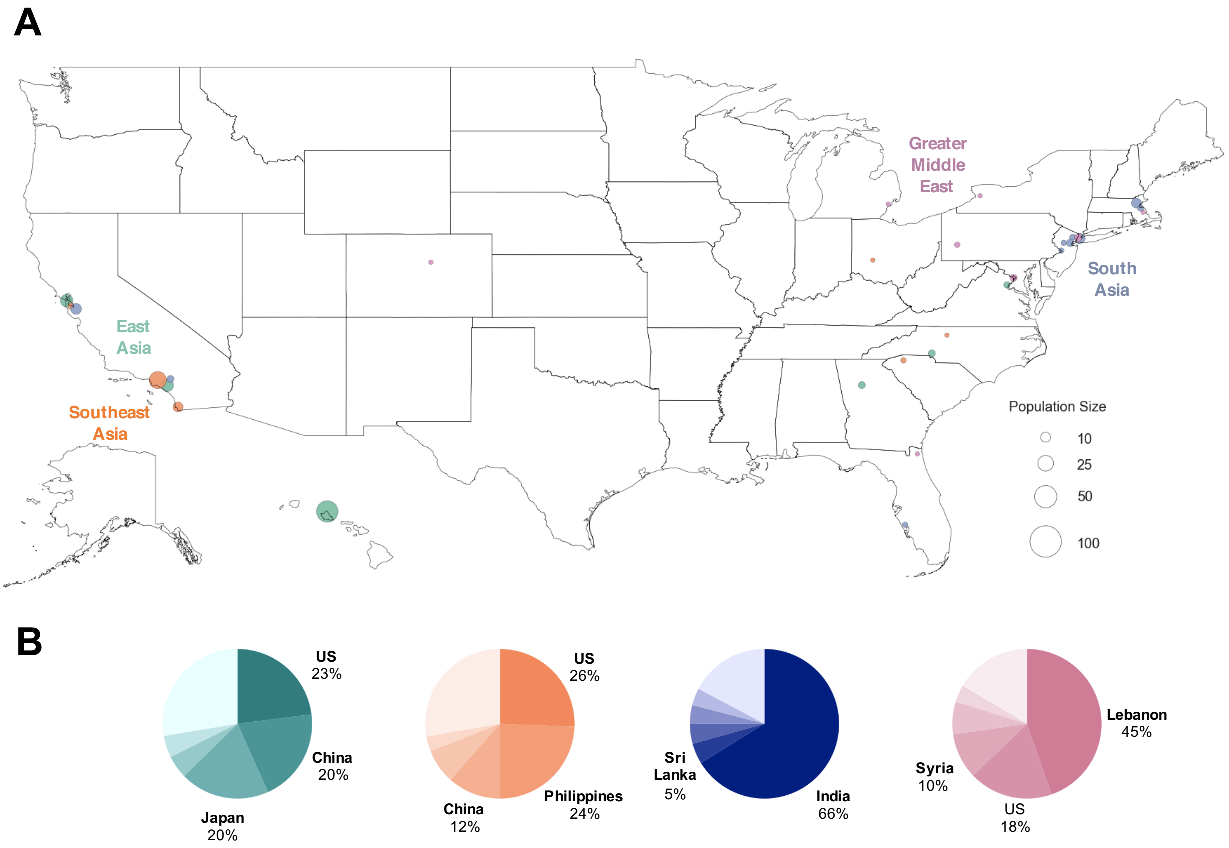
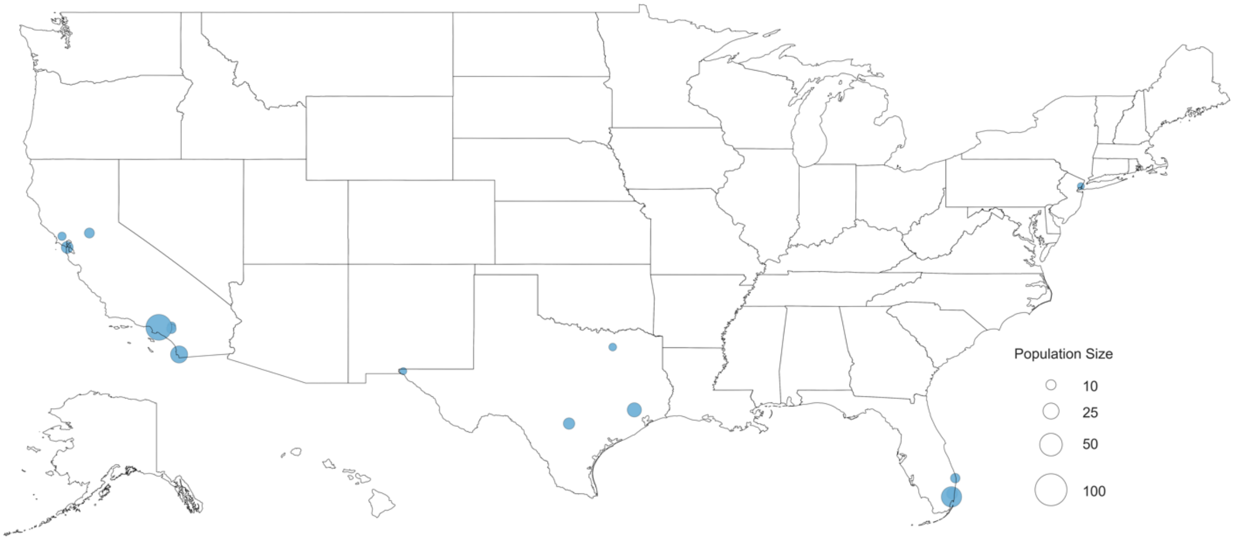


Figure S14. Distribution of Haplotype Clusters with Asian Ancestries

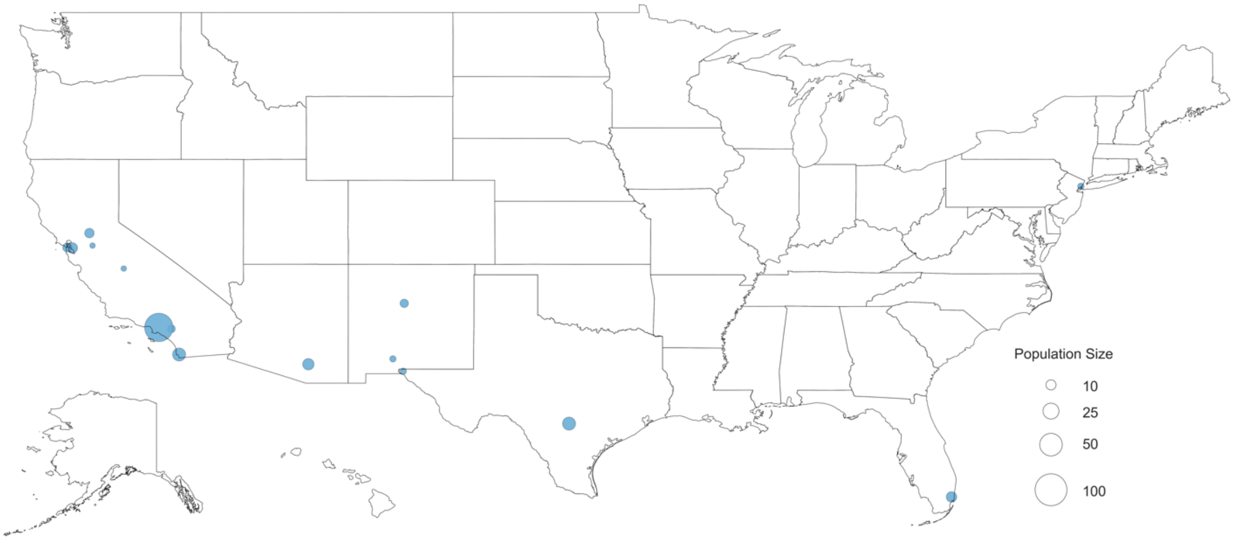
(A) Map of haplotype clusters corresponding to regional Asian ancestries. Each county containing present-day individuals is represented by a dot. The top 10 locations with the highest odds ratio are shown for each cluster. Maps showing the full distribution for each cluster can be found in **Figure S19**.

