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]*, Alicia R. Martin[7,8,9]*†

[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 MIT and Harvard, Cambridge, MA 02142, USA

[9] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

*These authors jointly supervised this work

†Corresponding author: armartin@broadinstitute.org

Abstract

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 over 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. 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.

Keywords: population genetics, human history, human genomics, USA

Main Text: 3,765 words (excluding Methods, References, and Figures), 6 figures, 1 table

Introduction

The United States population is a diverse collection of global ancestries shaped by migration from distant continents and admixture of 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 have led to genetic differentiation.[1–6]

Previous genetics studies of the US population have sought to disentangle the relationship between the genetic ancestry and population history of African Americans, European Americans, and Hispanics/Latinos. 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] European American ancestry is characterized by both mixing between different European populations as well as admixture with non-European populations.[6,8,9] Isolation and expansions in certain European population have also resulted in founder effects.[10,11] The mixing of European settlers with Native Americans have contributed to large variations in the admixture proportions of different Hispanic/Latino populations.[1,4,5] Among Hispanics/Latinos, Mexicans and Central Americans carry more Native American ancestry; Puerto Ricans and Dominicans have higher African ancestry; and Cubans have strong European ancestry.[1,4] 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 history.[12,13] 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.

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. Currently, these roles are best understood in European populations due to Eurocentric biases in studies.[14,15] Consequently, translational interpretability gaps are evident in non-European populations: more variants of unknown significance are identified via genetic testing;[16] polygenic risk scores for complex disease risks are much less accurate;[15,17] and false positive genetic misdiagnoses are more common.[18] Thus, studies of diverse, heterogeneous populations

offer substantial value to both our understanding of population history and biomedical outcomes.[19]

In this study, we comprehensively explore the population structure and migration history of over 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.[20] Here, we identify patterns of genetic ancestry and haplotype sharing among the project participants. We combine these patterns with ancestral birth origin records and geographic information to uncover recent demographic and migration trends. Taken together, we provide insights into the ancestral origins and complex population histories in the US.

Results

Genetic ancestry and diversity across the United States

To assess the diversity of ancestries among individuals in the Genographic Project, we first performed PCA and ADMIXTURE analysis (**Figure 1A-C**; **Figure S1-S2**).[21,22] Since self-reported ancestry does not always reflect genetic ancestry, we objectively assigned continental ancestry to each Genographic sample using the 1000 Genomes Project data as reference populations (**Methods and Materials**). We first trained a Random Forest classifier on the first 10 principal components (PCs) of the 1000 Genome Project samples with super population classifications as ancestry labels (EUR = European, AMR = Admixed American, AFR = African, EAS = East Asian, SAS = South Asian). We then used the trained model to assigned continent 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, although many have ancestry patterns similar to other European individuals (**Figure 1C; Table S1**). The inability to classify these individuals may be due to the complex and variable admixture profiles of certain populations such as Hispanics/Latinos.

Regional differences in genetic ancestry proportions correspond to historical demographic trends. We evaluated the admixture proportions of classified individuals across the four designated US Census regions: South, Northeast, Midwest, and West (**Figure 1C; Figure S2**). 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 S1**). Individuals classified as having African ancestry are most common in the South (3.2%), followed by the Northeast (3.0%). Individuals of Native American ancestry 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). East Asians mostly reside in the West (2.1%), while South Asians are most abundant in the Northeast (1.0%).

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.[23,24] By leveraging multiple PCs at once, UMAP can disentangle subcontinental structure (**Figure 1D-E; Figure S3-S4**). Similar to previous analysis,[24] populations in the 1000 Genomes Project form distinct clusters corresponding to ancestry and geography. The Genographic 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 on 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 populations of ACB and ASW were previously found to be similar to the Nigerian YRI and ESN populations.[2,17]

Population differentiation and migration rate inference across the United States

To better understand the relationship between genetics and geography, we investigated migration rates for genetically inferred Europeans, African Americans, and Hispanic/Latinos across the United States. We excluded East Asians and South Asians due to small sample size

and limited our analysis to the contiguous 48 states. We inferred effective migration rates with the estimating effective migration surfaces (EEMS) method,[25] which statistically characterizes genetic differentiation via resistance distance across non-homogenous landscapes. By overlaying a dense regular grid of demes and measuring genetic dissimilarities between neighboring demes, EEMS quantifies and visualizes areas with high relative rates of effective migration (colored in blue) and areas with low relative rates of effective migration (also called migration barriers and colored in dark orange).

The inferred migration rates for African Americans reveal genetic signatures of historical demographic events (**Figure 2A; Figure S5**). Along the Atlantic coast from the Florida Panhandle to southern Maine, we find high effective migration rates, 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 west across the rest of the country. This migration barrier, along with the migration barrier spanning Texas and New Mexico, reveals a pattern of isolation-by-distance 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.[7,26]

A highly complex pattern of migration exists amongst Hispanics/Latinos with varying migration rates across the country, capturing regional patterns of genetic similarity. Hispanics/Latinos in the southwestern states including two regions bordering Mexico--one in California and another extending from New Mexico to Texas--exhibit high effective migration rates and are separated by a migration barrier in Arizona (**Figure 2B; Figure S5**). These two distinct regions likely reflect known differences in northward migration from east versus west Mexico.[8,27] Along the Atlantic coast from Florida to New York, effective migration has also been fluid. However, barriers to migration are observed west of the Atlantic coast to the Mississippi River, likely resulting from varying admixture proportions.

The patterns of migration for Europeans capture subcontinental structure. Elevated migration rates are observed across most of the country, except for many states in the Midwest and along the Atlantic coast. We find low effective migration rates surrounding Minnesota and North Dakota, potentially due to the genetic dissimilarity of Finnish and Scandinavian ancestry abundant in the region (**Figure 2C; Figure S5**).[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 Chicago, Philadelphia, and Miami, are also barriers to migration, perhaps due to higher admixture proportions within cities. The migration barrier encompassing metropolitan New York City may be explained in part by the presence of divergent European populations, such as Ashkenazi Jews (**Figure 2C**).

