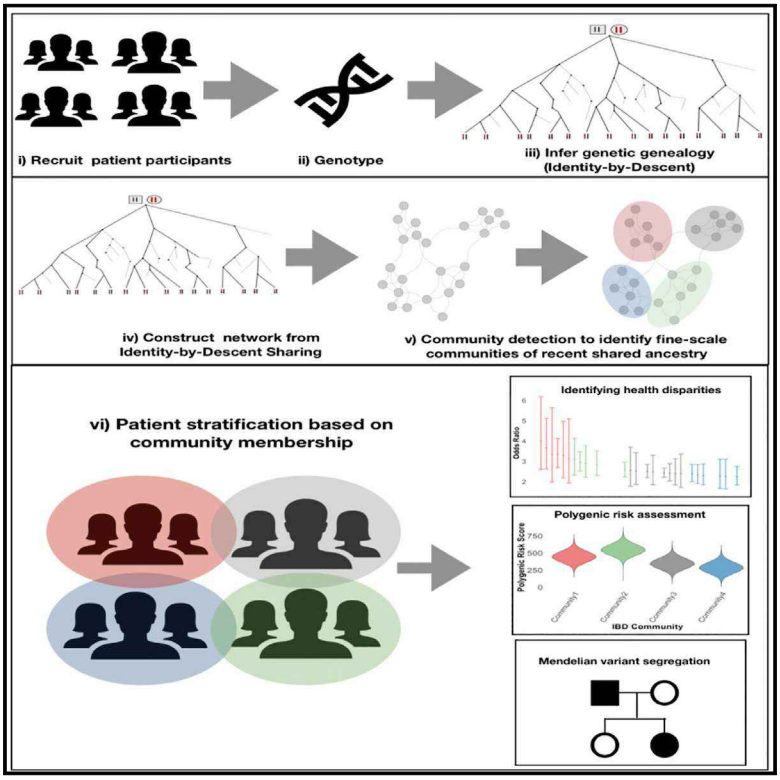
Cell

Toward a fine-scale population health monitoring system

Graphical abstract



Highlights

- Genomic data linked to health records capture demography in health systems
- Genetic networks reveal recent common ancestry in diverse populations
- Evidence of many founder populations in New York City
- Fine-scale population structure impacts genetic risk predictions

Authors

Gillian M. Belbin, Sinead Cullina, Stephane Wenric, ..., Noah A. Zaitlen, Christopher R. Gignoux, Eimear E. Kenny

Correspondence

eimear.kenny@mssm.edu

In brief

Taking a quantitative approach to genetic ancestry in health systems furthers understanding of disease burdens specific to fine-scale populations and the environmental and demographic ties that can impact disease.

Belbin et al., 2021, Cell *184*, 2068–2083
April 15, 2021
https://doi.org/10.1016/j.cell.2021.03.034

CellPress

CellPress

Cell

Article

Toward a fine-scale population health monitoring system

Gillian M. Belbin,[1,2,3] Sinead Cullina,[1,3] Stephane Wenric,[1,3] Emily R. Soper,[1] Benjamin S. Glicksberg,[4,5] Denis Torre,[4] Arden Moscati,[3] Genevieve L. Wojcik,[6] Ruhollah Shemirani,[7] Noam D. Beckmann,[4] Ariella Cohain,[4] Elena P. Sorokin,[6] Danny S. Park,[4,8] Jose-Luis Ambite,[7] Steve Ellis,[3] Adam Auton,[9] CBIPM Genomics Team,[1,3] Regeneron Genetics Center,[10] Erwin P. Bottinger,[5] Judy H. Cho,[3] Ruth J.F. Loos,[2,3] Noura S. Abul-Husn,[1,2,4] Noah A. Zaitlen,[11] Christopher R. Gignoux,[12] and Eimear E. Kenny[1,2,4,13,14,*]

[1]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[2]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[3]The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[4]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[5]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[6]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA
[7]Information Science Institute, University of Southern California, Marina del Rey, CA 90089, USA
[8]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA
[9]Department of Genetics, Albert Einstein College of Medicine, New York, NY 10461, USA
[10]Regeneron Genetics Center, Tarrytown, New York, NY 10591, USA
[11]Department of Neurology, University of California, Los Angeles, Los Angeles, CA 90033, USA
[12]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[13]Senior author
[14]Lead contact
*Correspondence: eimear.kenny@mssm.edu
https://doi.org/10.1016/j.cell.2021.03.034

SUMMARY

Understanding population health disparities is an essential component of equitable precision health efforts. Epidemiology research often relies on definitions of race and ethnicity, but these population labels may not adequately capture disease burdens and environmental factors impacting specific sub-populations. Here, we propose a framework for repurposing data from electronic health records (EHRs) in concert with genomic data to explore the demographic ties that can impact disease burdens. Using data from a diverse biobank in New York City, we identified 17 communities sharing recent genetic ancestry. We observed 1,177 health outcomes that were statistically associated with a specific group and demonstrated significant differences in the segregation of genetic variants contributing to Mendelian diseases. We also demonstrated that fine-scale population structure can impact the prediction of complex disease risk within groups. This work reinforces the utility of linking genomic data to EHRs and provides a framework toward fine-scale monitoring of population health.

INTRODUCTION

Populations around the world often have differential rates of disease due to a combination of genetic variation and environmental factors. Understanding the differences in disease burdens according to demographic factors is the basis of epidemiological research and is fundamentally important to clinical care and public health. Most studies of human disease begin by sampling from predefined populations, which are usually identified on the basis of race, ethnicity, cultural identity, or geography. However, these population categories are often too coarse to capture all of the environmental and demographic ties that can impact disease burdens. In the United States, individuals with roots from Latin America are often broadly classified as Hispanic and/or Latinx, but sub-groups with origins from different countries in the Americas may have different rates of disease. For example, populations of Puerto Rican (PR) descent have one of the highest asthma rates in the world, while populations of Mexican descent have one of the lowest (Carter-Pokras and Gergen, 1993; Homa et al., 2000).

With the advent of large-scale population-based DNA biobanks in health systems, new opportunities are available to characterize the links between demography and a broad range of health outcomes (Collins, 2012; Dewey et al., 2016; Belbin et al., 2017; Amendola et al., 2018; Abul-Husn and Kenny, 2019). Knowledge about genetic variation shared across human populations can aid in understanding the demographic events that might impact disease burden across populations. For

example, variants in the *APOL1* gene, which confer a substantially increased risk of kidney disease and cardiovascular disease, arose in Africa; were first discovered in African American (AA) populations (Kao et al., 2008; Parsa et al., 2013); and are mainly studied in African (Hassan et al., 2020; Ekrikpo et al., 2020; Thakoordeen-Reddy et al., 2020; Nqebelele et al., 2019) or AA (Miller et al., 2020; Umeukeje and Young, 2019; Gutiérrez et al., 2020) populations. However, *APOL1* risk variants exist at appreciable frequencies among many populations across the Americas that historically share African genetic ancestry, but may not self-identify as African or AA, and are subsequently underrepresented in *APOL1* research (Nadkarni et al., 2018; Kramer et al., 2017). This suggests that while self-reported race/ethnicity (R/E) information can be useful in assessing epidemiological risk, in some cases it may be limiting. Furthermore, this information may be inaccurately captured or missing in health systems and may not accurately recapitulate the inherent population structure actually impacting disease risk (Smith et al., 2010; Klinger et al., 2015).

High-density genome-scale data have long been used to examine genetic differences between populations that, in turn, can be used to infer population genetic history. Popular techniques are algorithmic-based methods such as principal-component analysis (PCA) (Price et al., 2006; Menozzi et al., 1978; Patterson et al., 2006) and model-based methods such as ADMIXTURE (Pritchard et al., 2000; Tang et al., 2005; Alexander et al., 2009), which estimate genetic ancestry by assessing genomic variants in aggregate. Other powerful approaches infer more fine-scale genetic ancestry by using local haplotypes along chromosomal segments (Gusev et al., 2009; Lawson et al., 2012; Maples et al., 2013). One such approach identifies residual signatures of distant genetic relationships that may be detected by inferring long-range haplotypes (Browning and Browning, 2012). Although genetic traces of our distant ancestors will decay rapidly over time, long haplotypes that have been co-inherited identical by descent (IBD) from some of our more recent ancestors may persist. At a population level, these can be analyzed in aggregate to infer distant relationships to a set of shared ancestors. We expect that individuals in a population who are distantly related to one another may also be more likely to share recent population history. This, in turn, may be linked to sharing of culture, environment, and correlations in disease risk.

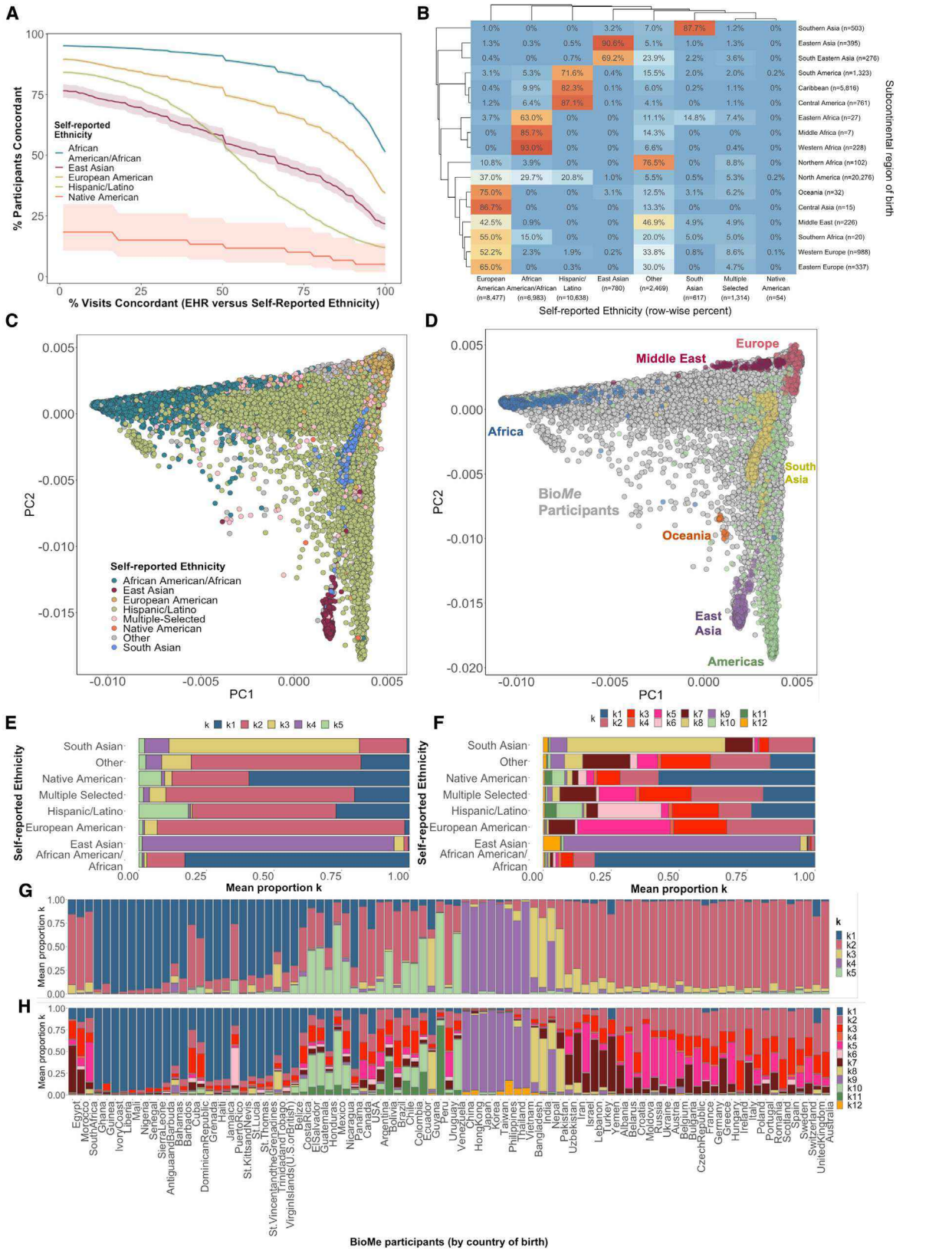
In this study, we examined fine-scale population structure in Bio*Me*, a highly diverse multi-ethnic biobank ascertained through the Mount Sinai Health System in New York City (NYC). When comparing EHR-recorded and self-reported R/E, we observed varying rates of discordance, demonstrating that R/E information can be inconsistently captured in EHRs, particularly for Hispanic/Latinx and Asian populations. Exploring genetic ancestry within Bio*Me* revealed further complexity, with distinct patterns of continental and subcontinental genetic substructure within self-reported R/E groups. To investigate this substructure further, we analyzed IBD sharing in Bio*Me* and applied an unsupervised, scale-free network modeling method to uncover clusters of populations informed by patterns of recent demography. We revealed 17 distinct communities highly correlated with recent migratory patterns to NYC and demonstrated that some of these communities harbor signatures of founder events, the timings of which coincided with the era of colonization of the Americas. We linked these communities to more than 1,700 health outcomes and found distinctive health patterns of disease risk, uncovering more than 1,100 examples of statistically significant differences in health outcomes between populations, some of which point to unknown or underappreciated population-specific disease risks. We then demonstrate elevated prevalence of founder variants for genetic disorders in two IBD founder communities relative to self-reporting R/E labels and country-of-origin information, suggesting that IBD communities allow for refined inference of prevalence of pathogenic variants in at-risk populations. Finally, we demonstrate significant differences in the distributions and predictive power of polygenic risk scores (PRSs) between two European ancestry IBD communities, indicating that fine-scale population structure also impacts our understanding of the genetics of complex traits. This work demonstrates the value of the application of genetic ancestry in medicine and how understanding fine-scale population structure could improve population health monitoring.