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

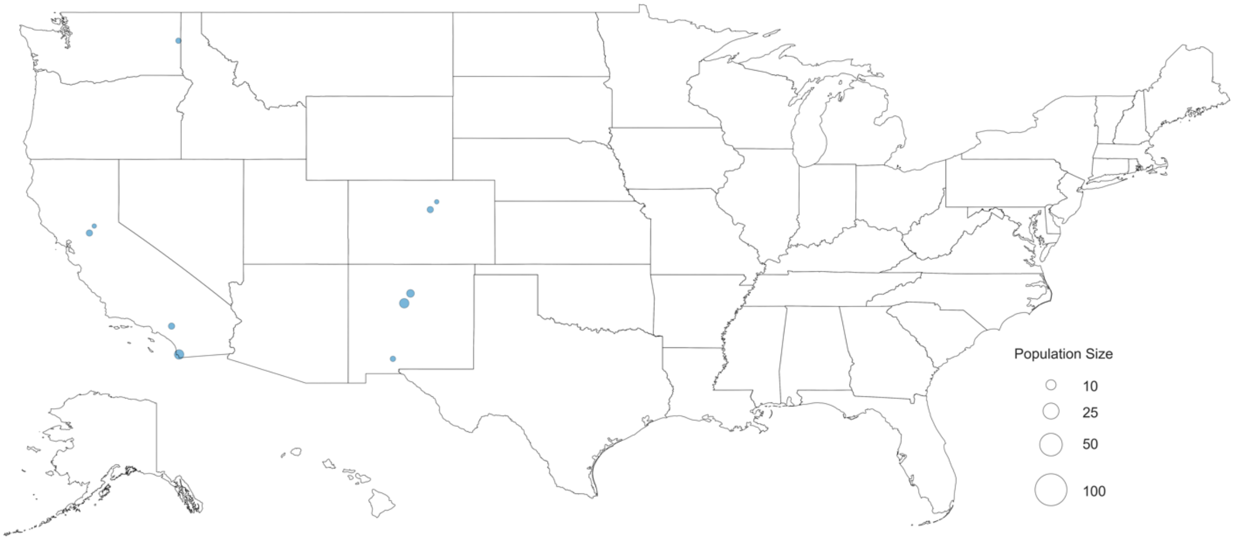
Hispanics/Latinos



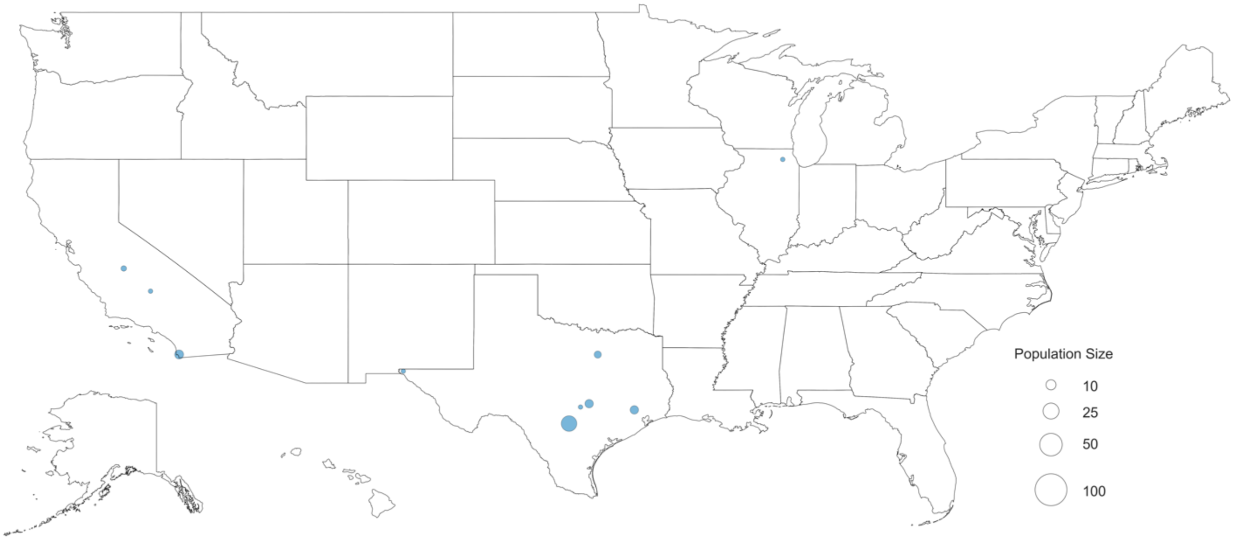
Hispanics/Latinos in California



Hispanics/Latinos in New Mexico



Hispanics/Latinos in Texas