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. In this network, 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 at each level of hierarchy.[8,28] 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 ($F_{ST} < 0.0001$) at the lowest level of hierarchy, 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 3**). 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 individuals, as unclassified individuals comprised of 8-11% of each cluster.

Genetic and geographic diversity is greatest amongst Hispanic/Latino haplotype clusters. We identified a total of five Hispanic-related clusters. The largest of these cluster (n=810) is strongly associated with south Florida (OR = 10.4; p = 2.5e-25; **Figure 4**, **Table S4**) but is also found in California, and Texas (OR ≥ 2; $p < 0.05$). No single ancestral birthplace characterizes this cluster, as the US, Mexico, and Cuba each make up more than 10% of the birth origin labels. Proportions of European ancestry tracts inferred with RFMix[29] are higher in this cluster (mean = 72.7%, sd=20.4%) than in the other Hispanic/Latino clusters (mean = 48.0% - 67.4%). 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, likely due to high proportions of African ancestry (mean = 11.2%, sd = 9.0%) among Puerto Ricans.

Three distinct clusters of Hispanics were found in the Southwest (**Figure 4**): 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 the genetic differentiation of these three clusters are subtle (F_{ST}=0.001-0.003), ancestral birth origin patterns and local ancestry proportions for these clusters reveal meaningful dissimilarities. Whereas the majority of Hispanics in New Mexico report US ancestral birth origins through grandparents, the recent ancestors of Hispanics 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 2B**). The reduced migration rate along the Texas-Mexico border may be caused by more recent immigrants. Unlike the Hispanic clusters associated with New Mexico and Texas, the Hispanics in California cluster contain greater proportions of ancestors from Central and South American (e.g., Colombia and El Salvador). Proportions of Native American ancestry is also highest in this cluster (**Figure 4**). Taken together, these two differences further explain the presence of the migration barrier in Arizona between the Hispanics in the California and the Hispanics 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 comprising of Southern and Eastern Europeans occurring from the 1880s to 1910s.[30] 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 (**Figure 5**). 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. 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 (**Figure 6**). 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, Ashkenazi Jews, and Admixed Jews are mostly located in the metropolitan area of New York City. Of the European clusters, median cumulative IBD sharing and cROH lengths are highest amongst Ashkenazi Jews (31.8cM and 11.3 Mb, respectively; **Table 1**). 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, was more consistently reported on both sides of the family in the larger Jewish cluster ("Ashkenazi Jewish"), suggesting that individuals are more admixed in the smaller cluster ("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 amongst the northern and western states ("African Americans North") while the other is distributed amongst the states southeast of the Appalachian Mountains ("African Americans South") (**Figure S7**). The proportion of US birth origin is higher in the northern cluster than the southern cluster, further evidence of isolation by distance amongst African Americans in the north.[7] These two clusters share similar cROH lengths but differ in admixture proportions and median IBD sharing, pointing to a cluster with consistent African American ancestors and a cluster with more admixed ancestors. Median IBD sharing is higher amongst African Americans in the south (median IBD = 19.6 cM, median cROH = 3.3 Mb) than in the north (median = 15.9 cM; **Table 1**) 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 S8**), which grew rapidly in the mid-20th Century after the elimination of national origin quotas.[31] The recency of immigration among these clusters is indicated by the less than 30% of ancestral birth origins coming from 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 Middle East individuals reflects many of the known 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. cROH lengths are particularly long for South Asians (median cROH = 10.3 cM), Southeast Asians (median cROH = 7.8 cM), and Middle Easterners (median cROH = 8.2 cM), potentially reflecting inbreeding patterns found in their ancestral regions.[32]

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 homogenous groups in isolation.[15] 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, EEMS is limited in identifying unique subpopulations, while haplotype clustering cannot assign partial membership for admixed individuals to multiple clusters. An integrative approach can thus enable greater insights into populations with complex histories.

Consistent with prior studies,[4,9] the recent demographic history of Hispanic/Latino populations is complex. Large variations in admixture proportions within and between subpopulations are reflected by US Census data and can likely be explained by numerous inferred migration barriers. For example, regional differences in the Southwest are highlighted by an inferred migration barrier in Arizona and distinct haplotype clusters surrounding this region. These differences are likely due to higher proportions of Native American ancestry as well as more Central and South American origins in the California Hispanics/Latinos compared to other southwestern Hispanic/Latinos. Interestingly, although the New Mexican cluster is distinct from the Texan cluster, high levels of gene flow are inferred from southern New Mexico to central Texas, suggesting that certain individuals in these two clusters are genetically similar and may share an ancestral origin (i.e. Mexico). In contrast, those in northern New Mexico are more genetically differentiated, as indicated by a migration barrier, but share the same cluster; these are likely *Nuevomexicanos*, descendants of Spanish colonial settlers.

The fine-scale population structure of African Americans also reflects known historical events following the transatlantic slave trade, during which millions of West Africans were forcibly moved to the Americas. Subsequently, the movement of African Americans during the Great Migration has been shown to correlate with current patterns of relatedness across US census regions.[7] Our results show barriers to migration and gene flow at fine-scale, particularly along the Appalachian Mountains. A north-south migration barrier is also present west of the Mississippi River, and is further supported by the north-south locations of two African American clusters that emphasize this divide. The southern African American cluster contains more recent ancestors outside the US, particularly of Caribbean origin, than the northern African American cluster. These genetic signatures illustrate the impact of recent migration patterns on modern population structure.