RESULTS

Evaluating the relationship between R/E, genetic ancestry, and geographical origin in the large Mount Sinai Health System in NYC

We evaluated R/E in a large (N = 36,061), diverse biobank (Bio*Me*) linked to EHRs in the Mount Sinai Health System in NYC. To understand how R/E information is captured in a large, urban health system, we compared the self-reported R/E surveyed during the Bio*Me* enrollment (self-reported R/E; Table S1) to R/E recorded in the Mount Sinai EHRs. We first restricted analysis to participants who reported only one of the following R/Es in the enrollment survey for one of five population groups: European American (EA; N = 9,830; see Star Methods), AA (N = 7,976), Hispanic/Latino (HL; N = 11,544), East and Southeast Asian (ESA; N = 965), and Native American (NA; N = 61) (total: 84.2% of Bio*Me*; N = 30,376 participants and n = 1,310,279 total visits). We mapped self-reported R/E for these individuals to R/E information extracted from the EHR at each independent visit. Self-reported and EHR-recorded R/E was concordant in 64.5% of total visits, or 71.6% of visits if excluding visits where R/E was designated as "unknown" (resulting in the exclusion of n = 129,678 healthcare visits), and we observed significant differences in concordance between population groups (Figure 1A). The remaining 15.8% of Bio*Me* participants self-reported a different R/E that did not directly correspond to an EHR-recorded R/E variable, precluding us from making a direct comparison between the two (Figure S1A). Overall, these analyses support that R/E data are often poorly and inconsistently captured, particularly for non-EA populations, during hospital visits (Banda et al., 2015; Klinger et al., 2015).

To explore the relationship between self-reported R/E and genetic ancestry, we estimated global genetic ancestry proportions for a subset of Bio*Me* participants (N = 31,705) genotyped on the Global Screening Array (GSA). We first examined the correlation between self-reported R/E and self-reported region of origin, revealing a complex relationship between R/E and subcontinental region (Figure 1B). We then merged Bio*Me* participants with a



(legend on next page)

reference panel of 87 populations representing ancestry from seven continental or subcontinental regions. Using PCA, we demonstrate that Bio*Me* participants represent a continuum of genetic diversity, even within self-reported R/E groups (Figure 1C). Bio*Me* participants fall between African and non-African reference panels on principal component 1 (PC 1) and between European, Asian, American, and Oceanian reference panels on PC 2 (Figure 1D). The first 10 PCs can also be represented as a low-dimensional topological map using the uniform manifold approximation and projection (UMAP) algorithm, presenting many clusters that are roughly grouped by continent and others with little or no overlap with reference panels, presenting a challenge for their interpretation (Figure S1B).

We also explored population demography using ADMIXTURE, a model-based approach that applies a pre-set number of putative ancestral populations to seek the best fit of ancestral clusters in the data (Figure S1C). Analysis, including reference panels from Thousand Genomes Project (TGP; N = 2,504), the Human Genome Diversity Panel (HGDP; N = 986), and the Population Architecture using Genomics and Epidemiology (PAGE) study (N = 700), that was fit to five ancestral populations (k = 5) recapitulates five continental-level ancestral components corresponding to African, European, East Asian, South Asian, and Amerindigenous ancestry. As expected, self-reported AA and HL participants exhibited varying degrees of ancestry from Europe, Africa, and the Americas; however, several individuals also appear to have appreciable (>10%) levels of South Asian (2.2% of individuals) or East Asian (0.6% of individuals) ancestry (Figure 1E). Participants who self-identify as ''other'' exhibit varying proportions of admixture from all 5 different continent groups. At k = 12, we observe an appreciable Oceanian component (6.3%; Figure 1F, orange) in participants who self-identify as East Asian, two distinct Amerindigenous components (Figure 1F, light and dark green) present in both self-reported NA and HL participants. By linking Bio*Me* participants born to self-reported country of birth, we were able to explore diversity linked to more recent demography (Figure 1G). We demonstrate European, African, and Amerindigenous ancestry proportions consistent with previous reports in participants born in Puerto Rico (59.3, 26.0, and 14.2%, respectively), Dominican Republic (50.3, 40.9, and 7.4%), Jamaica (13.0, 83.2, and 0.6%), and Cuba (66.4, 26.8, and 4.6%) (Moreno-Estrada et al., 2013). We observe appreciable levels of East and South Asian ancestry in participants born in Trinidad and Tobago (4.7 and 24.5%), the Bahamas (10.3 and 7.7%), and Guyana (6.1 and 50.6%), which is consistent with historical accounts of South Asian migration to the Caribbean. At k = 12 (Figure 1H). We also observe complex structure in the European component including the resolution of a component most predominant in participants who self-report as Jewish (83.9 versus 4.3% in participants who self-reported as ''white/Caucasian'' only) in a subset of the US-born participants, as well as participants born in North African and the Middle East. Furthermore, we observe two distinct European components that likely represent a northern-southern European cline in genetic ancestry across Europe (Novembre et al., 2008) and a distinct European component that appears predominantly in admixed populations from the Americas. This component is particularly notable in Caribbean-born participants and is present to a lesser extent in Portuguese- and Spanish-born Bio*Me* participants. Taken together, this information reveals Bio*Me* to be a rich source of ancestral genetic diversity, including many populations that are otherwise poorly represented in biomedical genomics research.

Detecting communities of recent shared ancestry in NYC

To further investigate the complex genetic ancestry in Bio*Me*, we explored patterns of distant relatedness reflecting fine-scale structure. We first detected pairwise genomic tracts inherited IBD (pairwise haplotypic tracts of 3 cM or longer) between all participants and 2,504 participants comprising 26 global populations of the 1000 Genomes Project phase 3 (Abecasis et al., 2012). We used this information to construct a network of pairwise IBD sharing between all individuals who were not inferred to be directly related (see STAR Methods; Figure S2A). To identify ''communities'' of individuals enriched for recent, shared

Figure 1. Evaluating the relationship between race/ethnicity, genetic ancestry, and geographical origin in the large Mount Sinai Health System in New York City (NYC)

(A) Plot of the percentage concordance between EHR-recorded and self-reported R/E for 5 categories for which there was a one-to-one mapping between the Bio*Me* survey and EHR R/E variables. The x axis represents the percentage of interactions a given individual has with the healthcare system that were concordant between EHR and self-reporting, while the y axis represents the percentage of individuals within a self-reported R/E category that meet that threshold. Ribbons around the lines represent the 95% confidence intervals.

(B) Heatmap representing the correlation between self-reported R/E and subcontinental region of birth for the subset of Bio*Me* participants genotyped on the GSA array for whom both self-reported R/E and country-of-birth information were available.

(C) Principal-component analysis (PCA) for N = 31,705 genotyped Bio*Me* participants colored by self-reported R/E.

(D) Unsupervised PCA plot of Bio*Me* participants (gray) with the addition of reference panels from 7 continental/subcontinental regions, namely Africa (blue), the Middle East (dark red), Europe (light red), South Asia (yellow), East Asia (purple), the Americas (green), and Oceania (orange), revealing complex and varied patterns of admixture and continental genetic ancestry among Bio*Me* participants.

(E) The mean ADMIXTURE components for Bio*Me* participants at k = 5 stratified by self-reported R/E. Each color represents an inferred ancestry proportion that at k = 5 corresponds to African (blue), European (red), South Asian (yellow), East Asian (purple), and Amerindigenous (green) ancestry. The proportions of each color per column represent the mean ancestry proportions of each of these five components within each self-reported R/E group.

(F) The mean admixture components for Bio*Me* participants at k = 12 stratified by self-reported R/E. At k = 12, ADMIXTURE returns patterns of subcontinental genetic ancestry among groups.

(G) The mean admixture components at k = 5 for Bio*Me* participants grouped by self-reported country of birth (only countries represented by N $\geq$ 10 participants are displayed).

(H) The mean admixture components at k = 12 for Bio*Me* participants based on self-reported country of birth.

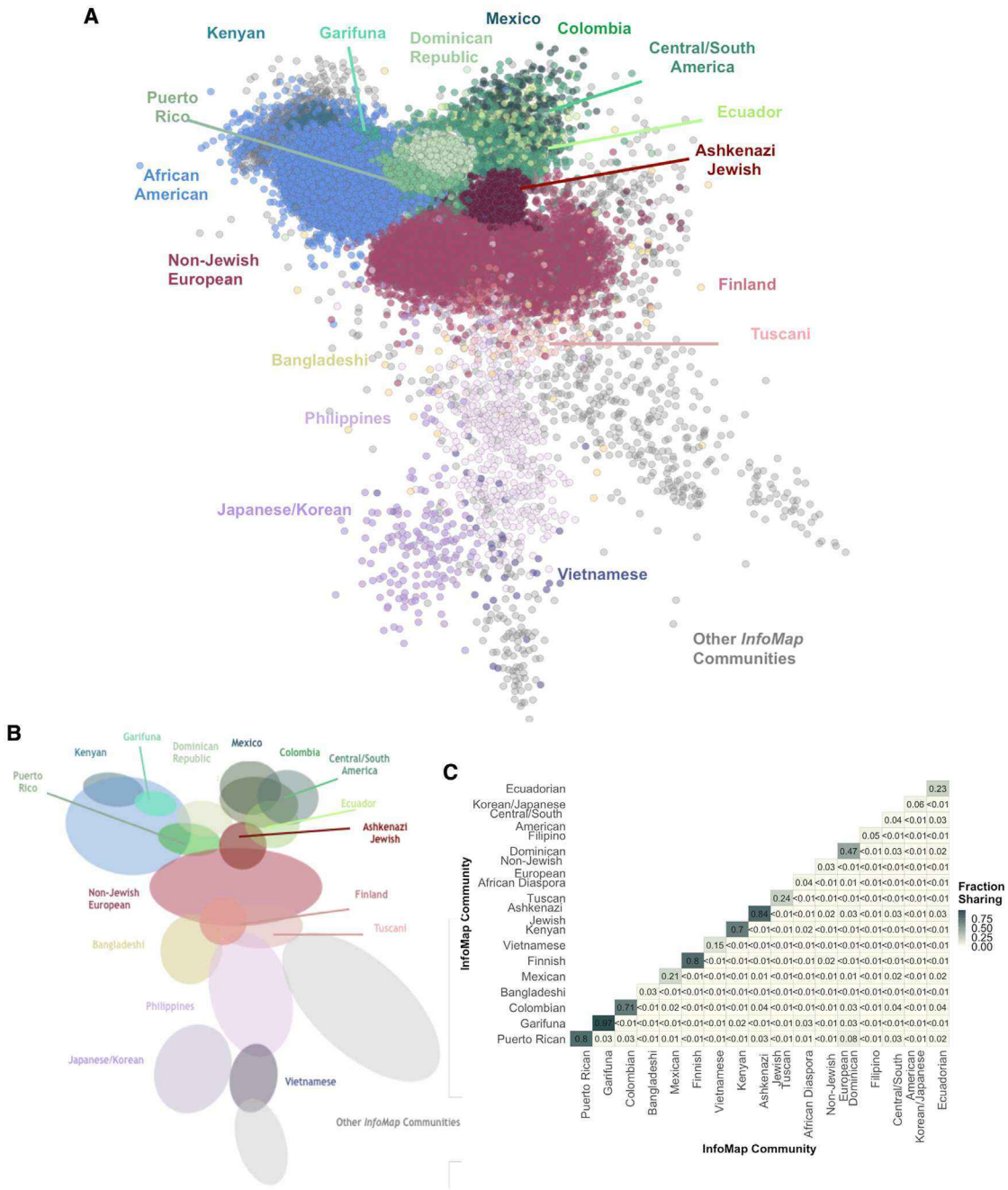


Figure 2. Detecting communities of recent shared ancestry in NYC

(A) Network of IBD sharing among Bio*Me* participants, colored by community membership as inferred by InfoMap (for the top 17 communities only and only showing nodes with >30 connections to other nodes). Returned communities reflect an enrichment of IBD sharing among individuals of recent, shared genetic ancestry and thus recapitulate fine-scale population substructure.

(B) Schematic representation of the 17 distinct IBD communities recovered by InfoMap.

(C) Heatmap representing the population-level fraction of IBD sharing within and between inferred IBD communities reveals a high degree of modularity.

genetic ancestry, we performed community detection via flow-based clustering using the InfoMap algorithm (Rosvall and Bergstrom, 2008, 2011) (Figures 2A and 2B).

We observed that 96% of Bio*Me* participants fall into one of 17 distinct clusters containing at least 100 individuals that we refer to as IBD communities. Analysis of the inter- and intra-population-level sharing of these communities revealed a higher probability of pairwise IBD sharing within communities compared with between communities (Figure 2C). Additionally, in instances where IBD sharing does occur, the pairwise sum of IBD sharing is

also higher within than between communities (Figure S2B). We examined the topology of the IBD network and determined community assignments to be non-random through a node-level comparison of similar edges between 10 instances of a network based on pairwise IBD sharing and 100 instances of topologically similar random networks created with the Erdős-Rényi (ER) model (Durrett, 2006). The quantile of the Jaccard similarity coefficient obtained by comparing these networks corresponds to a permutation p value. Therefore, the communities detected by using the InfoMap algorithm are non-random and statistically significant ($p < 3.25 \times 10^{-4}$).

We hypothesized that these communities represent geographical or ethnic substructure within the Bio*Me* population. To test this, we compared IBD community membership to survey-derived population level information for Bio*Me* participants and calculated the positive predictive values (PPVs) for each population label ("ground truth") versus each identified community (predictor) (Figures S2C and S2D). Many communities had high PPVs (>0.9) with a single country of origin (6/17), including Puerto Rico, the Dominican Republic, Colombia, Ecuador, Mexico, and Ethiopia, likely reflecting recent migrations to NYC. However, a number did not, including one community that notably consisted of 85% of individuals in Bio*Me* who self-reported having Jewish ancestry in the survey (Figure S2E). Other IBD communities were also detected that transcended self-reported R/E labels and mapped across various different groups. For example, one community that, based on a combination of self-reported country-of-origin information and PCA analysis, is likely to represent Garifuna (a population of admixed African and Amerindigenous ancestry with recent ties to Central America and St. Vincent and the Grenadines). Individuals in this community (N = 113) were either born in Europe, Central America, or in the United States and self-identified as AA, HL, or other, but cluster together tightly in PC space (Figure S2F).