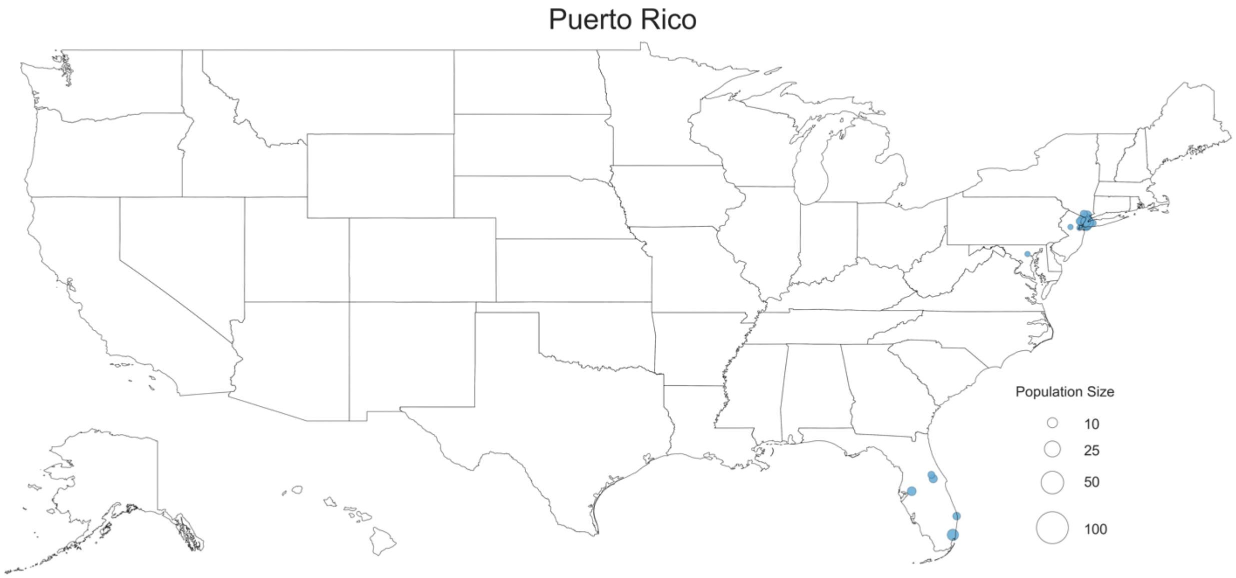
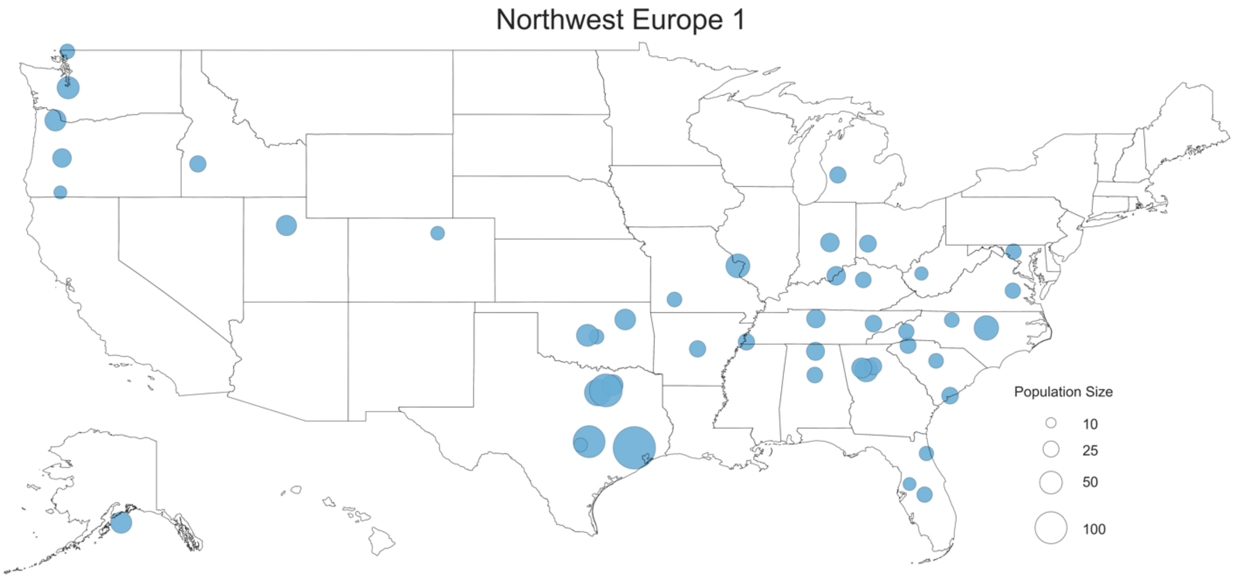
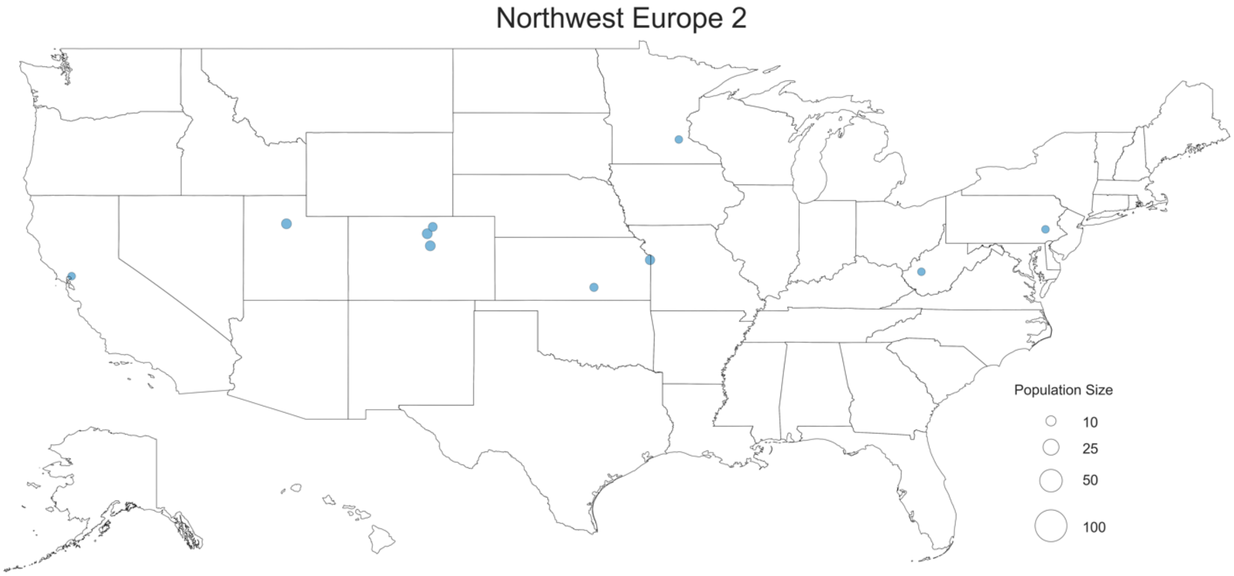


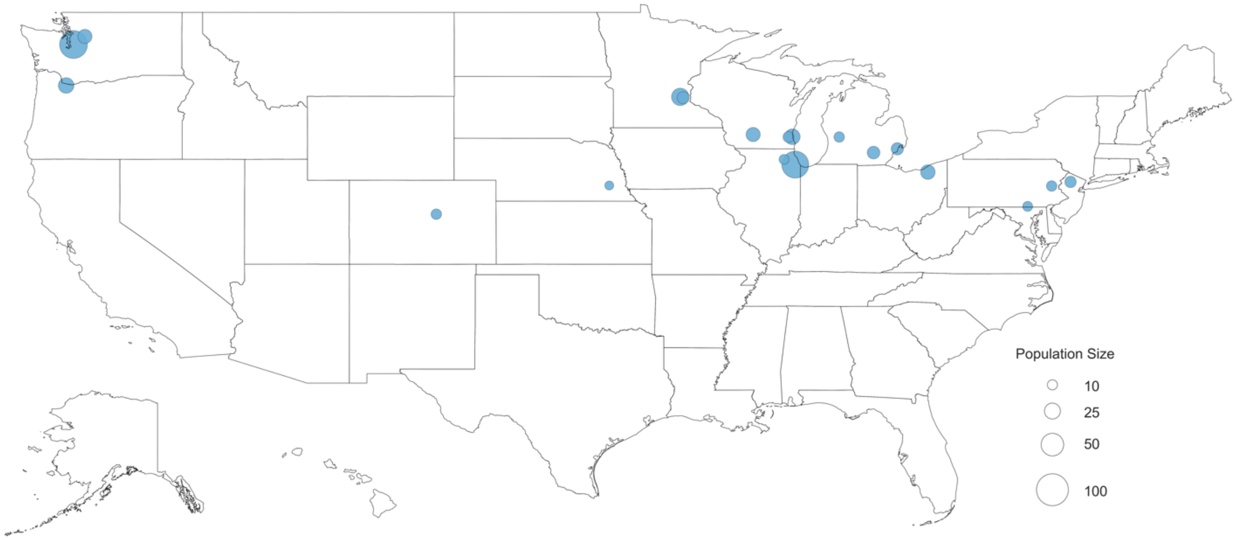
Figure S15. Geographical Distribution of Hispanic/Latino Haplotype Clusters

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.