Our ability to identify population structure for certain ancestries is subject to participation among individuals from those groups. In particular, individuals with Asian ancestries account for over 5% of US population, but they are underrepresented in US population genetics studies, hindering the investigation of their ancestry in prior studies.[8] Our analyses of East Asian, Southeast Asian, South Asian, and Middle Eastern populations therefore provide initial insights into their genetic structure. 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. Specifically, the long cROH in South Asians may reflect endogamy for example 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.[33–35] Given the small size of these clusters, however, further studies of more individuals are needed.

Population history in the US is best characterized among the most populous European descent individuals. Genetic diversity tends to be highest in more densely populated regions, likely due to the presence of multiple subpopulations in the same place. Many of the European subpopulations we identified are similar to those previously found—e.g., French Canadians, Acadians, Scandinavians, and Jews (Supplementary Discussion).[8] The geographic distribution of these subpopulations, particularly those that are more genetically diverged, overlap in the metropolitan areas in the Northeast, Midwest, and California.

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. 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 to the US has been relatively limited in population genetics studies, and the structure of this immigration will be learned from larger future studies. Additional considerations relating to population label precision are the accuracy of self-reported birth records and variable granularity of geopolitical boundaries.

The emergence of biobank-scale genomic data is enabling more complete pedigrees,[36] greater discoveries of fine-scale population structure, and more precise insights into health-related associations. An estimated 26 million people have taken a direct-to-consumer ancestry test,[37] indicating widespread interest in ancestry and heritable factors. As participation in genetic studies increase, especially in the US with the All of Us Research Program, so does the need for inferring more granular demographic histories in study cohorts. Understanding such structure is important to account for stratification, prevent the overgeneralization of results, and avoid exacerbating existing biases.[14,15] This study demonstrates the potential of coupling genetic data with geographic and birth origin data to reconstruct such demographic histories, particularly in a large and heterogeneous population.

Materials 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 data in this study came from participants that consented to have their results be used in scientific research. All data was deidentified.

Participants provided genotype data, geographic location (postal code), ancestral birth origin, and self-declared ethnicity. We limited our study to those individuals who provided valid geographic location. Ancestral birth origin and self-declared ethnicity data were collected up to the grandparents of the participants with ~60% of individuals provided complete pedigrees.

Genotyping and Quality Control

Participants of the Genographic project were sequenced with the GenoChip array,[20] an Illumina iSelect HD custom genotyping bead array with approximately 150,000 Ancestry Informative Markers. Quality control and phasing of data is described in Supplemental Materials and Methods. After QC, 32,589 individuals and 108,003 sites remained.

Principal Component Analysis

We performed principal component analysis on the quality-controlled samples using FlashPCA version 2.0.[22] We included the genotypes of all 2,504 individuals from the 1000 Genomes Project as reference samples. We computed PCs across 108,003 shared sites for 1000 Genome Project individuals and then projected the Genographic individuals on the same principal component space.

Continental Ancestry Assignment

We assigned continental ancestry to each Genographic sample by using a random forest classifier. Using the PCs and known super population assignment (AFR=African, EUR=European, EAS=East Asian, AMR=American, and SAS=South Asian) from the 1000 Genome 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).

Genetic Ancestry Proportion Estimation

We estimated admixture proportions using ADMIXTURE by first analyzing the 1000 Genomes Project in unsupervised mode to learn allele frequencies.[21] Then, we projected the learned allele frequencies onto the Genographic samples to obtain the admixture proportions. We ran ADMIXTURE with k=2-9 and chose k = 5 as the most stable representation.

UMAP

We applied the Uniform Manifold Approximation and Projection (UMAP) method to visualize subcontinental structure.[23,24] We first combined the PCs of the Genographic samples and the 1000 Genome 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 Materials and Methods.

Estimating Effective Migration Surfaces

We estimated migration and diversity relative to geographic distance using the estimating effective migration surfaces (EEMS) method for Genographic individuals that were classified under African, European, and Native American ancestries.[25] 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 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.

Haplotype Calling and Network Construction

We used IBDSeq version r1206 to generate shared identity-by-descent (IBD) segments from genotype data for all unrelated individuals.[38] Unlike other IBD detection algorithms, IBDseq does not reply 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 3cM. 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 to keep edges with cumulative IBD sharing is ≥12 cM and ≤72 cM, as previously described.[8]

Detection of IBD Clusters

To identify clusters of related individuals in the haplotype network, we used 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). 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 with 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) amongst 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 (African Americans, Acadians).

Mapping IBD Clusters

We mapped individuals using their present-day geographic location. We aggregated individuals from the same county using the postal code to county FIPS code mapping provided by the US Census. Longitude and latitude points of each county was found using the same data from the US Census. We identified enriched counties for each cluster by performing a Fisher's exact test on each county that had ≥30 individuals to obtain an odds ratio and significance value. We mapped only counties with statistically significant ($p<0.05$) enrichment and an odds ratio (OR) of greater than 1. The size of the circles is scaled to the number of individuals in each location.

Runs of Homozygosity

We used PLINK v1.90b3.39 to infer runs of homozygosity with a window of 25 SNPs.[39] We calculated the cumulative runs of homozygosity (cROH) size by summing the lengths of homozygous segments.

Local Ancestry Inference

We inferred local ancestry with RFMix v1.5.4 for Genographic samples in clusters that were annotated as Hispanics/Latinos and African Americans.[29] 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 as the reference population. We ran RFMix with the default minimum window size (0.2 cM) and a node size of 5 with the flags: -w 0.2, -n 5. Global ancestry proportions were derived by quantifying the proportions of total local ancestry tracts for each ancestry.

Genetic Divergence

We computed weighted Weir-Cockerham F_{ST} estimates for each pair of haplotype clusters using PLINK v1.90b3.39.[39] 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*.[40] We visualized the tree using Interactive Tree Of Life.[41]

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.

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

Acknowledgement

We thank the National Geographic Genographic Project participants who consented to research participations 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., 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.