We next determined whether IBD community detection was better able to classify individuals based on recent genetic ancestry than current best practices, i.e., *k*-means clustering over PCA eigenvectors. We performed *k*-means clustering from $k = 5$ to $k = 20$ over the first 5 PCs calculated across all Bio*Me* participants. To measure accuracy, we calculated PPV, negative predictive value (NPV), sensitivity, and specificity for each inferred cluster using country-of-origin labels. We compared *k*-means clustering to IBD communities with a PPV > 0.9 for country-of-origin population labels: the PR, Dominican, Ecuadorian, Colombian, Mexican, and Ethiopian communities. For each of the six communities, we observed that PPV, NPV, and specificity were always the same or higher in IBD communities compared with *k*-means clustering with any value of *k* (Figure S3). We noted that while sensitivity can also be higher for IBD communities, in some cases it was lower than *k*-means clustering at low *k* values for Colombian, Mexican, and Ethiopian groups. We speculate that this may reflect population substructure within a country of origin or subsequent patterns of migration to NYC. Nevertheless, the IBD-based approach was the most accurate genetic ancestry method for detecting fine-scale population structure and was selected for the downstream analysis of genetic ancestry in medicine using health record data.

Signatures of founder effects in Bio*Me* communities

Patterns in the distribution and abundance of population-level IBD sharing are influenced by demographic events such as migration, founder effects, and population bottlenecking (Browning and Browning, 2012; Thompson, 2013). We observed evidence of founder effects in the form of elevated IBD sharing in multiple IBD communities (Figure 3A). These included the Ashkenazi Jewish (AJ) community (sample size N = 4,415; median pairwise sum of IBD sharing = 21.79 cM [95% confidence interval (CI) = 21.77–21.80], sum of runs of homozygosity [ROH] = 11.62 MB [95% CI = 11.08–12.02]) and the Finnish community (N = 120, IBD = 10.61 cM [95% CI = 10.34–10.91], ROH = 9.04 MB [95% CI = 7.32–12.74]), both of which are known and well-studied founder populations. Five other communities exhibited high levels of IBD sharing (IBD > 8 cM) and enrichment for autozygosity (ROH > 5 MB), namely PR (N = 5,452), Dominican (N = 1,971), Garifuna (N = 113), Colombian (N = 234), and Ecuadorian (N = 438) communities. By contrast, communities similar to well-studied, non-founder populations such as non-Jewish EA exhibited lower levels of IBD sharing (IBD < 5 cM) and autozygosity (ROH < 5 MB; Table S2). Examining the network topology, IBD communities corresponding to founder populations exhibit specific characteristics: they have a high clustering coefficient (e.g., $C = 0.92$ for AJ; $C = 0.97$ for Garifuna; $C = 0.85$ for PR), while non-founder populations have much lower values (e.g., $C = 0.05$ for AA; $C = 0.09$ for non-Jewish EA), and they exhibit a strongly bimodal degree distribution, with a high distance between the intra-community degree distribution and the inter-community distribution as summarized by the Wasserstein metric (e.g., $W = 0.82$ for AJ; $W = 0.96$ for Garifuna; $W = 0.78$ for PR), while non-founder populations have much lower values (e.g., $W = 0.03$ for AA; $W = 0.03$ for non-Jewish Europeans) (Figure 3B; Figure S4A; Table S3).

To understand the timing of the bottlenecks leading to the observed founder effects, we used the IBD_{Ne} software (Browning and Browning, 2015) to model each community's effective population size (N_e) over antecedent generations. For the communities representing populations from the Americas, the results are consistent with a population bottleneck occurring approximately 10–15 generations ago, which is coincident with historical accounts of the timings of European contact and subsequent colonization (Figure 3C; Figure S4B). The most profound bottleneck was observed in the Garifuna community, with a minimum N_e of 321 individuals (95% CI = 286–372, 10 generations ago), consistent with a previous report (Mathias et al., 2016). This analysis demonstrates that founder effects are potentially more ubiquitous than previously thought, with >25% of Bio*Me* participants falling within a community exhibiting founder effects, and this was particularly notable among HL populations. This indicates an under-recognized potential for reducing the genetically driven health burden within HL communities, echoing similar work in South Asian populations (Nakatsuka et al., 2017).

Characterizing the health phenome in communities of distantly related individuals

To explore the effect of IBD community membership on predisposition to EHR-captured health outcomes, we tested each community for its relative enrichment of health-related traits.

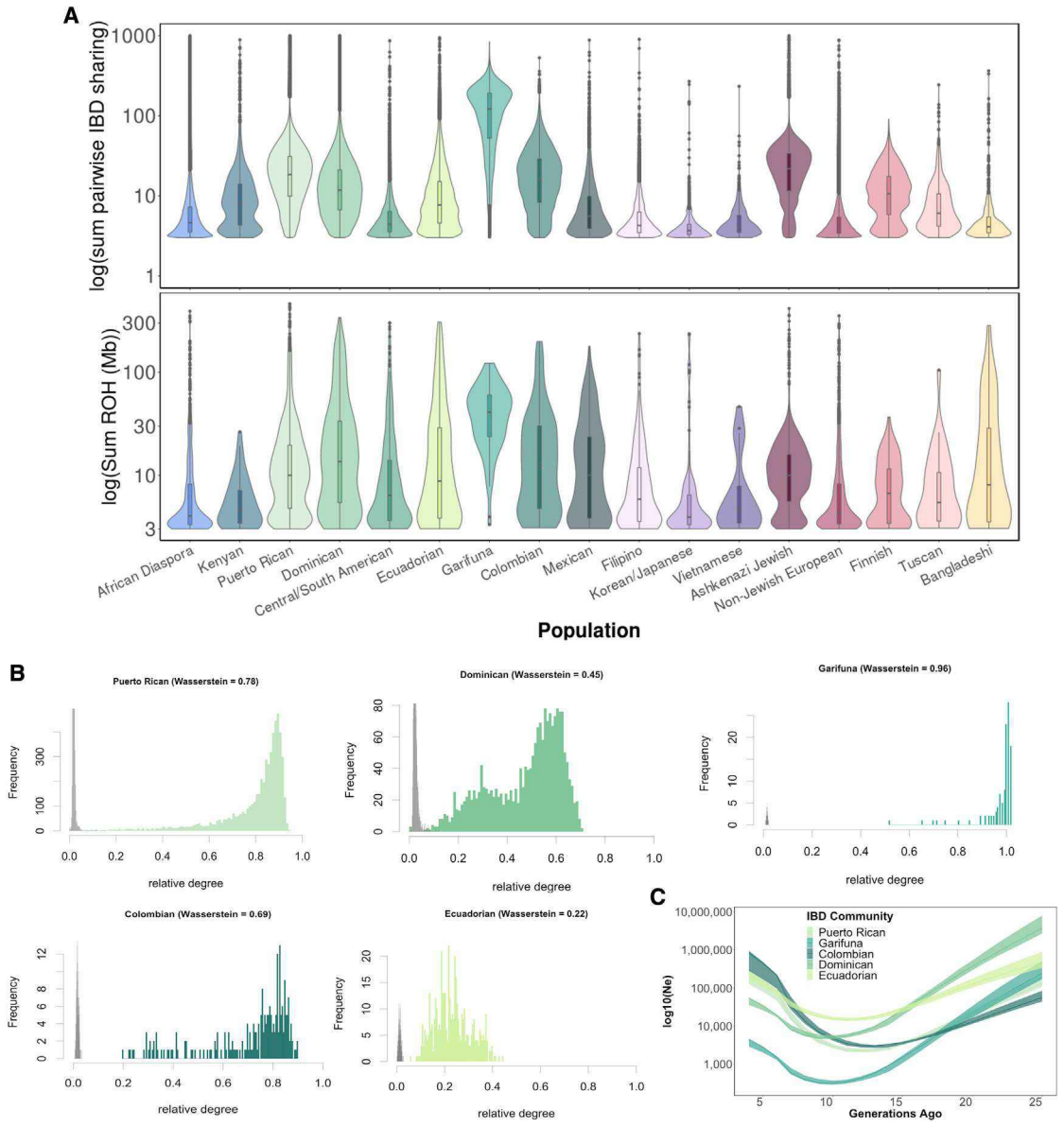


Figure 3. Signatures of founder effects in Bio*Me* communities

(A) (Top) Distribution of the mean sum of IBD sharing per inferred IBD community reveals the presence of elevated IBD sharing in several communities, including canonical founder populations such as the Ashkenazi Jewish and Finnish, and also in several Hispanic/Latino populations including the Puerto Rican, Dominican, and Garifuna. (Bottom) Distribution of the sum of runs of homozygosity (ROH) present within individuals per community. The violin and boxplots each represent the minima, maxima, and interquartile ranges of each distribution.

(B) Analysis of the distribution of degree sharing within versus between communities exhibiting elevated IBD revealed high levels of modularity, further suggestive of a founder effect.

(C) Using the tract-length distribution of IBD haplotypes to model the effective population size of communities over antecedent generations revealed evidence of population bottlenecks between 10 and 15 generations ago. Ribbons around the lines represent the 95% confidence intervals.

First, we extracted International Classification of Diseases 9 (ICD-9) (2007–2015) and ICD-10 (2015–present) billing codes from the EHR. We then applied an aggregation schema to convert ICD codes into 1,764 distinct diseases and traits called phecodes (Wu et al., 2019). We systematically performed logistic regression across all phecodes, where the membership of a given community was used as the primary predictor variable, adjusting for age and sex as covariates, and restricting analyses to communities containing $\geq$500 Bio*Me* participants (N = 7) in total. In all, 1,177 of 4,988 phecode associations tested were either significantly enriched or depleted across all seven of the top communities after Bonferroni correction. Summary statistics

for phecode associations per each of the largest seven communities are reported in Table S4.

We observed patterns of phecode enrichment across communities of shared continental ancestry. For example, three associations were significantly enriched in all three communities with appreciable African genetic ancestry (>20%), namely the PR, Dominican, and African diaspora communities (Table S4): essential hypertension, peripheral vascular disease, and type 2 diabetes (T2D). Higher prevalence of these diseases within AA populations have been widely reported in the epidemiological literature, but our findings and emerging evidence support these diseases as also being prevalent in some HL communities (Allison et al., 2015; Aguayo-Mazzucato et al., 2019; Campos and Rodriguez, 2019). Of the significantly associated phecodes, 21.7% (n = 274) were uniquely enriched in a single community, many of which had been previously reported to be at increased prevalence within those communities. For example, the phecode for asthma was observed to be most highly enriched in the PR ancestry (Oh et al., 2016) community (Table S4) (odds ratio [OR] = 2.91 [95% CI = 2.70–3.14]; $p < 1.13 \times 10^{-169}$). Furthermore, we observe elevated levels of ulcerative colitis (Bernstein et al., 2006) (OR = 2.61 [95% CI = 1.99–3.42]; $p < 2.90 \times 10^{-12}$) and Parkinson's disease (OR = 2.31 [95% CI = 1.78–3.00]; $p < 4.31 \times 10^{-10}$) in the AJ community (Figure 4A). We observe significantly elevated rates of sickle cell anemia (Makani et al., 2007) (OR = 6.92 [95% CI = 4.86–9.86]; $p < 8.20 \times 10^{-27}$) in the African diaspora community. Within the non-Jewish EA (Table S4) community, we observe elevated levels of multiple sclerosis (OR = 2.55 [95% CI = 2.01–3.237]; $p < 1.33 \times 10^{-14}$) and basal cell carcinoma (OR = 3.24 [95% CI = 2.50–4.20]; $p < 7.85 \times 10^{-19}$). Finally, in the Filipino community, both viral hepatitis B (Wong et al., 2013) (OR = 6.60 [95% CI = 5.01–8.69]; $p < 3.9 \times 10^{-41}$) and gout (Prasad and Krishnan, 2014) (OR = 2.94 [95% CI = 1.99–4.35]; $p < 6.55 \times 10^{-8}$) were at significantly higher prevalence (Table S4).

Among the community-specific phecodes, we also observe an enrichment of phecodes, suggesting increased prevalence of diseases within communities, that had not been previously reported. We observed a cluster of significantly enriched circulatory system (CS)-related phecodes, where 82% (9/11) were significantly enriched in the Dominican community. For example atherosclerosis of native arteries of the extremities (OR = 4.06 [95% CI = 3.13–5.28]; $p < 9.83 \times 10^{-26}$) and coronary atherosclerosis (OR = 1.64 [95% CI = 1.43–1.88]; $p < 1.61 \times 10^{-12}$) were the top two most significantly enriched codes (Table S4), suggestive of underlying increased prevalence of peripheral artery disease (PAD) in this population. Likewise, we observed a cluster of significantly enriched endocrine- and metabolic-related phecodes, with 28% (9/32) significantly enriched in AJs, among which the top phecodes were chronic lymphocytic thyroiditis (OR = 4.16 [95% CI = 3.62–4.78]; $p < 2.31 \times 10^{-89}$) and hypothyroidism (OR = 1.65 [95% CI = 1.49–1.83]; $p < 3.00 \times 10^{-21}$), suggesting an increased prevalence of autoimmune thyroid disease in this population. Taken together, this analysis suggests that identification of fine-scale genetic communities can help to elucidate the presence of population health disparities within healthcare systems.