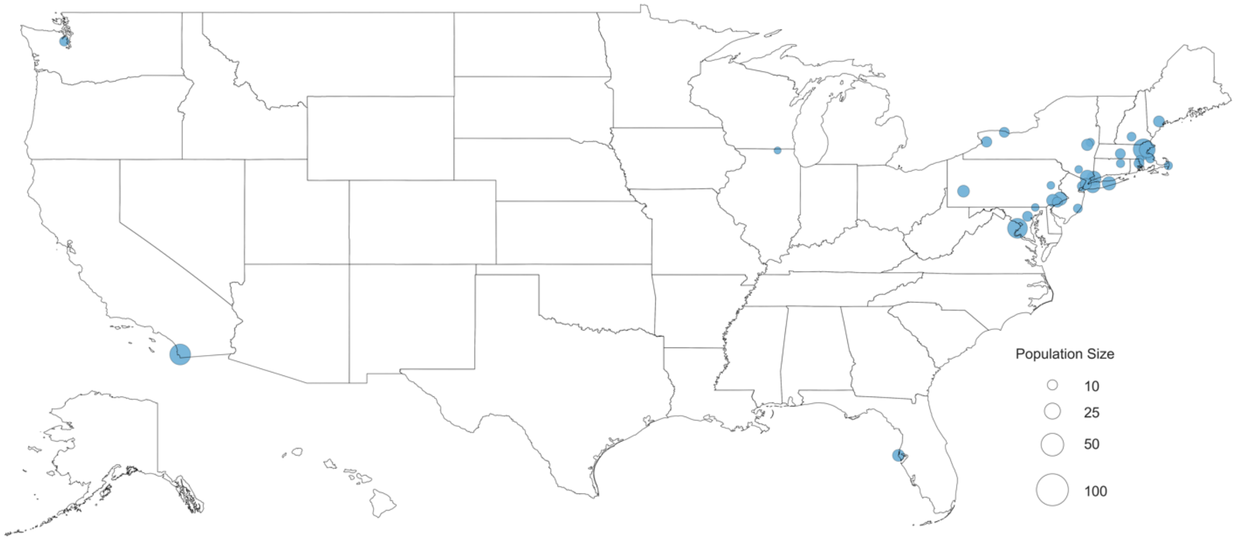

Northwest Europe 1
Population Size
10
25
50
100


Northwest Europe 2
Population Size
10
25
50
100

Central Europe



Ireland



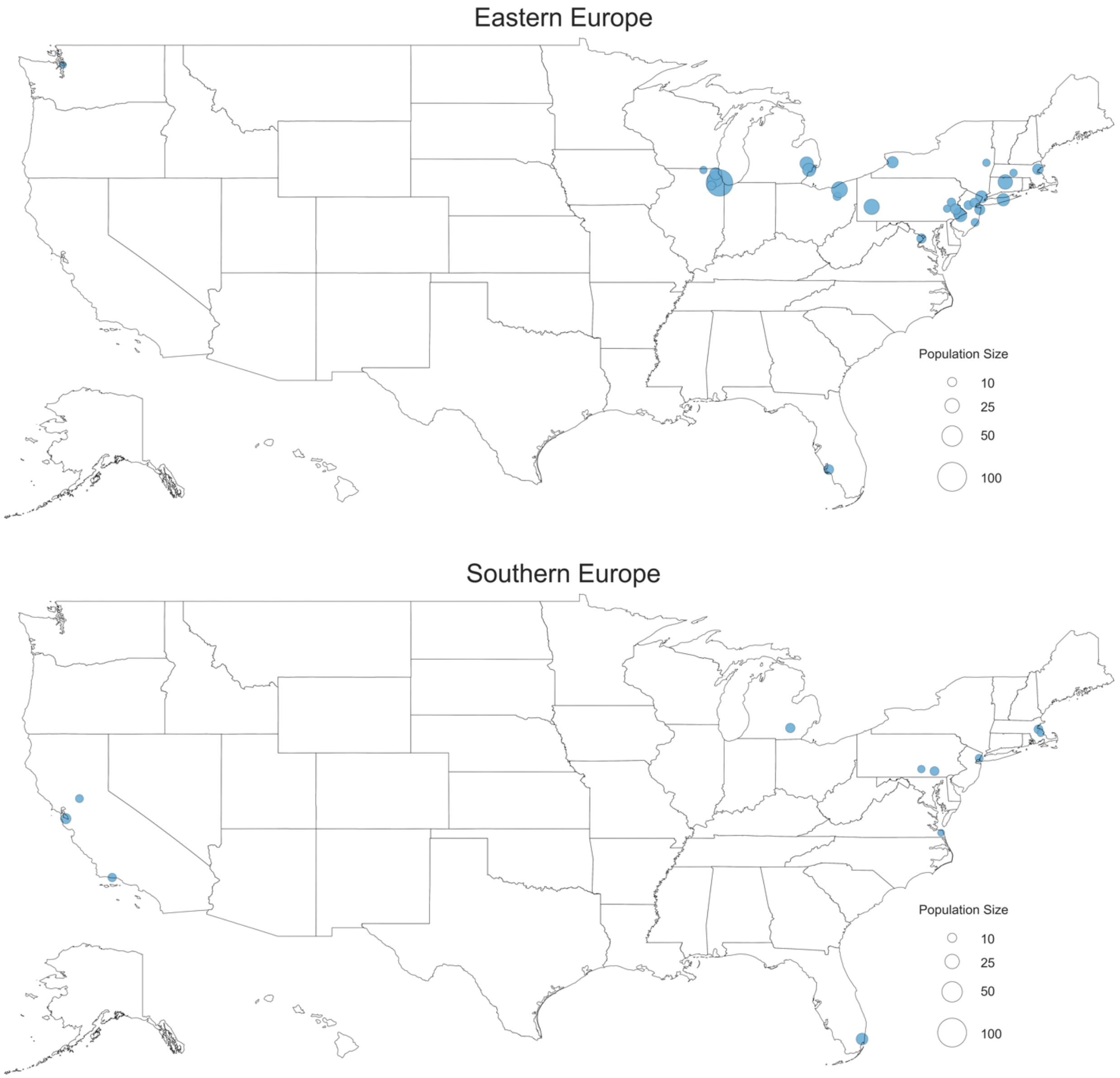
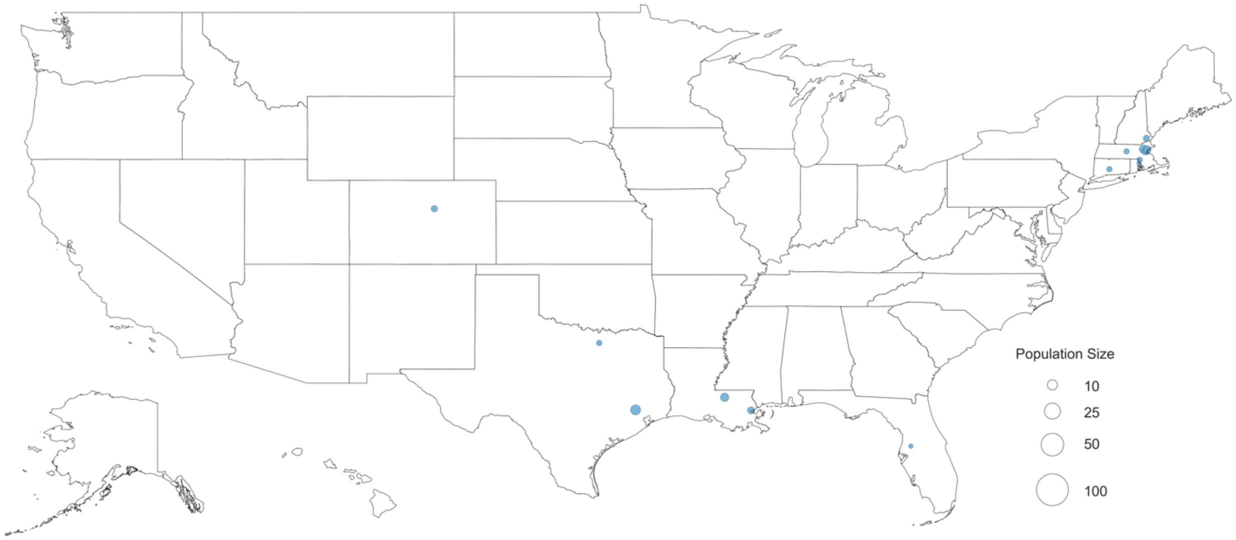
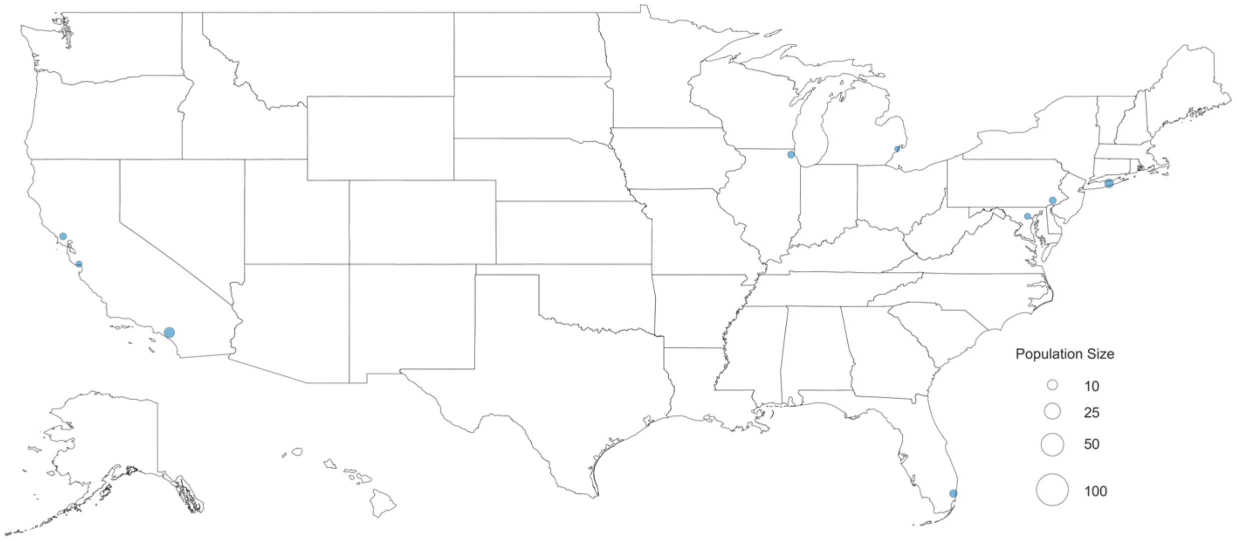