Author Contributions

C.L.D. and A.R.M. designed the study, performed research, and wrote the manuscript. R.S.W. founded and formerly directed the Genographic Project. M.G.V. coordinated and supervised the Genographic Project. M.M.V., C.H.Y., and R.T. contributed to the data aggregation and data analysis. A.R.M., C.R. and M.J.D. supervised research. All authors reviewed the manuscript.

Conflicts of Interest

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.

References

1. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Tishkoff, S. A. *et al.* The Genetic Structure and History of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
3. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci.* **107**, 786–791 (2010).
4. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc. Natl. Acad. Sci.* **107**, 8954–8961 (2010).
5. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
6. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
7. Baharian, S. *et al.* The Great Migration and African-American Genomic Diversity. *PLOS Genet.* **12**, e1006059 (2016).
8. Han, E. *et al.* Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat. Commun.* **8**, 14238 (2017).
9. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.* **96**, 37–53 (2015).
10. Wang, S. R. *et al.* Simulation of Finnish Population History, Guided by Empirical Genetic Data, to Assess Power of Rare-Variant Tests in Finland. *Am. J. Hum. Genet.* **94**, 710–720 (2014).
11. Bray, S. M. *et al.* Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proc. Natl. Acad. Sci.* **107**, 16222–16227 (2010).
12. Mooney, J. A. *et al.* Understanding the Hidden Complexity of Latin American Population Isolates. *Am. J. Hum. Genet.* **103**, 707–726 (2018).
13. Belbin, G. M. *et al.* Genetic identification of a common collagen disease in Puerto Ricans via identity-by-descent mapping in a health system. *eLife* **6**, e25060 (2017).
14. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nat. News* **538**, 161 (2016).
15. Martin, A. R. *et al.* Current clinical use of polygenic scores will risk exacerbating health disparities. *bioRxiv* 441261 (2019). doi:10.1101/441261
16. Caswell-Jin, J. L. *et al.* Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **20**, 234–239 (2018).

17. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
18. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).
19. Morales, J. *et al.* A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, (2018).
20. Elhaik, E. *et al.* The GenoChip: A New Tool for Genetic Anthropology. *Genome Biol. Evol.* **5**, 1021–1031 (2013).
21. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
22. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
23. McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2018).
24. Diaz-Papkovich, A., Anderson-Trocme, L. & Gravel, S. Revealing multi-scale population structure in large cohorts. (2018). doi:10.1101/423632
25. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
26. US Census Bureau. The Great Migration, 1910 to 1970. *U.S. Census* Available at: https://www.census.gov/dataviz/visualizations/020/. (Accessed: 21st February 2019)
27. Massey, D. S., Rugh, J. S. & Pren, K. A. The Geography of Undocumented Mexican Migration. *Mex. Stud. Mex.* **26**, 129–152 (2010).
28. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
29. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
30. Passel, J. S. & Fix, M. U.S. Immigration in a Global Context: Past, Present, and Future. *Indiana J. Glob. Leg. Stud.* **2**, 5–19 (1994).
31. Grieco, E. M., Trevelyan, E., Larsen, L., Acosta, Y. D. & Gambino, C. The Size, Place of Birth, and Geographic Distribution of the Foreign-Born Population in the United States: 1960 to 2010. *Popul. Div. Work. Pap. No 96 US Census Bur.* 38
32. Pemberton, T. J. *et al.* Genomic Patterns of Homozygosity in Worldwide Human Populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).

33. Moorjani, P. *et al.* Genetic Evidence for Recent Population Mixture in India. *Am. J. Hum. Genet.* **93**, 422–438 (2013).
34. Tadmouri, G. O. *et al.* Consanguinity and reproductive health among Arabs. *Reprod. Health* **6**, 17 (2009).
35. Hussain, R. & Bittles, A. H. Assessment of association between consanguinity and fertility in Asian populations. *J. Health Popul. Nutr.* **22**, 1–12 (2004).
36. Erlich, Y., Shor, T., Pe'er, I. & Carmi, S. Identity inference of genomic data using long-range familial searches. *Science* **362**, 690–694 (2018).
37. Regalado, A. More than 26 million people have taken an at-home ancestry test. *MIT Technology Review* Available at: https://www.technologyreview.com/s/612880/more-than-26-million-people-have-taken-an-at-home-ancestry-test/. (Accessed: 21st February 2019)
38. Detecting identity by descent and estimating genotype error rates in sequence data. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/24207118. (Accessed: 21st February 2019)
39. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
40. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
41. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242-245 (2016).