Improving estimates of prevalence of variants underlying Mendelian conditions

To understand how fine-scale IBD-communities impact our understanding of the prevalence of Mendelian conditions, we generated a curated dataset of founder variants previously reported in the literature. We focused on founder variants reported in AJ and PR populations, as these were the two largest communities in Bio*Me* exhibiting evidence of founder effects. We initially identified a total of n = 82 AJ and n = 11 PR variants that were observed in one or more individuals in the N = 27,727 unrelated (<2nd degree) Bio*Me* participants for whom exome sequence data were available (Table S5). Via literature review, we were able to determine that for N = 32 of the AJ and N = 4 of the PR variants, founder effect status was supported by haplotypic evidence. We excluded four variants due to evidence of moderate-to-low penetrance, *APC*.c.3920T>A (Boursi et al., 2013; Liang et al., 2013), *CHEK2*.c.1283C>T (Shaag et al., 2005), *PSEN1*.c.617G>C (Arnold et al., 2013), and *GBA*.c.1226A>G (Zuckerman et al., 2007; Balwani et al., 2010). Of the remaining variants, 28/30 had a pathogenic or pathogenic/likely pathogenic assertion in ClinVar, meaning the assertions were supported by two or more submitters with no conflicts. Two variants, *CLRN1*.c.144T>G and *ABCC8*.c3989-9G>A, were listed with conflicting evidence of pathogenicity in ClinVar; however, upon further inspection were interpreted as pathogenic for the specific conditions under examination and therefore were included in downstream analysis. The final dataset contained N = 27 AJ and N = 3 PR founder variants, seven linked to two autosomal dominant (AD) disorders (hereditary breast and ovarian cancer and Lynch syndrome) and 23 linked to 19 autosomal recessive (AR) disorders. There were 1,099 (3.96%) variant-positive Bio*Me* participants (N = 986 that carried one variant and N = 113 that carried two variants), for an overall prevalence of 1 in 313 for AD and 1 in 27 for AR founder variants (Boursi et al., 2013; Liang et al., 2013).

To evaluate the population specificity and accuracy of prevalence estimates of Mendelian founder variants in Bio*Me*, we examined the participant-reported answers to survey questions compared with IBD community membership. We restricted analysis to the subset of N = 27,627 participants for whom we had complete survey information for both self-reported R/E and self-reported country of origin. For AJ founder variants, we examined answers to the question "What ethnic groups are part of your family background?" for variant-positive Bio*Me* participants (Table S6). Of the 946 participants who harbored an AJ founder variant, 182 (19.2%) only self-reported as Jewish, 471 (49.8%) only as Caucasian/white, 168 (17.8%) as both, and 125 (13.2%) selected a different answer. By contrast, 919 (97.1%) were members of the AJ IBD community. We repeated this analysis for the three PR founder variants. Of the 145 participants who harbored a PR founder variant, 135 (93.1%) only self-reported as HL, 0 (0%) only self-reported as Caucasian/white, 1 (0.7%) as both, and 9 (6.2%) selected a different answer. When examining prevalences based on answers to the survey question "Where were you born?", we observed that 63 (43.4%) self-reported being born in PR, 78 (53.8%) in the United States, and 4 (2.8%) selected a different answer. By contrast, 141 (97.2%) were members of the PR IBD community. We then used IBD communities to calculate per-variant prevalence, yielding rates

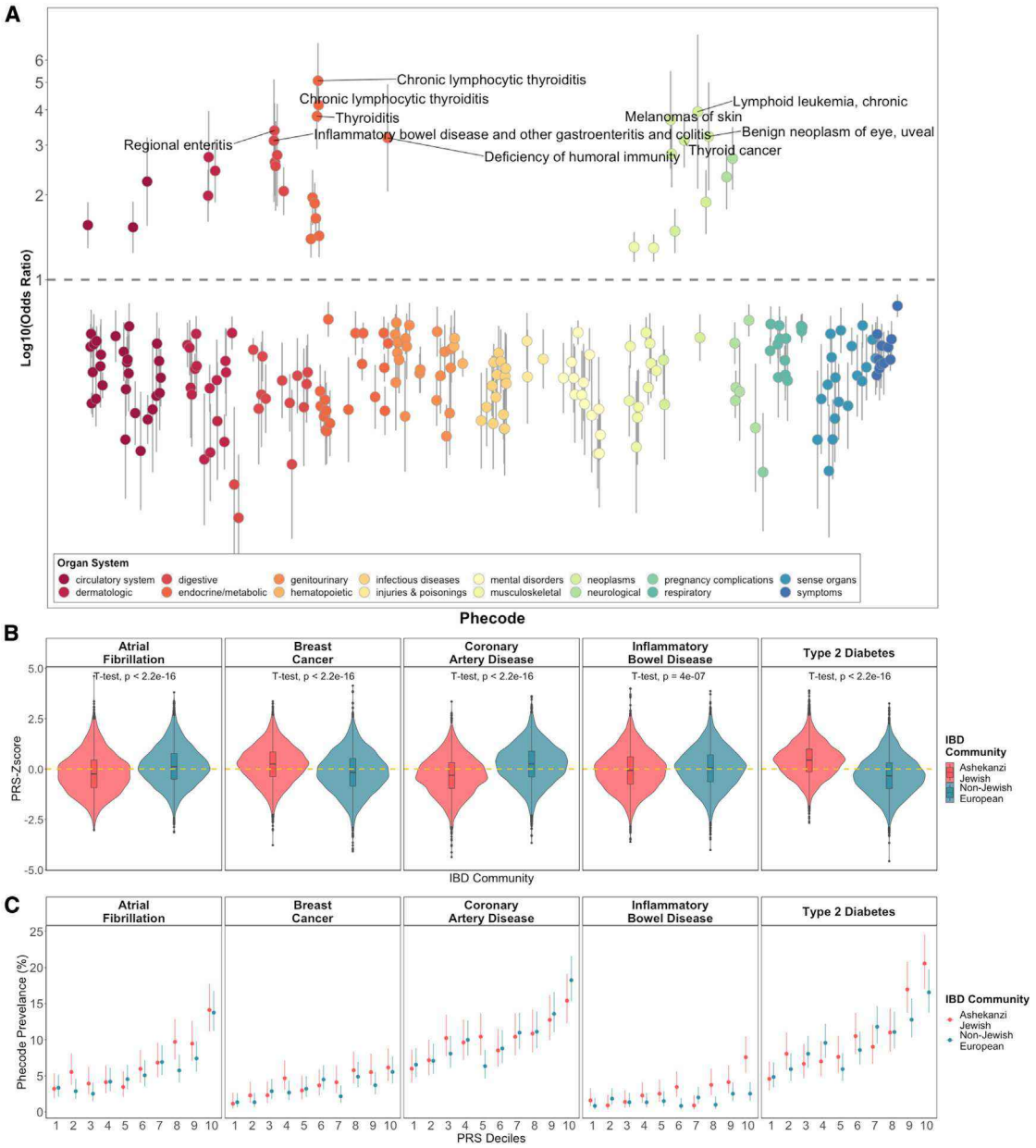


Figure 4. Exploring the impact of fine-scale IBD communities on complex disease

(A) Odds ratios (ORs) for the phecodes identified as being significantly higher or lower prevalence among the EHR of participants in the AJ IBD community (N = 4,409) colored by organ system. Lines represent the 95% confidence interval of the OR. Only phecodes for which the association met the Bonferroni threshold for statistical significance are displayed. Annotations are provided for phecodes with the highest ORs.

(B) *Z*-score-normalized distributions for 5 phenotypic traits in the AJ and non-AJ EA fine-scale IBD communities. The yellow dashed line represents the mean of the distribution prior to stratification based on community membership. Boxplots and violin plots represent the interquartile range, minima, and maxima of each distribution.

(C) Prevalence of disease-specific case status, stratified by PRS decile for the AJ and non-Jewish EA IBD communities. Bars represent 95% confidence intervals.

ranging from 1:108 to 1:4,102 for AD and from 1:23 to 1:820 for AR variants (Table 1). This analysis demonstrates that by linking disease variants to IBD communities and modeling the genetic estimates of the specific founder events that give rise to disease variants, we can improve our understanding of the population specificity and accuracy of prevalence rates.

Exploring the impact of fine-scale IBD communities and complex disease

Next, we investigated the impact of fine-scale population structure, as captured by IBD communities, on complex disease. We focused on PRSs, which predict complex disease risk by aggregating the contributions of many variants associated with

Table 1. IBD communities refine estimates of prevalence of variants underlying Mendelian conditions

Gene	Mode of inheritance	Physical position (GRCh38)	c.DNA position	Protein modification	Self-reported ethnic group	Self-reported country of birth	IBD community	Condition	IBD community prevelance
AJ founder variants (N = 27)									
MSH2	AD	2: 47475171	c.1906G>C	p.Ala636Pro	C(1); O(1)	US(2)	AJ(2)	Lynch syndrome	1 in 2,051
MSH6	AD	2: 47806630-47806631	c.3984_3987dupGTCA	p.Leu1330Valfs	B(1)	US(1)	AJ(1)	Lynch syndrome	1 in 4,102
MSH6	AD	2: 47806606-47806609	c.3959_3962delCAAG	p.Ala1190fs	C(1)	US(1)	AJ(1)	Lynch syndrome	1 in 4,102
BRCA1	AD	17: 43124028-43124029	c.68_69delAG	p.Glu23fs	AJ(12); C(13); B(9); O(4)	US(33); O(5)	AJ(38)	hereditary breast and ovarian cancer	1 in 108
BRCA1	AD	17: 43057062-43057063	c.5266dupC	p.Gln1756Profs	AJ(1); C(4); B(1)	US(5); O(1)	AJ(5); O(1)	hereditary breast and ovarian cancer	1 in 820
BRCA2	AD	13: 32340301	c.5946delT	p.Ser1982fs	AJ(4); C(17); B(8); O(5)	US(31); O(3)	AJ(34); O(1)	hereditary breast and ovarian cancer	1 in 124
F11	AR	4: 186274193	c.403G>T	p.Glu135Ter	AJ(15); C(66); B(26); O(20)	US(111); O(16)	AJ(123); O(4)	factor XI deficiency	1 in 33
F11	AR	4: 186280258	c.901T>C	p.Phe301Leu	AJ(40); C(84); B(35); O(28)	US(159); O(28)	AJ(182); O(5)	factor XI deficiency	1 in 23
MTTP	AR	4: 99622756	c.2593G>T	p.Gly865Ter	AJ(2); C(7); B(3); O(2)	US(12); O(2)	AJ(14)	abetalipoproteinemia	1 in 293
BLM	AR	15: 90766923-90766928	c.2207_2212delATCTGAinsTAGATTC	p.Tyr736fs	AJ(1); C(11); B(1); O(4)	US(13); O(4)	AJ(16); O(1)	Bloom syndrome	1 in 256
MPL	AR	1: 43337929	c.79+2T>A	–	AJ(5); C(20); B(11); O(5)	US(33); O(8)	AJ(41); O(1)	congenital amegakaryocytic thrombocytopenia	1 in 103
RTEL1	AR	20: 63695619	c.3791G>A	p.Arg1264His	AJ(8); C(15); B(4); O(3)	US(27); O(3)	AJ(29); O(1)	dyskeratosis congenita	1 in 141
ELP1	AR	9: 108899816	c.2204+6T>C (IVS20+6T>C)	–	AJ(16); C(42); B(20); O(14)	US(75); O(17)	AJ(88); O(4)	familial dysautonomia	1 in 47
ABCC8	AR	11: 17397055	c.3989-9G>A	–	AJ(10); C(23); B(9); O(5)	US(41); O(6)	AJ(49)	familial hyperinsulinism	1 in 87
ABCC8	AR	11: 17395888-17395890	c.4160_4162delTCT	p.Phe1387del	AJ(1); C(3); B(1)	US(5); O(0)	AJ(5)	familial hyperinsulinism	1 in 820
FANCC	AR	9: 95172033	c.456+4A>T	–	AJ(4); C(24); B(4); O(5)	US(33); O(4)	AJ(37)	Fanconi anemia	1 in 111
GBA	AR	1: 155240660-155240661	c.84dupG	p.Leu29Alafs*18	AJ(4); C(4); B(1); O(2)	US(10); O(1)	AJ(10); O(1)	Gaucher disease	1 in 410
TMEM216	AR	11: 61393965	c.218G>T	p.Arg73Leu	AJ(11); C(15); B(6); O(5)	US(35); O(2)	AJ(37)	Joubert syndrome	1 in 111
BCKDHB	AR	6: 80168945	c.548G>C	p.Arg183Pro	AJ(4); C(17); B(3); O(5)	US(23); O(6)	AJ(29)	maple syrup urine disease	1 in 141
MCOLN1	AR	19: 7526759	c.406-2A>G	–	AJ(6); C(13); B(1); O(3)	US(20); O(3)	AJ(23)	mucolipidosis IV	1 in 178
GJB2	AR	13: 20189415	c.167delT	p.Leu56fs	AJ(32); C(74); B(19); O(15)	US(117); O(23)	AJ(133); O(7)	nonsyndromic hearing loss	1 in 31
TCIRG1	AR	11: 68041392	c.117+4A>T	–	AJ(7); C(4); B(4); O(2)	US(14); O(3)	AJ(15); O(2)	osteopetrosis	1 in 273

(*Continued on next page*)