Figure S16. Geographical Distribution of European American Haplotype Clusters Present-day location of individuals in each cluster. Each county is represented by a dot and only the counties with a significant odds ratio ($p < 0.05$) are shown for each cluster. These European haplotype clusters reflect 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.

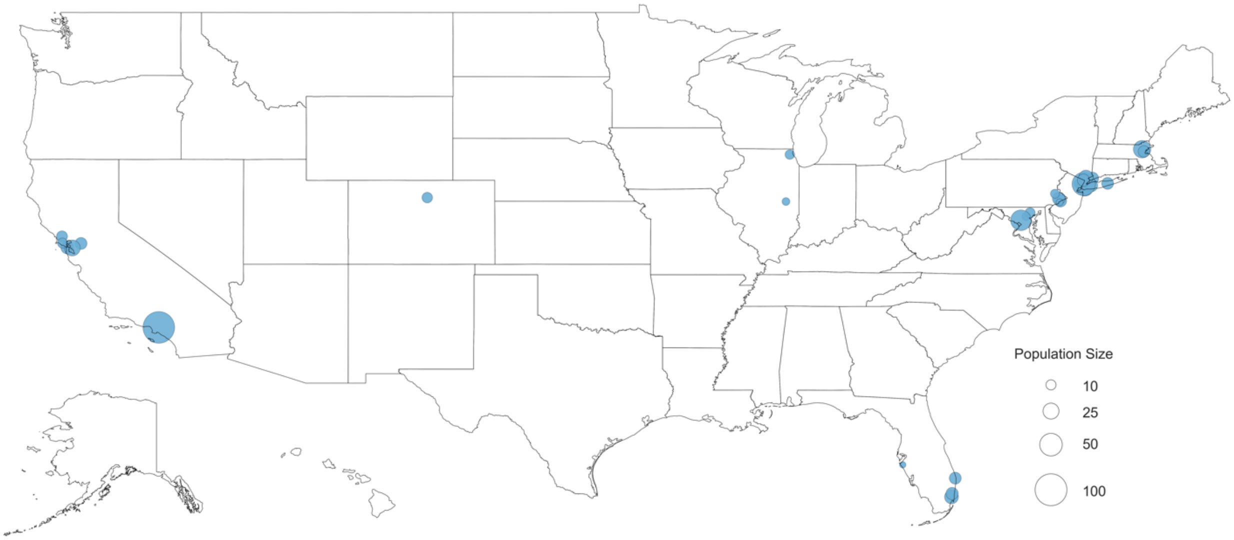


Acadia

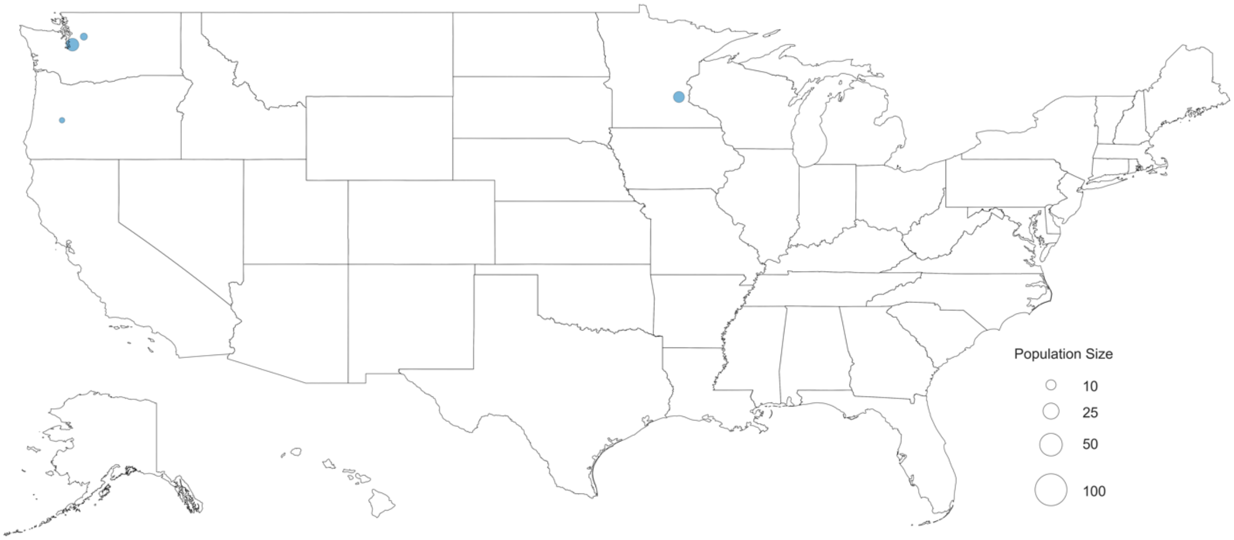


Admixed Jewish

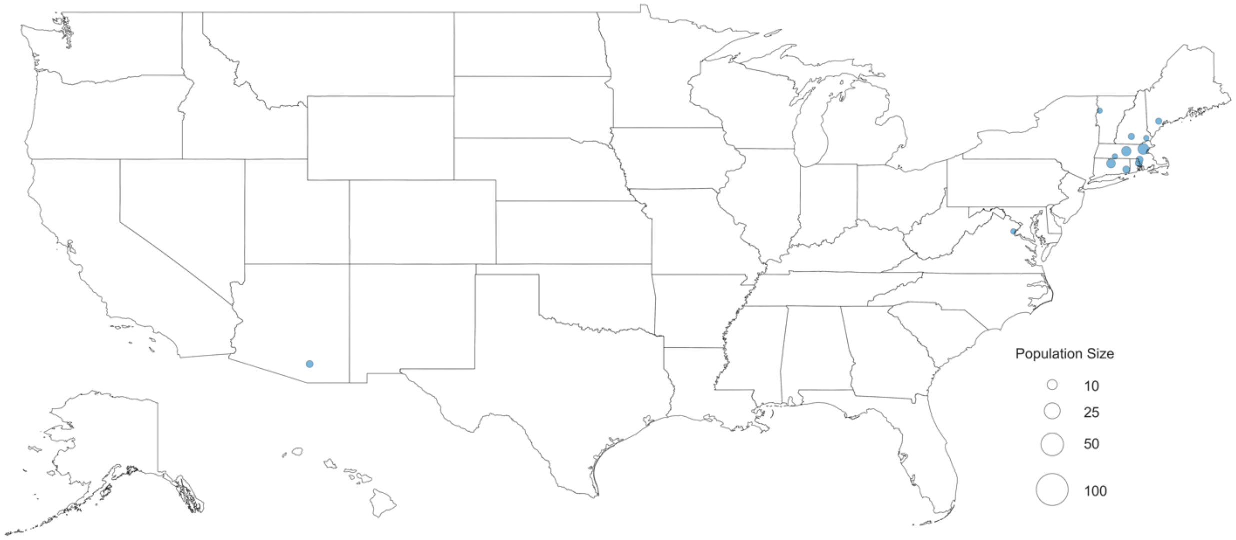
Ashkenazi Jewish



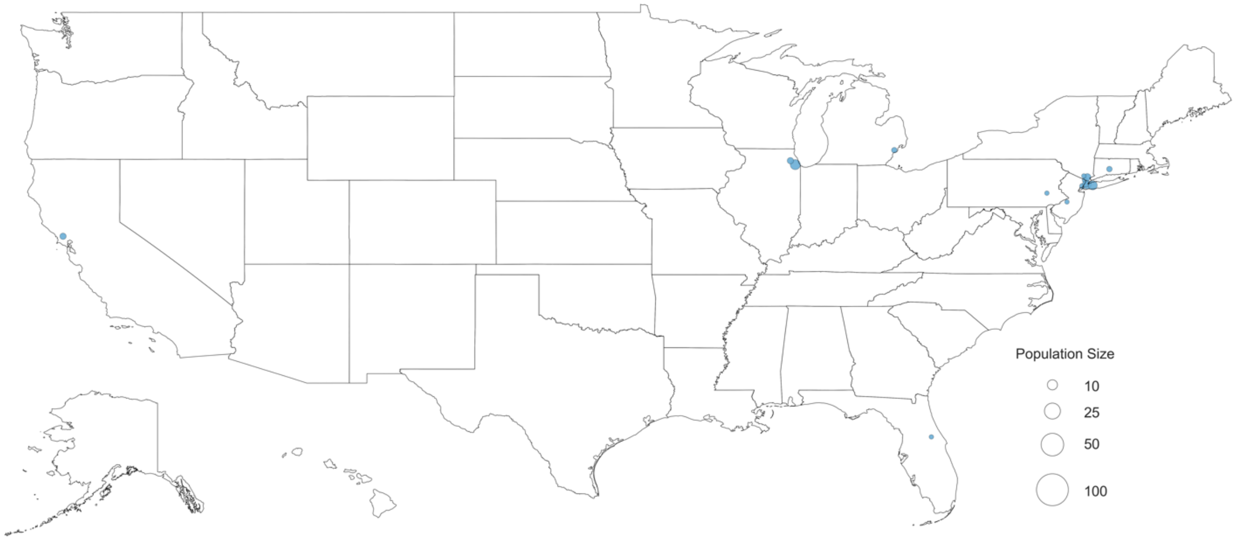
Finland



French Canadian



Greece-Italy