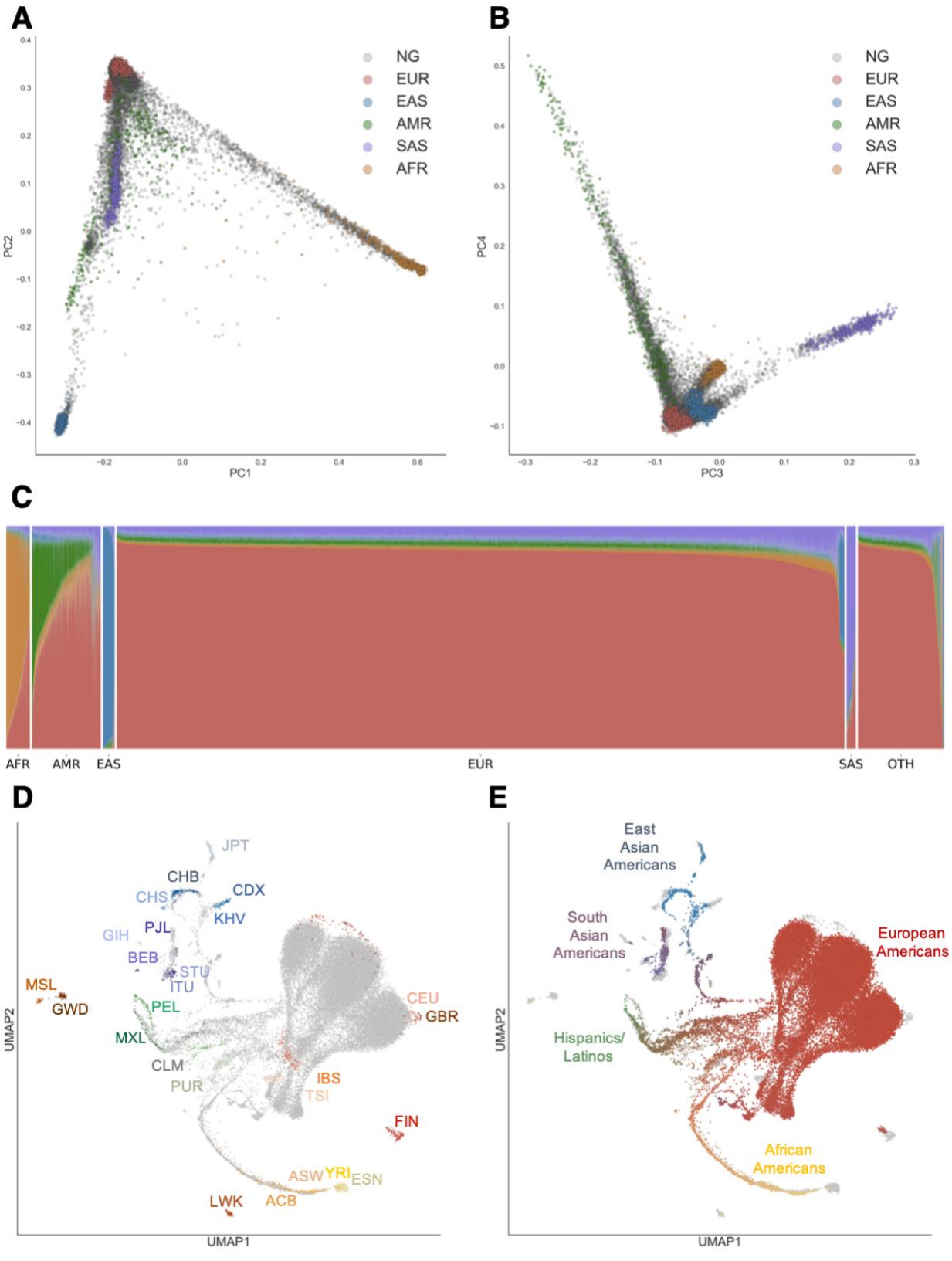


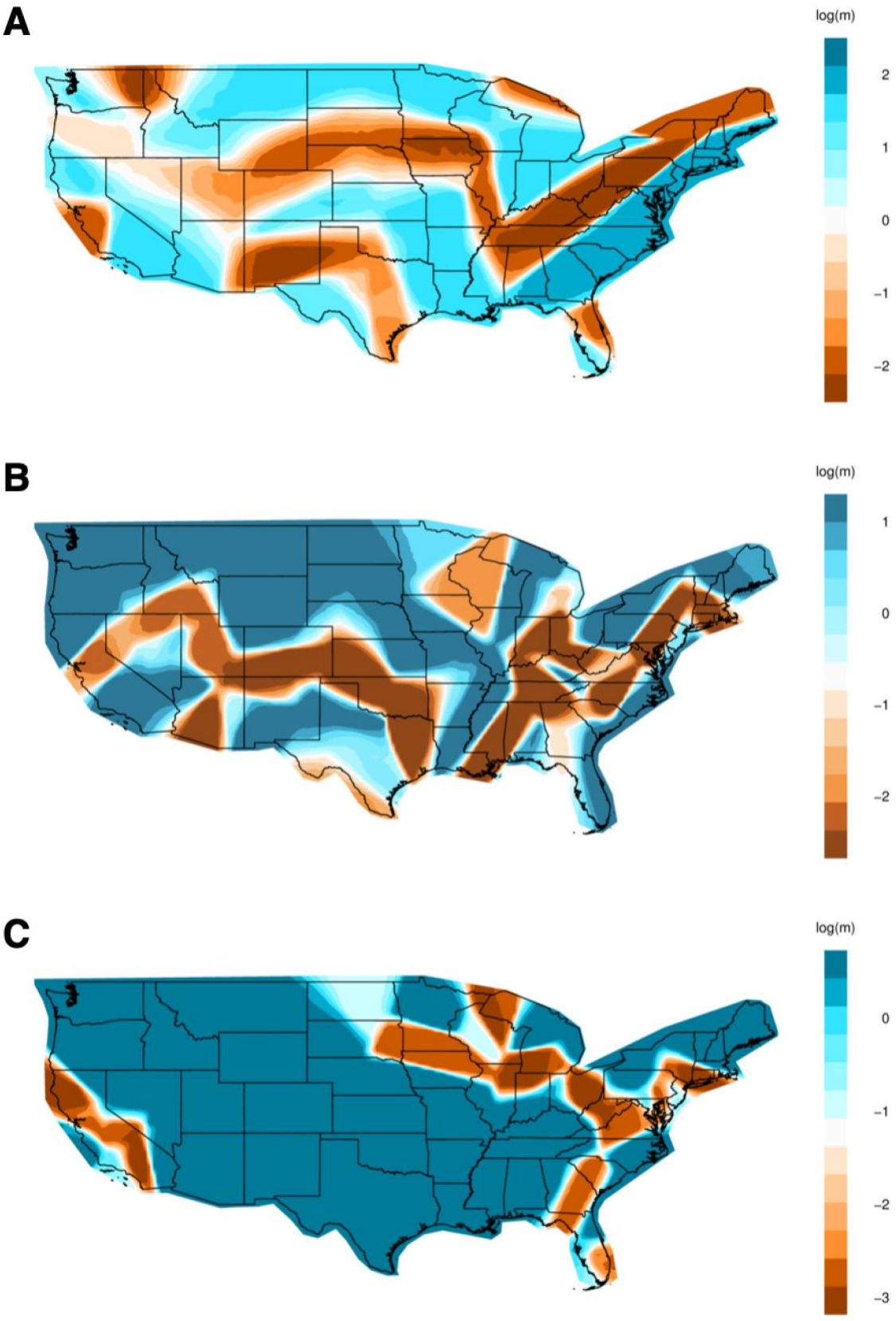
Figure 1. Genetic Diversity of the US Population

(A) Principal Components Analysis of individuals in the United States and in the 1000 Genome Project. Each individual is represented by a single dot. Individuals in this study are colored in grey while 1000 Genome Population individuals are colored by super population (EUR = European, AFR = African, AMR = Admixed American, EAS = East Asian, SAS = South Asian). Principal components (PC) 1 and PC2 are shown.