Table 1. *Continued*

Gene	Mode of inheritance	Physical position (GRCh38)	c.DNA position	Protein modification	Self-reported ethnic group	Self-reported country of birth	IBD community	Condition	IBD community prevelance
DNAI2	AR	17: 74309345	c.1304G>A	p.Trp435Ter	AJ(5); C(9); B(4); O(1)	US(17); O(2)	AJ(19)	primary ciliary dyskinesia	1 in 216
CFAP298	AR	21: 32602299	c.735C>G	p.Tyr245Ter	AJ(5); C(6); B(1); O(2)	US(12); O(2)	AJ(15)	primary ciliary dyskinesia	1 in 293
DHDDS	AR	1: 26438228	c.124A>G	p.Lys42Glu	AJ(4); C(18); B(7); O(6)	US(31); O(4)	AJ(35)	retinitis pigmentosa, nonsyndromic	1 in 117
PCDH15	AR	10: 54317414	c.733C>T	p.Arg245Ter	AJ(5); C(10); B(1); O(1)	US(15); O(2)	AJ(17)	Usher syndrome type 1	1 in 241
CLRN1	AR	3: 150972565	c.144T>G	p.Asn48Lys	AJ(8); C(29); B(5); O(3)	US(41); O(4)	AJ(44); O(1)	Usher syndrome type 3A	1 in 93
PR founder variants (N = 3)									
BRCA2	AD	13: 32338277	c.3922G>T	p.Glu1308Ter	HL(7)	US(3); PR(4)	PR(7)	hereditary breast cancer	1 in 729
COL27A1	AR	9: 114195977	c.2089G>C	p.Gly697Arg	HL(82); B(1); O(4)	US(47); PR(38); O(2)	PR(87); O(2)	Steel syndrome	1 in 60
SGCG	AR	13: 23324452	c.787G>A	p.Glu263Lys	HL(49); O(6)	US(31); PR(22); O(2)	PR(54); O(2)	limb girdle dystrophy	1 in 96

disease across an individual's genome into a single risk score. We selected previously validated and published PRS for five common diseases with major public health impact: cardiovascular disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer (Khera et al., 2018). Since these PRSs were derived and validated in individuals of primarily European ancestry, we restricted our analyses to European ancestry participants in Bio*Me*, to avoid confounding due to genetic ancestry (Martin et al., 2017; De La Vega and Bustamante, 2018; Martin et al., 2019). This allowed us to calculate PRS distributions for each disease in European ancestry participants (N = 10,273) and normalize the distributions as Z scores (mean = 0, standard deviation = 1). We then stratified the participants into members of the AJ IBD community (N = 4,337) and non-Jewish EA IBD communities (N = 5,936) and re-examined the normalized distributions. We observed that the distribution of the PRS between communities were significantly different for all five diseases examined (Figure 4B; two-sided t test), with the largest difference observed for T2D (Δmean = 0.77, $p < 1 \times 100^{-300}$) and the smallest for inflammatory bowel disease (Δmean = 0.11, $p < 4.0 \times 10^{-7}$). Notably, we saw a significantly elevated prevalence of inflammatory bowel disease within the topmost decile in the AJ IBD community, relative to the non-Jewish EA (Figure 4C; chi-square test $p < 0.0006$). Furthermore, we observed significant differences in predictive power for the PRS for inflammatory bowel disease between the non-Jewish EA and AJ IBD communities, respectively (area under the curve [AUC] = 56.56% [95% CI = 50.43–62.69], AUC = 65.57% [95% CI = 60.50–70.63]; $p < 0.013$). These observed differences across communities mirror previously observed differences in PRS performance across continental ancestry groups attributable to the impact of genetic drift in the calculation of PRS, rather than differences in genetic contributions (Martin et al., 2019). This highlights the growing importance of understanding fine-scale population structure in PRS models.

DISCUSSION

Here, we demonstrate how genetic ancestry may be used for fine-scale population health monitoring in medicine. We examine the intersection of R/E, captured both as EHR and self-reported, and genetic ancestry in a large urban health system and show how EHR reporting imperfectly captures R/E compared with self-reporting, while genetic ancestry reveals additional complexity in population structure compared with self-reporting. We apply a framework to detect fine-scale population structure by characterizing a network of distant relatedness within patients in the Bio*Me* biobank and show 17 distinct IBD communities that are highly correlated with culturally endogamous groups and recent diaspora to NYC from countries around the world. By linking to EHRs and testing for enrichment of ICD-9/10-based health outcomes within uncovered communities, we demonstrate a significant community-specific enrichment of both anticipated and novel health-related traits. This suggests that IBD communities could be used to identify patient populations at elevated risk for diseases that may not be otherwise captured via available population labels. Furthermore, we demonstrate significant differences in the distribution and accuracy of PRS for common

disease between two IBD communities of shared continental ancestry, suggesting that fine-scale genetic structure elucidated by these community definitions may also have important implications for improving genomic risk prediction for common diseases.

This work also elucidates the extent and complexity of founder populations in NYC. We show that many of the IBD communities exhibit evidence of founder effects, as demonstrated by elevated levels of autozygosity, median within-community IBD sharing, and application of methods that measure the degree of distance between communities in the network topology. This approach resulted in more accurate prevalence estimates for founder variants linked to genetic disorders compared with those derived using populations defined by self-reported R/E or geographical region of origin. This suggests that IBD communities could be used to identify patient populations at elevated risk for genetic disorders that may not be otherwise captured via available population labels. We identified both canonical founder populations with known historical evidence of founder events, including AJ (Atzmon et al., 2010) and Finnish (Martin et al., 2018) populations, as well as less well-characterized founder populations in NYC, namely populations of Garifuna (Atzmon et al., 2010; Herrera-Paz et al., 2010), PR (Belbin et al., 2017), Colombian (Carvajal-Carmona et al., 2003; Mooney et al., 2018), and Ecuadorian and Dominican (Browning et al., 2018) descents. Overall approximately one-quarter of Bio*Me* participants harbor genetic signatures of founder effects, and extrapolating this observation to the demographics of NYC, we estimate that approximately 15% of New Yorkers could be genetically linked to one or more founder populations. This finding mirrors similar observations of founder effects in large, predominantly European ancestry biobanks in Finland (Martin et al., 2018) and rural Pennsylvania (Staples et al., 2018), where founder effects were found to be ubiquitous in the former and in approximately one-fifth of the latter cohort. A recent study of a direct-to-consumer genetic database of approximately 770,000 customers across the United States also revealed myriad signatures of founder effects that could be attributed to pre-diaspora population structure and/or post-diaspora isolation, i.e., multiple Irish ancestry groups in Boston (Han et al., 2017). This suggests that as more massive-scale population-based biobanks emerge, evidence will increasingly show that founder effects and founder populations may be more common, and their origins more complex, than previously thought.

A better understanding of population structure may also impact the use of genomic information to inform patient care of genetically driven disease, such as cancer detection and treatment (Deng and Nakamura, 2017), pre- and perinatal testing (Peters et al., 2015; Hui and Bianchi, 2017), and new applications of routine genomic screening for preventive health (Trivedi, 2017). For example, in recent work in the same populations, we showed that individuals with AJ founder variants in *BRCA1* and *BRCA2* genes were twice as likely to have undergone clinical genetic testing compared with other groups with founder variants in these genes, despite similar rates of cancer (Abul-Husn et al., 2019). A lack of patient and/or provider awareness about population-specific risks may impact rates of genetic testing (Williams et al., 2019). For common diseases, which are largely influenced by non-genetic factors, but nevertheless may also have a substantial genetic component, this is particularly true. Understanding complex patterns of distant relatedness at a population level can provide simultaneous insights into both genetic and environmental factors underlying disease. For example, the observed elevation of risk for asthma in PR populations can be linked to a host of environmental and socioeconomic contributing factors differentially impacting PRs compared with other Hispanic/Latinx groups, while PR-specific genetic determinants have yet to be identified (Szentpetery et al., 2016). This demonstrates how genomics can be an extremely useful tool to help uncover non-genetic disease factors contributing to common disease and also cautions against oversimplified assumptions that one population group is more genetically predisposed to disease versus another.

Our approach has a number of limitations that we highlight here. The fine-scale population structure in the Bio*Me* biobank uncovered by our approach is often strongly correlated with certain population labels and demographic information captured by our survey instruments. However, as no ground truth information exists by which to evaluate the accuracy and efficacy of community assignments, the community labels must be considered supported by evidence rather than definitive. In some cases, we demonstrated weak evidence of correlation with collected population labels, which was challenging to interpret, most notably for the putative Garifuna community. Evidence to support the community assignment was limited to the observations of (1) tight clustering of individuals in PCA space, indicating uniform proportions of African and Amerindigenous genetic ancestry in community members, which are each indicative of a founder population; and (2) evidence of a very tight bottleneck using IBD tract-length modeling, the timing of which is coincident with the known historical founder events in the Garifuna population. However, we demonstrate that even in the absence of conclusive evidence to ascribed population labels, consistent and well-defined genetic communities can be clinically useful. Finally, the success of defining IBD-based communities is contingent on existing patterns of assortative mating in a given population dataset, which is influenced by cultural endogamy, patterns of migration, geographical proximity, and other demographic forces. In addition, a person could have ancestry from two or more communities or self-reported groups (6.5% of participants in this study). In these instances, any one particular community assignment may be sub-optimal for precision medicine applications and indicates future directions to develop methods that can model population structure jointly as continuous and discrete processes.

This work contributes to the ongoing conversation about the role of R/E in medicine. Some have argued that R/E is not biologically meaningful (King and Owens, 2001; Cooper et al., 2003) and have demonstrated potential harms connected to the use of race in patient care (Vyas et al., 2020). Others contend that R/E categories are correlated with underlying genetic and/or socioeconomic factors impacting disease and that capturing R/E information is biologically and clinically useful (Burchard et al., 2003). Here, we demonstrate that embedding genomic data in health systems, and using it to infer genetic ancestry, will allow the development of evidence-based means to utilize R/E, genetic ancestry, and the socioeconomic determinants of

health for both rare and common diseases. Finally, the network-based machine learning approaches applied here are highly scalable to very large datasets of individuals. As more genomic data become available in health systems globally (Stark et al., 2019), and via large research projects (Bycroft et al., 2018; All of Us Research Program Investigators et al., 2019), we anticipate approaches such as ours will uncover increasing nuances in the population structure. Furthermore, as EHRs evolve and increase in resolution of longitudinal phenome data (Abul-Husn and Kenny, 2019), they will offer enhanced opportunities to monitor health in real time, allowing for agile programs of discovery, prediction, and intervention.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Study population
- METHOD DETAILS
  - Ascertainment of Race/Ethnicity Information in BioMe
  - Genotype Quality Control
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Comparison of Electronic Health Record versus self-reported Race/Ethnicity
  - Collapsing of Self-Reported Race/Ethnicity Questionnaire Data to explore the relationship with genetic ancestry
  - Genetic Relatedness Estimation
  - Global Ancestry Estimation
  - Phasing and Identity-by-Descent Inference
  - Network Construction and Community Detection
  - IBD Network Analysis
  - Analysis of IBD Community Membership Using Population Labels
  - Inference of Runs of Homozygosity
  - Phenotype ontology
  - Analysis of Enrichment of Phecodes within IBD-Communities
  - Genotype Imputation and Polygenic Risk Score Estimation
  - Annotation of Clinically Relevant Founder Variation from Exome Data

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2021.03.034.

ACKNOWLEDGMENTS

Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award numbers S10OD018522 and S10OD026880. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

Conceptualization, G.M.B., C.R.G., N.A.Z., A.A., and E.E.K.; methodology, G.M.B., R.S., C.R.G., J.A., N.D.B., A.C., and E.E.K.; formal analysis, G.M.B., S.C., S.W., E.S., D.T., and B.S.G.; resources, E.E.K., J.H.C., E.P.B., R.J.F.L., and Regeneron Genetics Center; data curation, G.M.B., E.S., S.E., and A.M.; writing – original draft, G.M.B., N.A.Z., C.R.G., and E.E.K.; review and editing, G.M.B., B.S.G., N.D.B., E.S., N.A.H., E.P.S., N.A.Z., C.R.G., and E.E.K.; visualization, G.M.B., S.C., S.W., and D.T.; funding acquisition, E.E.K.

DECLARATION OF INTERESTS

N.S.A.-H. was previously employed by Regeneron Pharmaceuticals and has received a speaker honorarium from Genentech. E.E.K. has received speaker honoraria from Illumina and Regeneron Pharmaceuticals. S.W. is currently employed at Tempus. A.M. is currently employed at Regeneron. E.P.S. is currently employed at Calico. D.S.P. is currently employed at Ancestry. A.A. is currently employed at 23&Me. The remaining authors declare no competing interests.

Received: September 23, 2019
Revised: November 18, 2020
Accepted: March 12, 2021
Published: April 15, 2021

REFERENCES

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Abul-Husn, N.S., and Kenny, E.E. (2019). Personalized Medicine and the Power of Electronic Health Records. Cell *177*, 58–69.

Abul-Husn, N.S., Soper, E.R., Odgis, J.A., Cullina, S., Bobo, D., Moscati, A., Rodriguez, J.E., Loos, R.J.F., Cho, J.H., Belbin, G.M., et al. (2019). Exome sequencing reveals a high prevalence of BRCA1 and BRCA2 founder variants in a diverse population-based biobank. Genome Med. *12*, 2.

Aguayo-Mazzucato, C., Diaque, P., Hernandez, S., Rosas, S., Kostic, A., and Caballero, A.E. (2019). Understanding the growing epidemic of type 2 diabetes in the Hispanic population living in the United States. Diabetes Metab. Res. Rev. *35*, e3097.

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

Allison, M.A., Gonzalez, F., 2nd, Raij, L., Kaplan, R., Ostfeld, R.J., Pattany, M.S., Heiss, G., and Criqui, M.H. (2015). Cuban Americans have the highest rates of peripheral arterial disease in diverse Hispanic/Latino communities. J. Vasc. Surg. *62*, 665–672.

All of Us Research Program Investigators, Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., and Dishman, E. (2019). The "All of Us" Research Program. N. Engl. J. Med. *381*, 668–676.

Amendola, L.M., Berg, J.S., Horowitz, C.R., Angelo, F., Bensen, J.T., Biesecker, B.B., Biesecker, L.G., Cooper, G.M., East, K., Filipski, K., et al. (2018). The Clinical Sequencing Evidence-Generating Research Consortium: Integrating Genomic Sequencing in Diverse and Medically Underserved Populations. Am. J. Hum. Genet. *103*, 319–327.