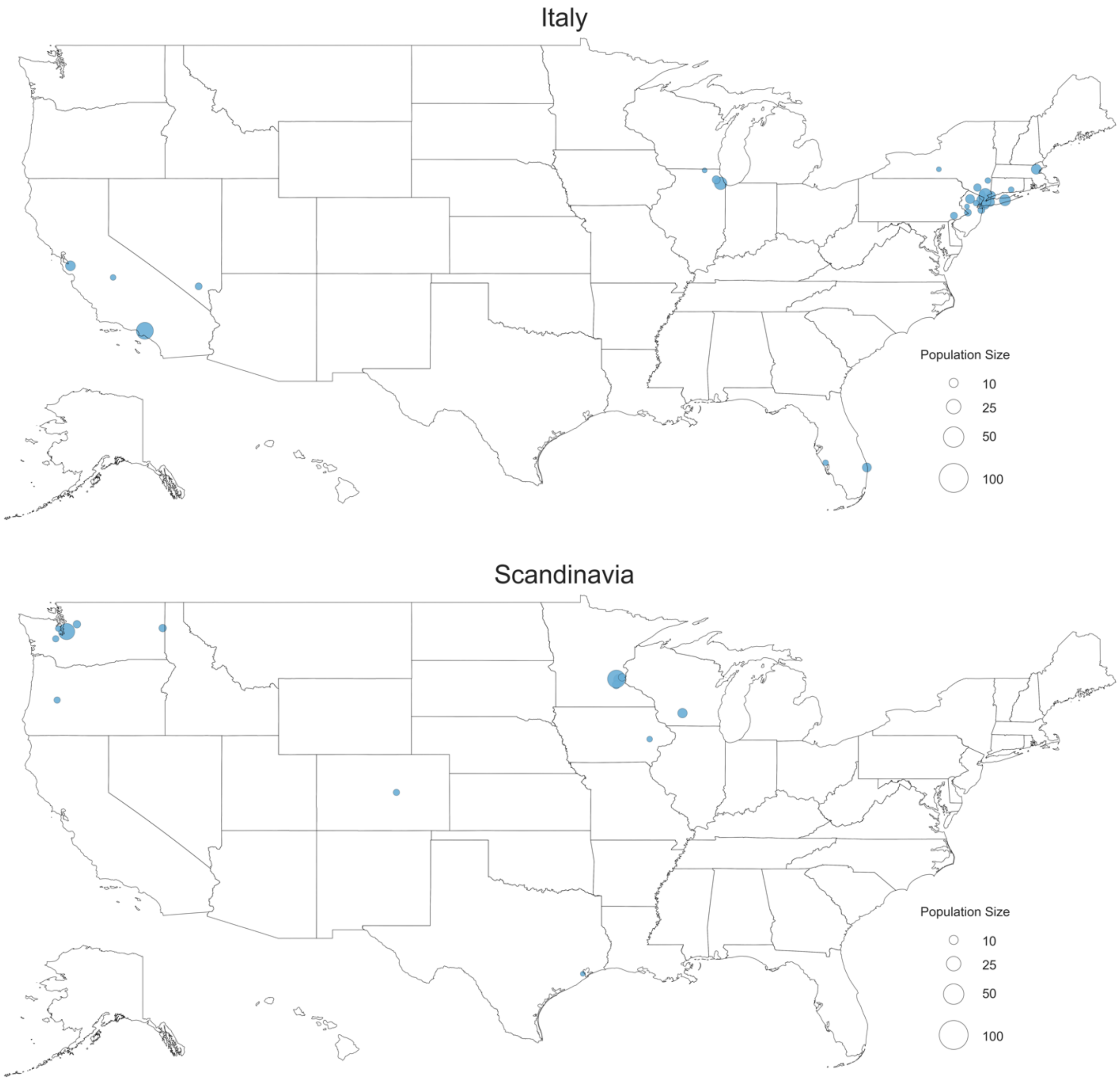


Figure S17. Geographical Distribution of Genetically-Differentiated European American Haplotype Clusters

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.

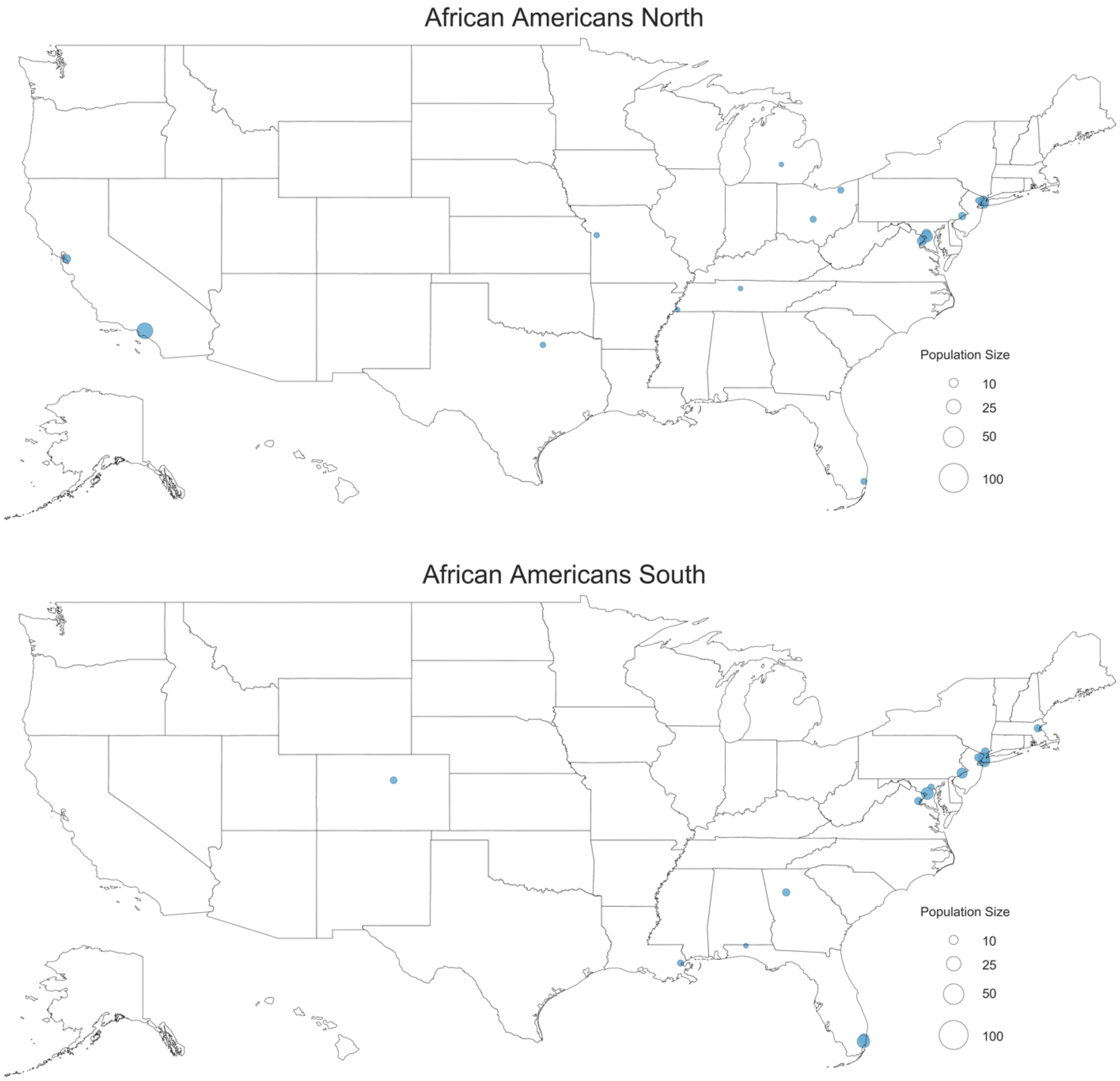
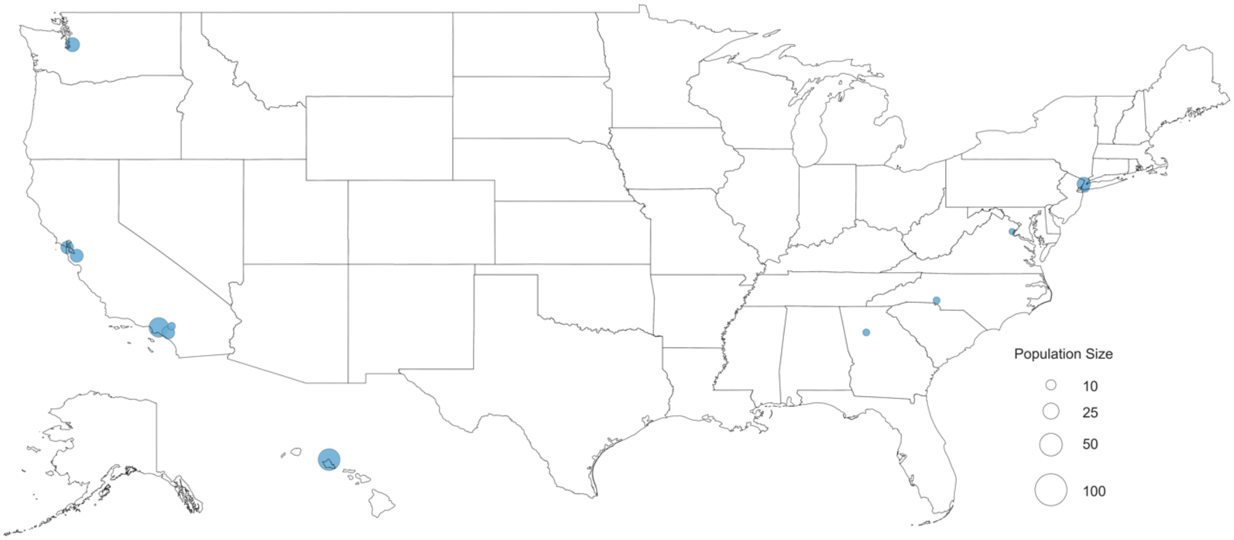
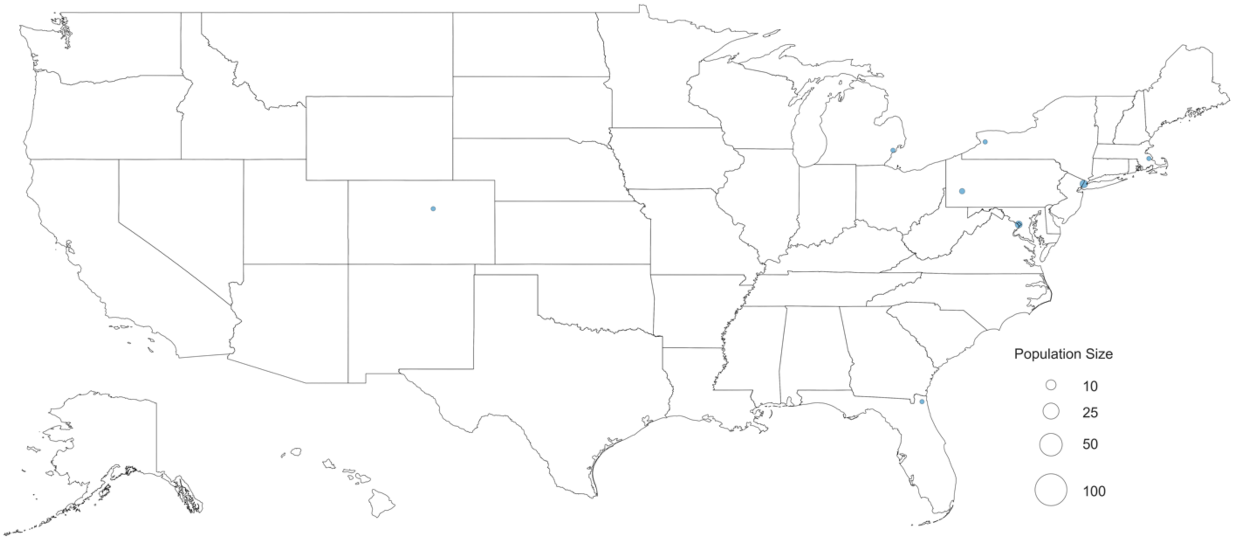