(B) Similar to (A), with PC3 and PC4 shown.

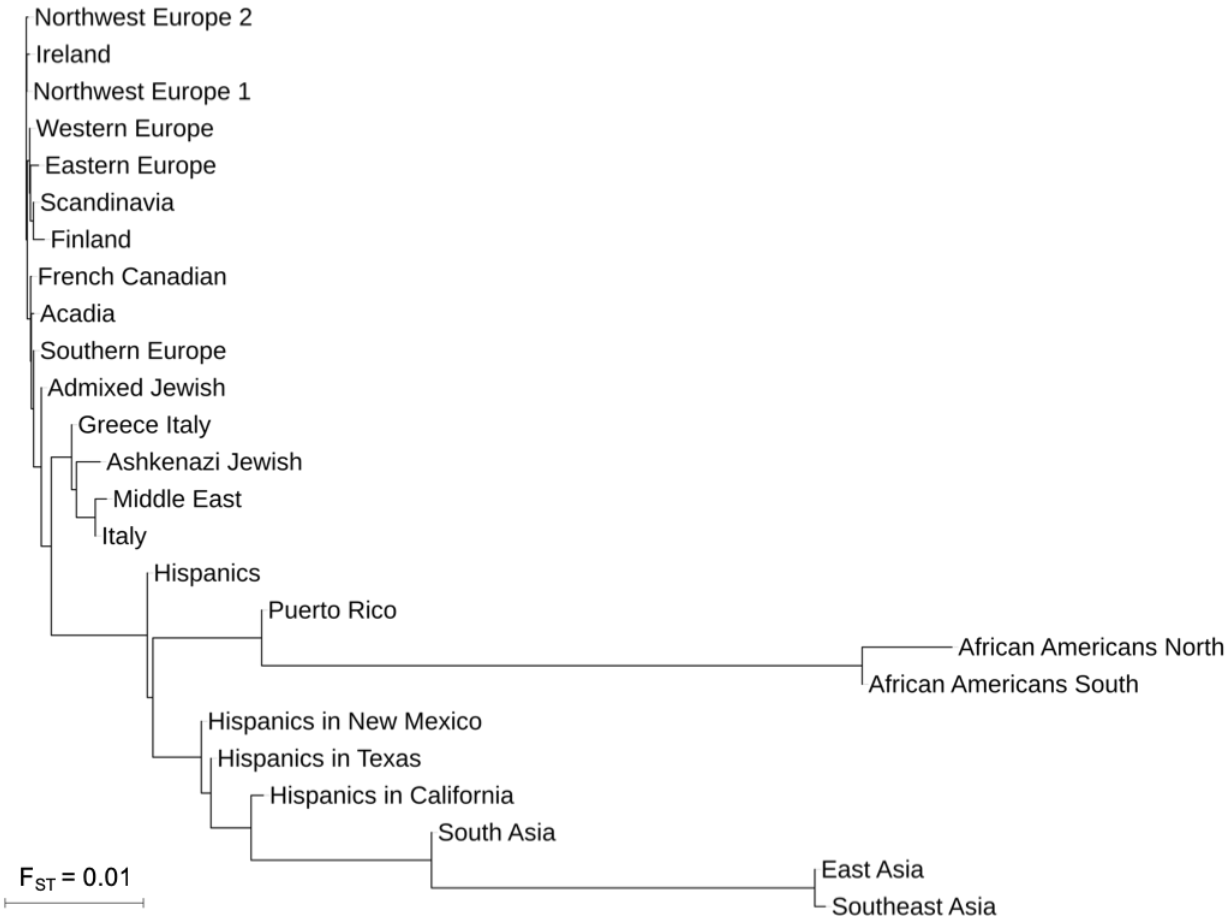
(C) ADMIXTURE analysis at K=5 of individuals in this study. Each individual was assigned a continent-level ancestry label using a Random Forest model trained on the super population labels and the first 10 PCs of the 1000 Genome Project dataset. OTH = individuals who did not meet the 90% confidence threshold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

(D) 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 Genographic Project individuals from this study are in grey. In (E), 1000 Genome Project individuals are colored in grey while Genographic Project 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, NAT or Native American = green, EAS = blue, SAS = purple). Distances in UMAP do not directly correspond to genetic distance. See Materials and Methods for specific population labels.



645

646 **Figure 2. Migration Rates of African Americans, Hispanics/Latinos, and Europeans within**
647 **the United States.**
648 (A) - (C) Migration rates inferred with EEMS for African Americans (A), Hispanics/Latinos (B),
649 and Europeans (C). Colors and values correspond to inferred rates, *m*, relative to the overall
650 migration rate across the country. Shades of blue indicate logarithmically higher migration (i.e.
651 log(m) = 1 represents effective migration that is ten-fold faster than the average) while shades
652 of orange indicate migration barriers.
653



654
655 **Figure 3. Genetic differentiation of haplotype clusters**
656 Unrooted phylogenetic tree of haplotype clusters was constructed using the neighbor joining
657 method with F_{ST} as genetic distance. Negative branch lengths were converted to zero.