Arnold, S.E., Vega, I.E., Karlawish, J.H., Wolk, D.A., Nunez, J., Negron, M., Xie, S.X., Wang, L.S., Dubroff, J.G., McCarty-Wood, E., et al. (2013). Frequency and clinicopathological characteristics of presenilin 1 Gly206Ala mutation in Puerto Rican Hispanics with dementia. J. Alzheimers Dis. *33*, 1089–1095.

Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P.F., Morrow, B., Friedman, E., Oddoux, C., Burns, E., et al. (2010). Abraham's children in the

genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. Am. J. Hum. Genet. *86*, 850–859.

Balwani, M., Fuerstman, L., Kornreich, R., Edelmann, L., and Desnick, R.J. (2010). Type 1 Gaucher disease: significant disease manifestations in "asymptomatic" homozygotes. Arch. Intern. Med. *170*, 1463–1469.

Banda, Y., Kvale, M.N., Hoffmann, T.J., Hesselson, S.E., Ranatunga, D., Tang, H., Sabatti, C., Croen, L.A., Dispensa, B.P., Henderson, M., et al. (2015). Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics *200*, 1285–1295.

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.

Bernstein, C.N., Rawsthorne, P., Cheang, M., and Blanchard, J.F. (2006). A population-based case control study of potential risk factors for IBD. Am. J. Gastroenterol. *101*, 993–1002.

Boursi, B., Sella, T., Liberman, E., Shapira, S., David, M., Kazanov, D., Arber, N., and Kraus, S. (2013). The APC p.I1307K polymorphism is a significant risk factor for CRC in average risk Ashkenazi Jews. Eur. J. Cancer *49*, 3680–3685.

Browning, S.R., and Browning, B.L. (2012). Identity by descent between distant relatives: detection and applications. Annu. Rev. Genet. *46*, 617–633.

Browning, S.R., and Browning, B.L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am. J. Hum. Genet. *97*, 404–418.

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.

Burchard, E.G., Ziv, E., Coyle, N., Gomez, S.L., Tang, H., Karter, A.J., Mountain, J.L., Pérez-Stable, E.J., Sheppard, D., and Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. N. Engl. J. Med. *348*, 1170–1175.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Campos, C.L., and Rodriguez, C.J. (2019). High blood pressure in Hispanics in the United States: a review. Curr. Opin. Cardiol. *34*, 350–358.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.

Carter-Pokras, O.D., and Gergen, P.J. (1993). Reported asthma among Puerto Rican, Mexican-American, and Cuban children, 1982 through 1984. Am. J. Public Health *83*, 580–582.

Carvajal-Carmona, L.G., Ophoff, R., Service, S., Hartiala, J., Molina, J., Leon, P., Ospina, J., Bedoya, G., Freimer, N., and Ruiz-Linares, A. (2003). Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. Hum. Genet. *112*, 534–541.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Collins, R. (2012). What makes UK Biobank special? Lancet *379*, 1173–1174.

Cooper, R.S., Kaufman, J.S., and Ward, R. (2003). Race and genomics. N. Engl. J. Med. *348*, 1166–1170.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

De La Vega, F.M., and Bustamante, C.D. (2018). Polygenic risk scores: a biased prediction? Genome Med. *10*, 100.

Deng, X., and Nakamura, Y. (2017). Cancer Precision Medicine: From Cancer Screening to Drug Selection and Personalized Immunotherapy. Trends Pharmacol. Sci. *38*, 15–24.

Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205–1210.

Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C.V., Staples, J., Gonzaga-Jauregui, C., et al. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science *354*, aaf6814.

Durrett, R. (2006). Erdös–Rényi Random Graphs. In Random Graph Dynamics (Cambridge University Press), pp. 27–69.

Ekrikpo, U.E., Mnika, K., Effa, E.E., Ajayi, S.O., Okwuonu, C., Waziri, B., Bello, A., Dandara, C., Kengne, A.P., Wonkam, A., and Okpechi, I. (2020). Association of Genetic Polymorphisms of TGF-β1, HMOX1, and APOL1 With CKD in Nigerian Patients With and Without HIV. Am. J. Kidney Dis. *76*, 100–108.

Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326.

Gutiérrez, O.M., Irvin, M.R., Zakai, N.A., Naik, R.P., Chaudhary, N.S., Estrella, M.M., Limou, S., Judd, S.E., Cushman, M., Kopp, J.B., and Winkler, C.A. (2020). APOL1 Nephropathy Risk Alleles and Mortality in African American Adults: A Cohort Study. Am. J. Kidney Dis. *75*, 54–60.

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.

Hassan, M.O., Duarte, R., Dickens, C., Dix-Peek, T., Naidoo, S., Vachiat, A., Grinter, S., Manga, P., and Naicker, S. (2020). APOL1 Genetic Variants Are Associated with Serum-Oxidized Low-Density Lipoprotein Levels and Subclinical Atherosclerosis in South African CKD Patients. Nephron *144*, 331–340.

Herrera-Paz, E.-F., Matamoros, M., and Carracedo, A. (2010). The Garífuna (Black Carib) people of the Atlantic coasts of Honduras: Population dynamics, structure, and phylogenetic relations inferred from genetic data, migration matrices, and isonymy. Am. J. Hum. Biol. *22*, 36–44.

Homa, D.M., Mannino, D.M., and Lara, M. (2000). Asthma mortality in U.S. Hispanics of Mexican, Puerto Rican, and Cuban heritage, 1990-1995. Am. J. Respir. Crit. Care Med. *161*, 504–509.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

Hui, L., and Bianchi, D.W. (2017). Noninvasive Prenatal DNA Testing: The Vanguard of Genomic Medicine. Annu. Rev. Med. *68*, 459–472.

Kao, W.H.L., Klag, M.J., Meoni, L.A., Reich, D., Berthier-Schaad, Y., Li, M., Coresh, J., Patterson, N., Tandon, A., Powe, N.R., et al. (2008). MYH9 is associated with nondiabetic end-stage renal disease in African Americans. Nat. Genet. *40*, 1185–1192.

Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.

King, M.C., and Owens, K. (2001). Genomic views of human history. Pathol. Biol. (Paris) *286*, 384–387.

Klinger, E.V., Carlini, S.V., Gonzalez, I., Hubert, S.S., Linder, J.A., Rigotti, N.A., Kontos, E.Z., Park, E.R., Marinacci, L.X., and Haas, J.S. (2015). Accuracy of race, ethnicity, and language preference in an electronic health record. J. Gen. Intern. Med. *30*, 719–723.

Kramer, H.J., Stilp, A.M., Laurie, C.C., Reiner, A.P., Lash, J., Daviglus, M.L., Rosas, S.E., Ricardo, A.C., Tayo, B.O., Flessner, M.F., et al. (2017). African Ancestry-Specific Alleles and Kidney Disease Risk in Hispanics/Latinos. J. Am. Soc. Nephrol. *28*, 915–922.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1–26.

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

Liang, J., Lin, C., Hu, F., Wang, F., Zhu, L., Yao, X., Wang, Y., and Zhao, Y. (2013). APC polymorphisms and the risk of colorectal neoplasia: a HuGE review and meta-analysis. Am. J. Epidemiol. *177*, 1169–1179.

Makani, J., Williams, T.N., and Marsh, K. (2007). Sickle cell disease in Africa: burden and research priorities. Ann. Trop. Med. Parasitol. *101*, 3–14.

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.

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.

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.

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.

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.

Mathias, R.A., Taub, M.A., Gignoux, C.R., Fu, W., Musharoff, S., O'Connor, T.D., Vergara, C., Torgerson, D.G., Pino-Yanes, M., Shringarpure, S.S., et al. (2016). A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat. Commun. *7*, 12522.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. https://arxiv.org/abs/1802.03426.

Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. Science *201*, 786–792.

Miller, A.K., Azhibekov, T., O'Toole, J.F., Sedor, J.R., Williams, S.M., Redline, R.W., and Bruggeman, L.A. (2020). Association of preeclampsia with infant APOL1 genotype in African Americans. BMC Med. Genet. *21*, 110.

Mooney, J.A., Huber, C.D., Service, S., Sul, J.H., Marsden, C.D., Zhang, Z., Sabatti, C., Ruiz-Linares, A., Bedoya, et al.; 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.

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.

Nadkarni, G.N., Gignoux, C.R., Sorokin, E.P., Daya, M., Rahman, R., Barnes, K.C., Wassel, C.L., and Kenny, E.E. (2018). Worldwide Frequencies of APOL1 Renal Risk Variants. N. Engl. J. Med. *379*, 2571–2572.

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.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics *32*, 1749–1751.

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.

Nqebelele, N.U., Dickens, C., Dix-Peek, T., Duarte, R., and Naicker, S. (2019). JC Virus and *APOL1* Risk Alleles in Black South Africans With Hypertension-Attributed CKD. Kidney Int. Rep. *4*, 939–945.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. *10*, e1004234.

Oh, S.S., White, M.J., Gignoux, C.R., and Burchard, E.G. (2016). Making Precision Medicine Socially Precise. Take a Deep Breath. Am. J. Respir. Crit. Care Med. *193*, 348–350.

Parsa, A., Kao, W.H., Xie, D., Astor, B.C., Li, M., Hsu, C.Y., Feldman, H.I., Parekh, R.S., Kusek, J.W., Greene, T.H., et al. (2013). APOL1 risk variants, race, and progression of chronic kidney disease. N. Engl. J. Med. *369*, 2183–2196.

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

Peters, D.G., Yatsenko, S.A., Surti, U., and Rajkovic, A. (2015). Recent advances of genomic testing in perinatal medicine. Semin. Perinatol. *39*, 44–54.

Prasad, P., and Krishnan, E. (2014). Filipino gout: a review. Arthritis Care Res. (Hoboken) *66*, 337–343.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

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

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., 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.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

Rosvall, M., and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA *105*, 1118–1123.

Rosvall, M., and Bergstrom, C.T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PLoS ONE *6*, e18209.

Shaag, A., Walsh, T., Renbaum, P., Kirchhoff, T., Nafa, K., Shiovitz, S., Mandell, J.B., Welcsh, P., Lee, M.K., Ellis, N., et al. (2005). Functional and genomic approaches reveal an ancient CHEK2 allele associated with breast cancer in the Ashkenazi Jewish population. Hum. Mol. Genet. *14*, 555–563.

Shemirani, R., Belbin, G.M., Avery, C.L., Kenny, E.E., Gignoux, C.R., and Ambite, J.L. (2019). Rapid detection of identity-by-descent tracts for mega-scale datasets. BioRxiv. https://doi.org/10.1101/749507.

Smith, N., Iyer, R.L., Langer-Gould, A., Getahun, D.T., Strickland, D., Jacobsen, S.J., Chen, W., Derose, S.F., and Koebnick, C. (2010). Health plan administrative records versus birth certificate records: quality of race and ethnicity information in children. BMC Health Serv. Res. *10*, 316.

Staples, J., Maxwell, E.K., Gosalia, N., Gonzaga-Jauregui, C., Snyder, C., Hawes, A., Penn, J., Ulloa, R., Bai, X., Lopez, A.E., et al. (2018). Profiling and Leveraging Relatedness in a Precision Medicine Cohort of 92,455 Exomes. Am. J. Hum. Genet. *102*, 874–889.

Stark, Z., Dolman, L., Manolio, T.A., Ozenberger, B., Hill, S.L., Caulfied, M.J., Levy, Y., Glazer, D., Wilson, J., Lawler, M., et al. (2019). Integrating Genomics into Healthcare: A Global Responsibility. Am. J. Hum. Genet. *104*, 13–20.

Szentpetery, S.E., Forno, E., Canino, G., and Celedón, J.C. (2016). Asthma in Puerto Ricans: Lessons from a high-risk population. J. Allergy Clin. Immunol. *138*, 1556–1558.

Tang, H., Peng, J., Wang, P., and Risch, N.J. (2005). Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. *28*, 289–301.

Thakoordeen-Reddy, S., Winkler, C., Moodley, J., David, V., Binns-Roemer, E., Ramsuran, V., and Naicker, T. (2020). Maternal variants within the apolipoprotein L1 gene are associated with preeclampsia in a South African cohort of African ancestry. Eur. J. Obstet. Gynecol. Reprod. Biol. *246*, 129–133.

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

Thompson, E.A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. Genetics *194*, 301–326.

Trivedi, B.P. (2017). Medicine's future? Science *358*, 436–440.

Umeukeje, E.M., and Young, B.A. (2019). Genetics and ESKD Disparities in African Americans. Am. J. Kidney Dis. *74*, 811–821.

Villani, C. (2009). The Wasserstein distances. In Optimal Transport. Grundlehren der mathematischen Wissenschaften (A Series of Comprehensive Studies in Mathematics), Volume 338 (Springer), pp. 93–111.

Vyas, D.A., Eisenstein, L.G., and Jones, D.S. (2020). Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. N. Engl. J. Med. *383*, 874–882.

Williams, C.D., Bullard, A.J., O'Leary, M., Thomas, R., Redding, T.S., 4th, and Goldstein, K. (2019). Racial/Ethnic Disparities in BRCA Counseling and Testing: a Narrative Review. J. Racial Ethn. Health Disparities *6*, 570–583.

Wong, S.N., Ong, J.P., Labio, M.E., Cabahug, O.T., Daez, M.L., Valdellon, E.V., Sollano, J.D., Jr., and Arguillas, M.O. (2013). Hepatitis B infection among adults in the philippines: A national seroprevalence study. World J. Hepatol. *5*, 214–219.

Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J.C., et al. (2019). Developing and Evaluating Mappings of ICD-10 and ICD-10-CM Codes to PheCodes. JMIR Med. Inform. *7*, e14325.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.