Figure S18. Geographical Distribution of African American Haplotype Clusters
Present-day location of individuals in each African American cluster. Each county is represented by a dot and only the counties with significant odds ratios ($p < 0.05$) are shown for each cluster.

East Asia



Greater Middle East



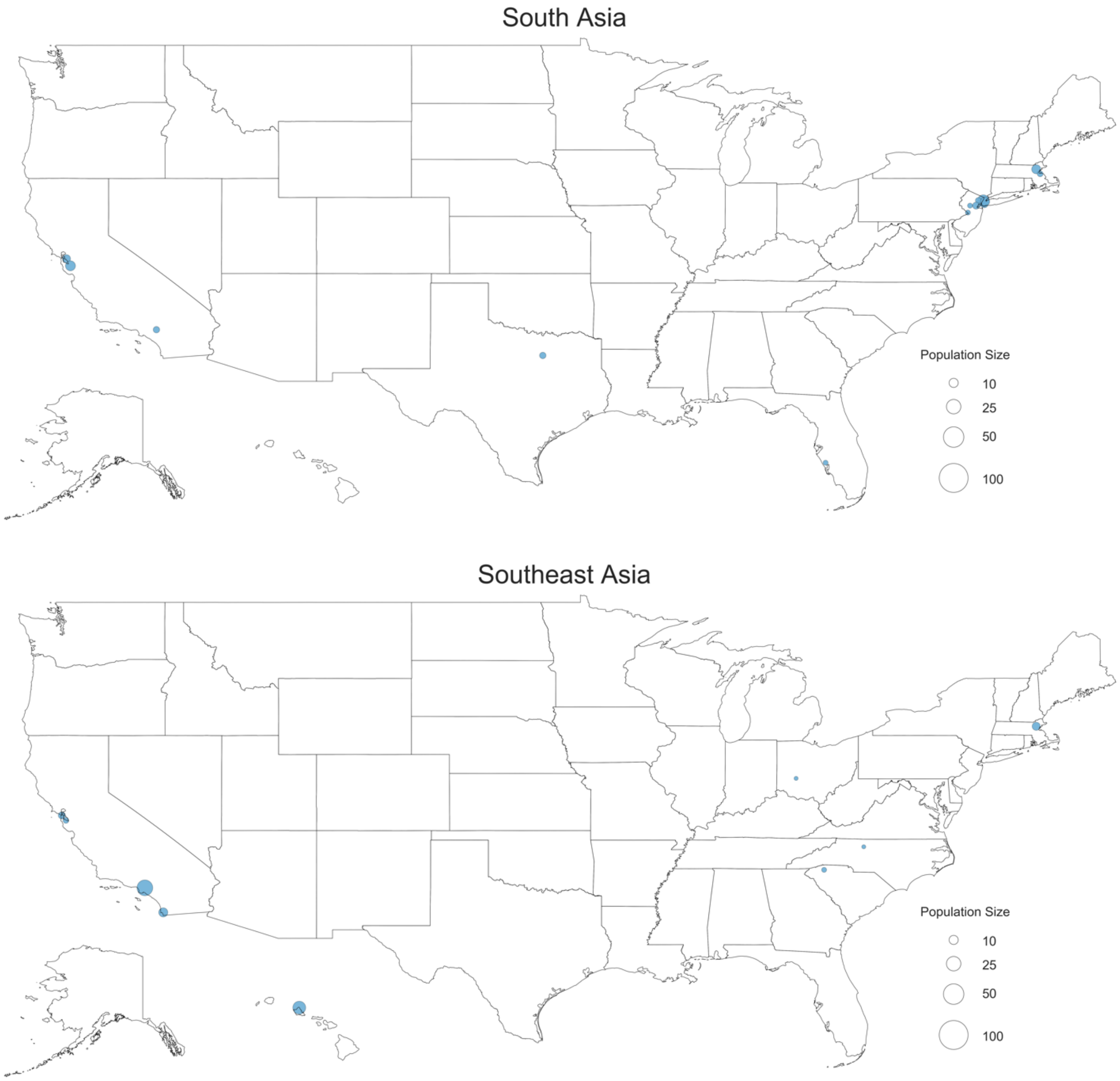


Figure S19. Geographical Distribution of Asian Haplotype Clusters

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.