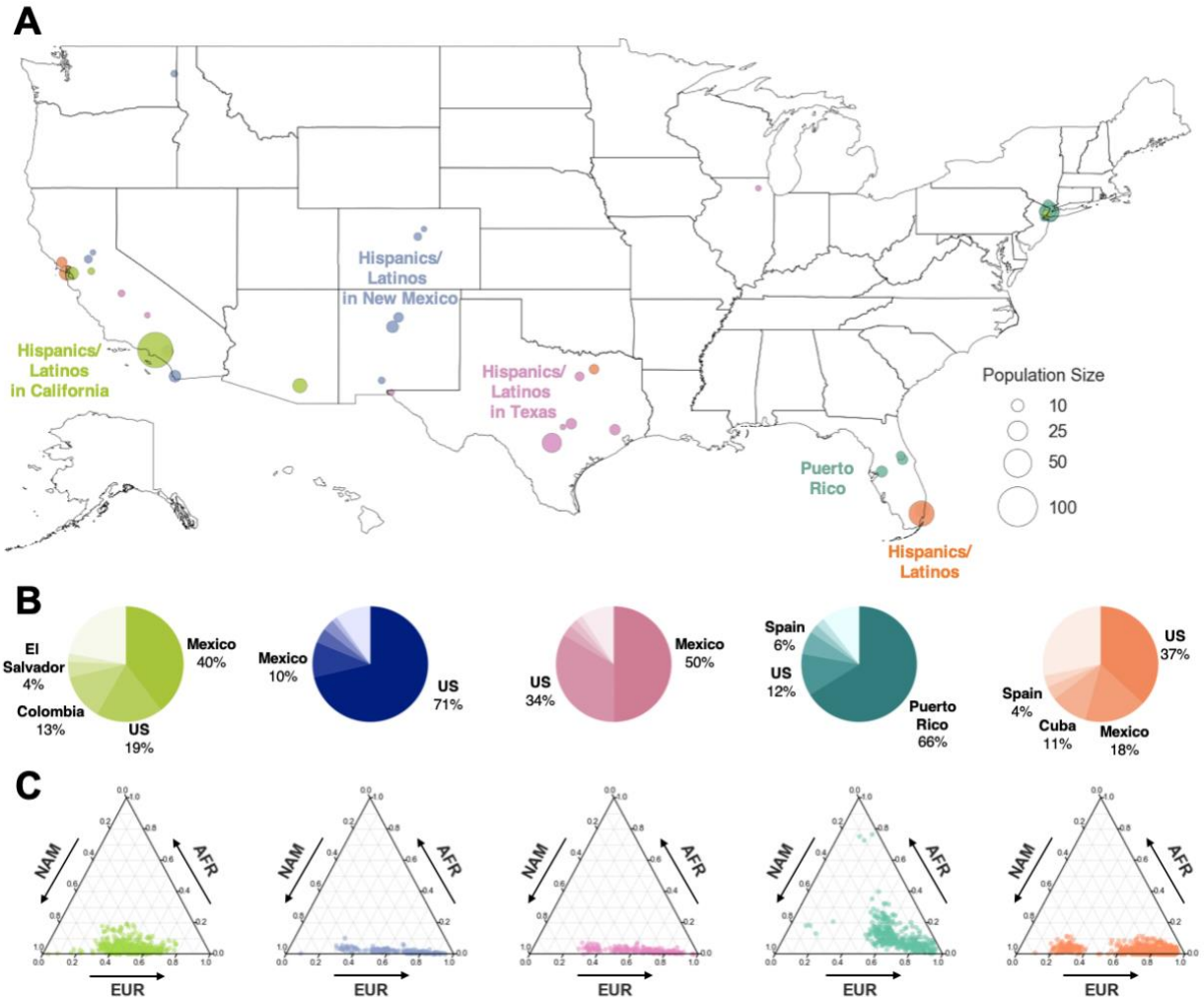


Figure 4. Distribution of Hispanic/Latino Haplotype Clusters

(A) Map of counties in which Hispanic/Latino haplotype clusters are enriched. Each dot corresponds to a county, 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 only the top ten locations with the highest odds ratios are shown for each cluster. Maps showing the full distribution for each haplotype cluster can be found in the supplement (**Figure S8**).

(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 slices; 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.