Zuckerman, S., Lahad, A., Shmueli, A., Zimran, A., Peleg, L., Orr-Urtreger, A., Levy-Lahad, E., and Sagi, M. (2007). Carrier screening for Gaucher disease: lessons for low-penetrance, treatable diseases. JAMA *298*, 1281–1290.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Bio*Me* Biobank	This study	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000388.v1.p1
The Thousand Genomes Project	1000 Genomes Project; The 1000 Genomes Project Consortium, 2015	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp
The Human Genome Diversity Panel	The Human Genome Diversity Project; Cann et al., 2002	http://www.cephb.fr/hgdp/
Thousand Genomes Imputation Panel	1000 Genomes Project	https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
Genetic Maps	1000 Genomes Project	https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
Software and algorithms		
PLINK	Purcell et al., 2007	https://www.cog-genomics.org/plink2
KING	Manichaikul et al., 2010	http://people.virginia.edu/~wc9c/KING/Download.htm
ADMIXTURE	Alexander et al., 2009	http://dalexander.github.io/admixture/download.html
iLASH	Shemirani et al., 2019	https://github.com/roohy/IBD
iGraph/infomap	Rosvall and Bergstrom, 2008	https://github.com/igraph/rigraph
SHAPEIT	Delaneau et al., 2011	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#download
IMPUTE2	Howie et al., 2009	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#download
bcftools/RoH	Narasimhan et al., 2016	https://github.com/samtools/bcftools
GCTA	Yang et al., 2011	https://cnsgenomics.com/software/gcta/#Download
Umap	McInnes et al., 2018	https://cran.r-project.org/web/packages/umap/index.html

RESOURCE AVAILABILITY

Lead contact

Requests for resources and materials should be directed to lead contact Eimear Kenny (eimear.kenny@mssm.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The exome sequencing datasets used in this study were generated by Regeneron and are not publicly available. The data will be made available for purposes of replicating the results by contacting the corresponding author and appropriate collaboration and/or data sharing agreements.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Study population

The Bio*Me* Biobank is an electronic health record (EHR)-linked biobank of over 60,000 participants from the Mount Sinai Health System (MSHS) in New York, NY. Participant recruitment into Bio*Me* has been ongoing since 2007, and occurs predominantly through ambulatory care practices across the MSHS. Bio*Me* participants consent to provide DNA and plasma samples linked to their de-identified EHRs. Participants provide additional information on self-reported ancestry, personal and family history through questionnaires administered upon enrollment. This study was approved by the Icahn School of Medicine at Mount Sinai's Institutional Review Board (Institutional Review Board 07–0529). All study participants provided written informed consent.

METHOD DETAILS

Ascertainment of Race/Ethnicity Information in BioMe

Participant R/E was solicited in the form of the multiple-choice question with nine options to choose from (the exact phrasing of the question and corresponding response options are delineated in Table S6). Prior to 2014, in addition to a multiple-choice question about R/E, participants were also given the option to report their country of birth. After 2014, enrolling participants were provided with options to report the country of birth of both their parents and grandparents as well. All participants recorded answers to survey question 1, and 43.6% of participants also provided responses to survey question 2.

Genotype Quality Control

Bio*Me* participants (N = 32595) were genotyped on the Illumina Global Screening Array (GSA) platform. Quality control (QC) of the GSA data for N = 32595 participants and n = 635623 variants was performed stratified by R/E category. Individuals with an R/E-specific heterozygosity rate that surpassed ± 6 standard deviations of the population-specific mean, along with individuals with a call rate of < 95% were removed (N = 684 participants in total). N = 80 individuals were then removed for exhibiting persistent discordance between EHR-recorded and genetic sex. A further N = 126 duplicate individuals were also excluded from downstream analysis. In total 31705 passed sample level QC for downstream analysis. All quality control steps were conducted using Plink(v1.90b3.43) (Purcell et al., 2007; Chang et al., 2015). Sites with a call rate below 95% were excluded (n = 19253), along with sites that were seen to significantly violate Hardy-Weinberg equilibrium (HWE) when calculated stratified by ancestry. HWE thresholds for site exclusion varied by R/E, specifically we set a threshold of $p < 1x10^{-5}$ in for all populations except HL, where it was set to $p < 1x10^{-13}$ (n = 11503 SNPs in total). This resulted in the retention of n = 604869 sites.

QUANTIFICATION AND STATISTICAL ANALYSIS

Comparison of Electronic Health Record versus self-reported Race/Ethnicity

Race/ethnicity (R/E) information was extracted from the Electronic Health Records (EHR) for all Bio*Me* participants for every available patient visit between January 2007 and December 2014. This window of time was selected as R/E was recorded using consistent categories. We do not have information whether the R/E was inputted by a medical professional or the patient. The possible R/E designations within this time frame consisted of "African American (Black)," "Asian," "Caucasian (White)," "Hispanic/Latino," "Native American," "Other," "Pacific Islander" or "Unknown." For individuals who had greater than one interaction with the healthcare system and conflicting EHR recorded R/E, we selected the R/E designation assigned at their earliest visit for downstream comparison to self-reported R/E. We mapped EHR recorded R/E to self-reported R/E for N = 36061 Bio*Me* participants across 1492428 healthcare visits in total. For individuals who only self-reported one R/E, in instances where that category had a direct mapping to one of the EHR R/E variables, we made a direct comparison between self-reported and EHR recorded R/E to calculate the percentage of individuals who had been miss-classified in the EHR, with the exception that we collapsed individuals who self-identified as "*Caucasian/White*" (N = 7691), "*Jewish*" (N = 934), or both (N = 935) into one group that we describe as "*European American*," and mapped this to the EHR category of "Caucasian (White)." We also mapped individuals who self-identified as *"East or Southeast Asian (i.e. China Japan Korea Indonesia)"* only to the EHR category "*Asian*." The additional R/E categories derived from the Bio*Me* survey were specifically namely "*South Asian*," "*Mediterranean*," "*Other*," or "*Multiple*" selected ethnicities at enrollment (Table S1; N = 5685 individuals across 182149 healthcare visits). Individuals who self-reported as South Asian were most often classified as "Other" (45.9%) or "Asian" (25.8%). Individuals who self-identified as "*Mediterranean*" were most often classified as "Caucasian (White)" (59.6%) or "Other" (15.7%), while for individuals who either self-reported as "*Other*" or who checked multiple categories there was no clear majority designation (Figure S1A).

Collapsing of Self-Reported Race/Ethnicity Questionnaire Data to explore the relationship with genetic ancestry

To explore the relationship between self-reported R/E and genetic ancestry for the N = 31705 Bio*Me* participants genotyped on the GSA array, we collapsed self-reported R/E information into 8 categories: "African American/African" (N = 7055), "East/South-East Asian" (N = 784), "South Asian" (N = 622), "Native American" (N = 55) ; participants who selected "Caucasian/White," "Jewish" or both were designated "European American" (N = 8619), and participants who selected "Hispanic/Latin American" and any other category were designated as "Hispanic/Latino" (N = 10682). Participants who selected "Mediterranean," "Other" or both were designated as "Other" (N = 2559). Finally, participants who selected any other combination of multiple categories were designated as "Multiple Selected" (N = 1329).

Genetic Relatedness Estimation

Pairwise kinship coefficients were estimated for all Bio*Me* participants (N = 31705) using all N = 604869 SNPs that passed QC using the KING software (Manichaikul et al., 2010) (v1.4) by passing the *–kinship* flag.

Global Ancestry Estimation

Prior to calculating PCA we restricted analysis to common (minor allele frequency (MAF) > 0.01), autosomal sites. We also removed regions of the genome known to be under recent selection, specifically *HLA* (chr6: 27032221-35032223, hg38), *LCT*

(chr2:134242429-136242430), an inversion on chromosome 8 (chr8:6142478-16142491), a region of extended LD on chromosome 17 (chr17:41843748-46922634), *EDAR* (chr2:108383544-109383544), *SLC2A5* (chr15:47707803-48707803), and *TRBV9* (chr7:142391891-142392412). Finally, the GSA data was intersected with and merged with genome sequence data from 26 populations in 1000 Genomes Project phase 3 reference panels (The 1000 Genomes Project Consortium, 2015) (TGP; N = 2504), and 53 populations in the Human Genome Diversity Panel (Cann et al., 2002) (HGDP; N = 986) and 8 additional reference panels genotyped as part of the PAGE consortium (Bari, Khomani, Nama, Oaxacan, Peru Warao, Yukpa, Zapotec) from the Population Architecture using Genomics and Epidemiology (PAGE; N = 700) study, both genotyped on the Multi-Ethnic Genotyping Array (MEGA). This resulted in a total of n = 260502 snps and N = 35854 individuals. The first 20 PCs were calculated using PLINK (v1.9). We also ran ADMIXTURE (Alexander et al., 2009) with 5-fold cross validation from $k = 2$ to $k = 12$ across all individuals inferred to be unrelated (N = 32354 in total, including reference panels), by randomly removing one of each individual in a pairwise relationship defined by KING to be greater than 3rd degree relatives (as defined by a pairwise kinship coefficient of > = 0.0442 in the KING output; this resulted in the exclusion of N = 3500 Bio*Me* participants in total). To visualize fine-scale population substructure we applied the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to the first 10 principal components across all samples using the ''umap'' library in R using the default parameters.

Phasing and Identity-by-Descent Inference

Prior to phasing and inference of Identity-by-Descent, the GSA array data for Bio*Me* participants was lifted over to GRCh37/hg19, before additional quality control was performed on variants, including the removal of SNPs with a call rate below 99% (n = 48436) and variants with a MAF < 1% (n = 135011). Palindromic variants were also excluded at this stage (n = 4375). The data was subsequently merged with the TGP reference panels (N = 2504 individuals), and only intersecting sites were retained, resulting in the retention of n = 402042 SNPs in total. Phasing was subsequently performed per autosome on all N = 34209 individuals with the SHAPEIT software (O'Connell et al., 2014) (*v2.r790*) using the hapmapII genetic map (build: GRCh37/hg19) using default flags and –output-max –force. Phased haplotypes were subsequently converted to plink format and IBD was called using the iLASH software(Shemirani et al., no date) using the following flags:

slice_size 400, step_size 400, perm_count 12, shingle_size 20, shingle_overlap 0, bucket_count 4, max_thread 20, match_threshold 0.99, interest_threshold 0.70, max_error 0, min_length 3

For quality control, IBD tracts that overlapped with low complexity regions were excluded, along with IBD tracts that fell within regions of excessive IBD sharing, which we defined as regions of the genome where the level of pairwise IBD sharing exceeded 3 standard deviations above the genome-wide mean (Figure S4C).

Network Construction and Community Detection

To construct the IBD network, IBD tracts along the genome were summed between each pair of individuals inferred to be 2nd degree relatives or less from Bio*Me* and the 1000 genomes reference panel (N = 31688 individuals in total (N = 29184 of which were Bio*Me* participants inferred to be < 2nd degree relatives, and N = 2504 samples from TGP)) to generate the total sum of IBD sharing per pairwise relationship. This was used to construct an adjacency matrix where each node represents an individual and each weighted edge represents the pairwise sum of IBD sharing. To detect the presence of structure within the IBD network we used the implementation of InfoMap (Rosvall and Bergstrom, 2008) in the iGraph package (R version 3.2.0). Visualization of the IBD network was also performed using iGraph using a Fruchtermen-Reingold layout (with n = 1000 iterations), after excluding poorly connected nodes (< 30 connections).

IBD Network Analysis

We define neighborhood matrices as a matrix-representation of the community assignments determined by using InfoMap (InfoMap assigns each individual one and only one community). A neighborhood matrix is an n x n order matrix, where n is the total number of individuals in the IBD network. Individuals i and j are assigned a 1 for the same community, and 0 for different communities, corresponding to positions N_{ij} and N_{ji}. We can extend this scheme across multiple runs of clustering estimation, particularly crucial when measuring concordance across runs. To compare the ''neighborhood'' of individual k across multiple runs, we use a Jaccard similarity coefficient to measure positive concordance: J = Intersection(row(k) = 1)/Union(row(k) = 1)

The Jaccard similarity coefficient obtained while comparing the k-th row of N_1 and the k-th row of N_2 constitutes a natural measure of community consistency across runs, with J = 1 if k has exactly the same neighbors in the two matrices N_1 and N_2, J = 0 if k does not share any neighbors between the two matrices. This then naturally can be averaged across the entire matrix to determine overall similarity of community assignments between different runs of InfoMap. We use this as our statistic for our permutation testing, by comparing our observed network structure to topologically similar random networks with consistent numbers of edges created with the Erdős–Rényi (ER) model. An empirical p value ascertaining the stability of community assignments represents the probability of randomly obtaining an identical neighborhood matrix as the neighborhood matrix obtained by running InfoMap with the original IBD network.

Community assignments were ascertained by using InfoMap both for stability and robusticity on the original IBD-based network (10 times) and on 100 sets of topologically similar random networks generated with the ER model. We computed the clustering

coefficient *(i.e.* the ratio of existing edges between the neighbors of a specific node relative to the total number of potential edges between said neighbors) for each community. In addition, we used the Wasserstein distance (Villani, 2009).

Analysis of IBD Community Membership Using Population Labels

To explore the correlation between self-reported country of origin, subcontinent and R/E and IBD community membership for Bio*Me* participants we calculated Positive Predictive Values (PPVs), Negative Predictive Values (NPVs), sensitivity and specificity for each self-reported label versus membership of each IBD community using the ''caret'' library in R (Kuhn, 2008). For each of these metrics we treated the self-reported information as the 'ground truth' and the IBD-community designation as the predictor. For the analyses by country and subcontinental origin, US-born Bio*Me* participants were excluded from the calculations.

Inference of Runs of Homozygosity

Runs of homozygosity were also calculated using bcftools/1.0 (Narasimhan et al., 2016). Analysis was performed stratified by IBD community, and using IBD community specific allele frequencies and including genetic maps. Post ROH inference, analysis was restricted to tracts of greater than 3MB in length.

Phenotype ontology

15665 unique ICD-9 and ICD-10 billing codes from the Mount Sinai Bio*Me* biobank were collapsed into 1764 phenotype codes (Phecodes) (Wu et al., 2019; Denny et al., 2010) on the basis of the PheWAS catalog (https://www.phewascatalog.org) using an in-house R script (R version 3.4.1).