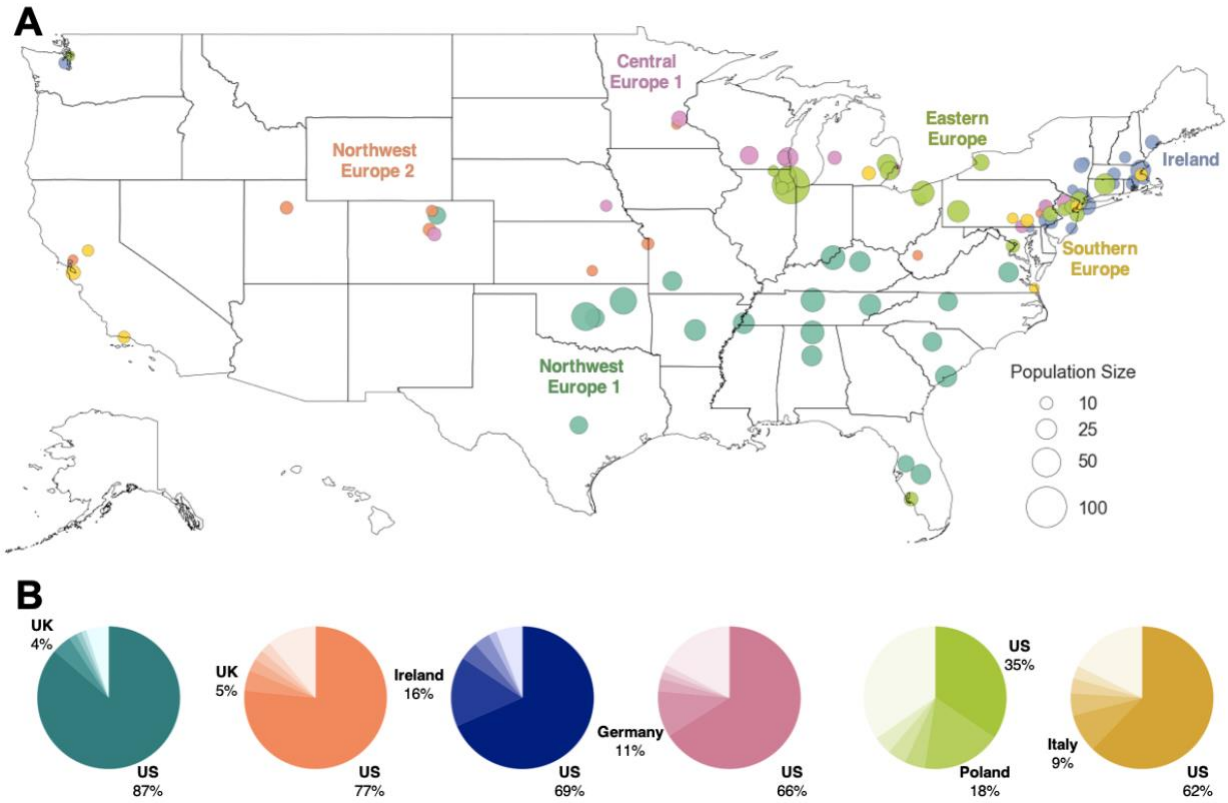


Figure 5. Distribution of European American Haplotype Clusters

(A) Geographic distributions of haplotype clusters corresponding to regional European ancestries. Each county containing present-day individuals is represented by a dot. The top 20 locations with the highest odds ratio are shown for each cluster. Maps showing the full distribution for each cluster can be found in the supplement (**Figure S8**).

(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 visualized as individual slices; the remaining birth origins are aggregated into one slice (lightest color).

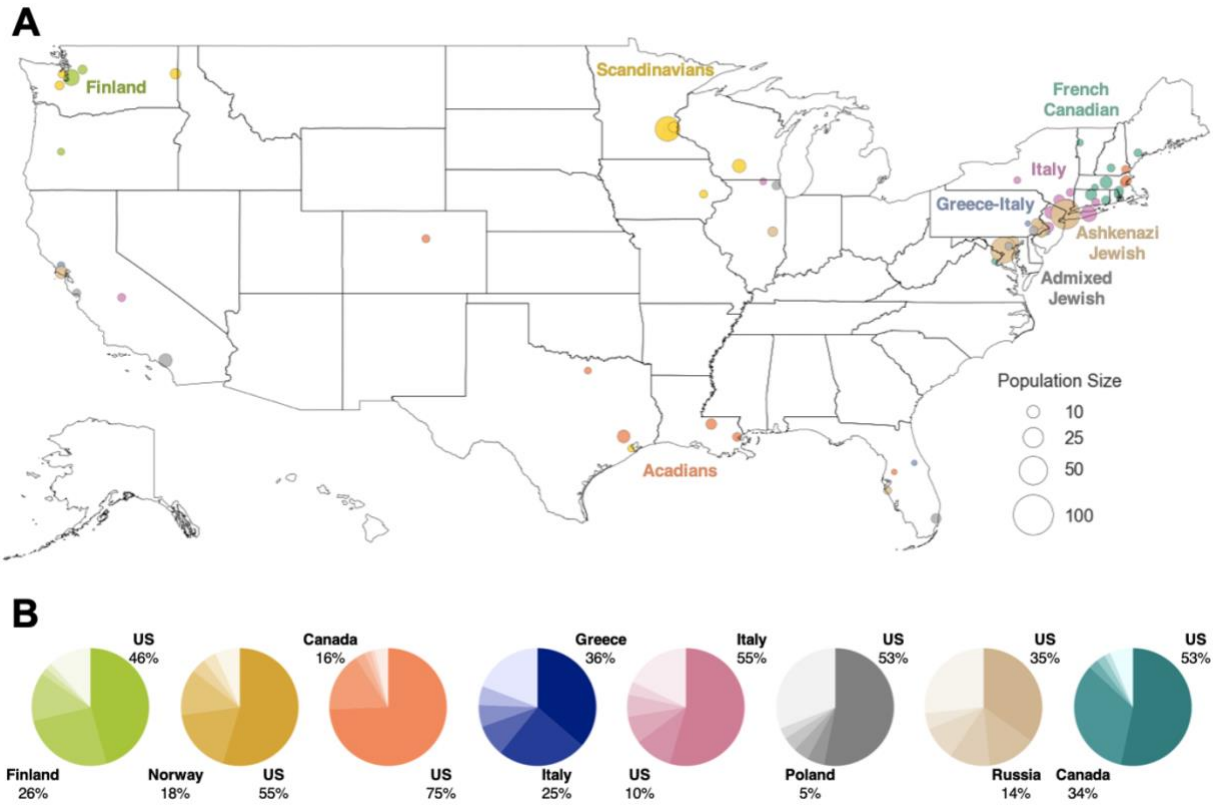


Figure 6. Distribution of European American Haplotype Clusters

(A) Present-day location of individuals in clusters of more genetically isolated European populations, similar to Figure 5A. For clarity, the top ten locations with the highest odds ratio are shown for each cluster.

(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 slices; the remaining birth origins are aggregated into one slice (lightest color).

Cluster	Samples	Median Cumulative ROH	Median Cumulative IBD
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

Table 1. Summary of Haplotype Clusters

Cumulative runs of homozygosity (cROH) was calculated by summing the regions of continuous homozygous segments. Cumulative IBD was determined by summing IBD segments of ≥ 3 cM and filtering for only pairs ≥ 12cM and ≤ 72 cM. Statistics were determined within haplotype clusters, rather than across the ancestrally heterogeneous and imbalanced full network.