Analysis of Enrichment of Phecodes within IBD-Communities

We performed logistic regression systematically across 1764 phecodes and all IBD communities with > = 500 members who were also Bio*Me* participants (N = 7 in total), (i.e., excluding communities predominantly composed of individuals from the 1000 Genomes Reference panel). To avoid spurious association, phecodes that were present in less than 10 instances per community were excluded from the analyses. To perform the regression we encoded IBD community membership as a binary predictor variable and generated Plink format ''ped'' files for each community, where community membership was encoded as ''1'' and non-membership encoded as ''2.'' We then performed logistic regression for each phecode and for each community using Plink(v1.9), adjusting for age and sex as covariates, and excluding 2nd degree relatives and above. Statistical significance was determined for each IBD community-wide association via Bonferroni correction.

Genotype Imputation and Polygenic Risk Score Estimation

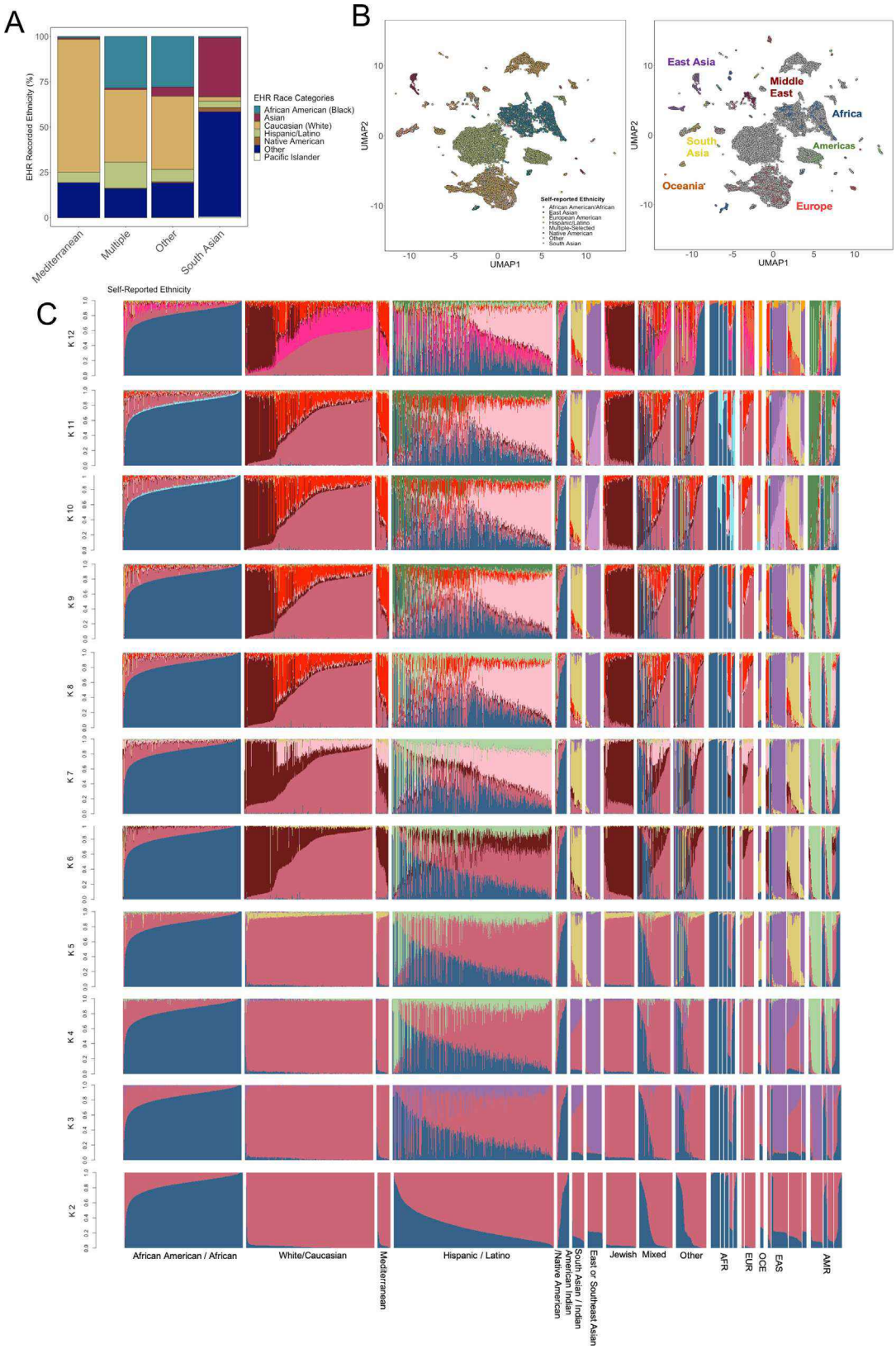
Imputation was performed on phased haplotypes (described previously) using IMPUTE(v2.3.2) (Howie et al., 2009) with the phase III Thousand Genomes data as reference panel, and the addition of the following flag: *''-filt_rules_1 'ALL<0.0002' 'ALL>0.9998'.* Weights for polygenic risk scores (PRS) for five diseases and body mass index (BMI) were downloaded from: http://kp4cd.org/dataset_downloads/mi

PRS were calculated for each of the five diseases across all N = 31705 Bio*Me* participants by summing genotype data previously imputed to the Thousands Genomes Phase III reference panel using the ''*–score sum*'' flag in PLINK(v1.9) and participants were subsequently stratified into groups based on their IBD community membership for further analysis. PRS performance was assessed by calculating the Area Under the Curve (AUC) using the pROC package (Robin et al., 2011) in R using disease case status for a given phecode as the outcome variable, and PRS as the predictor variable.

Annotation of Clinically Relevant Founder Variation from Exome Data

Founder variants for both the AJ and PR communities were curated via literature review and extracted from exome data available for the N = 27727 Bio*Me* participants that were a part of the community detection analysis. The number of heterozygous carriers for each variant was then ascertained for the subset of N = 27627 participants for whom the self-reported ancestry survey information was complete. This analysis was performed stratified by self-reported R/E, self-reported country of origin, and IBD-community membership using PLINK(v1.9)

Supplemental figures



(legend on next page)

Figure S1. Exploration of R/E and genetic ancestry in a health system, related to Figure 1

(A) EHR R/E designations for individuals whose self-reported R/E did not correspond to one of the EHR categories. The x axis represents self-reported R/E categories from Bio*Me* and the y axis represents the percentage of EHR recorded R/E designations per group by R/E category.

(B) UMAP projection of the first 10 principal components for Bio*Me* participants. (A) UMAP colored by self-reported ethnicity for the Bio*Me* participants genotyped on the GSA array colored by self-reported R/E. (B) UMAP of all Bio*Me* participants (gray) along with reference samples from 87 global populations colored by Continental region of origin.

(C) ADMIXTURE runs from k = 2 to k = 12 in N = 31705 unrelated Bio*Me* participants and reference samples from 87 global populations. ADMIXTURE output of Bio*Me* participants stratified by their self-reported R/E (left) and reference samples from 87 global populations (right). Populations labeled ''AFR'' correspond to reference samples from Africa, ''EUR'' from Europe, ''OCE'' from Oceania, ''EAS'' from East Asia, ''SAS'' from South Asia and ''AMR'' from the Americas.

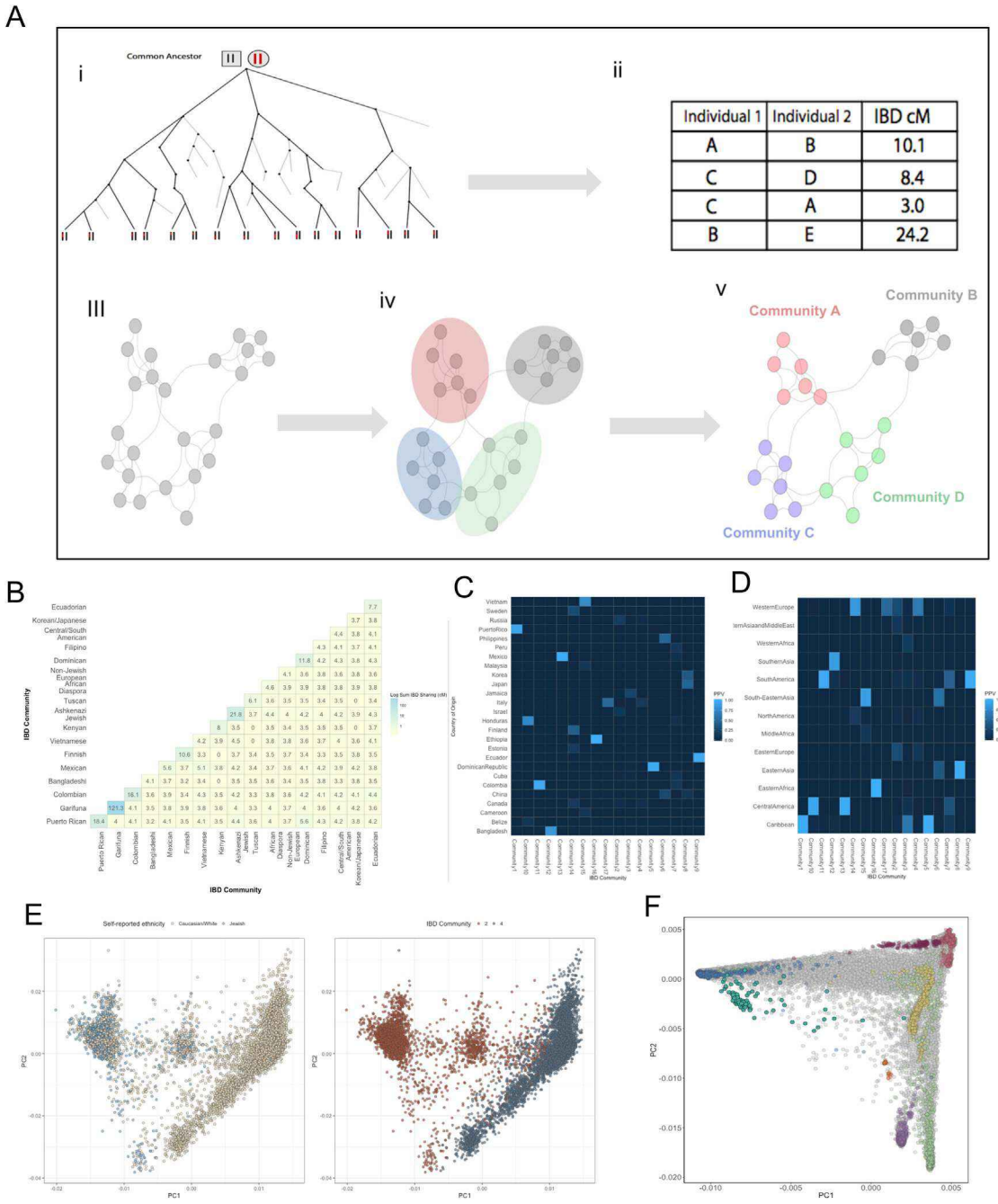


Figure S2. Network-based inference of fine-scale populations, related to Figure 2

(A) Schematic of the IBD-community detection workflow. (i) Haplotypes inherited Identical-by-Descent (IBD) from a recent, common ancestor are present and readily detectable in genomic data. (ii) Detected IBD segments can be used to construct an adjacency matrix where the pairwise relationship between each Bio*Me* participant is represented by the total sum of IBD segments they share across their genome (cM). (iii) This adjacency matrix can be used to construct a network where every node represents an individual and each (weighted) edge represents the sum of IBD sharing between a given pair. (iv) Running the community detection algorithm InfoMap over the IBD network allows for the detection of 'communities' of individuals that are statistically enriched for the sharing of IBD. (v) InfoMap returns community membership status for each node in the network. (B). Median sum of IBD sharing within and between the largest IBD Communities in Bio*Me*. Each tile within the heatmap represents the median sum of IBD haplotype sharing (log_{10} scale) within and between- IBD communities, with blue representing higher IBD sharing and yellow representing lower levels of sharing. (C) Positive Predictive Values (PPVs) for geographical origin versus IBD Communities. PPVs for country of origin versus the 17 largest IBD communities. (D) PPVs for sub-continental region of origin. Results are only shown for population labels with a PPV > = 0.1 for at least one IBD community. (E) PCA analysis of Bio*Me* European Americans reveals IBD-Community membership reflects genetic Jewish versus non-Jewish ancestry. (Left) PCA plot of Bio*Me* participants who self-identify as "Jewish" (blue) and "Caucasian/White" (yellow) reveal clustering in

(legend continued on next page)

PCA space. (Right) The same PCA plot colored by IBD-community membership, where community "2" (red) appears to represent genetic Jewish ancestry, while community "4" (dark blue) represents Non-Jewish European ancestry. (F) PCA analysis with of the Bio*Me* IBD community inferred to be Garifuna. Bio*Me* participants who belong to the IBD community that we infer are likely to be Garifuna (teal) cluster on a cline between the African (blue) and Amerindigenous (green) reference panels.

Figure S3. Comparison of IBD communities and k-means clustering to explore genetic ancestry, related to Figure 2

IBD based community detection versus k-means clustering over principal component analysis. Comparison of population designation using k-means clustering versus IBD communities with strong PPVs (> 0.9) for a particular country of origin. The first panel of each plot represents classification metrics for IBD community

(legend continued on next page)

detection against a given country of origin, specifically Puerto Rican (A), Dominican (B), Ecuadorian (C), Colombian (D), Mexican (E) and Ethiopian (F). The following panels represent the same metrics for the cluster obtained via k-means clustering over PCA with the highest PPV for a given specification of *k* (ranging from k = 5 to k = 20). In each instance the IBD-community detection is constantly able to better classify individuals in recent diaspora populations based on Positive Predictive Value, Negative Predictive Value, Sensitivity and Specificity.

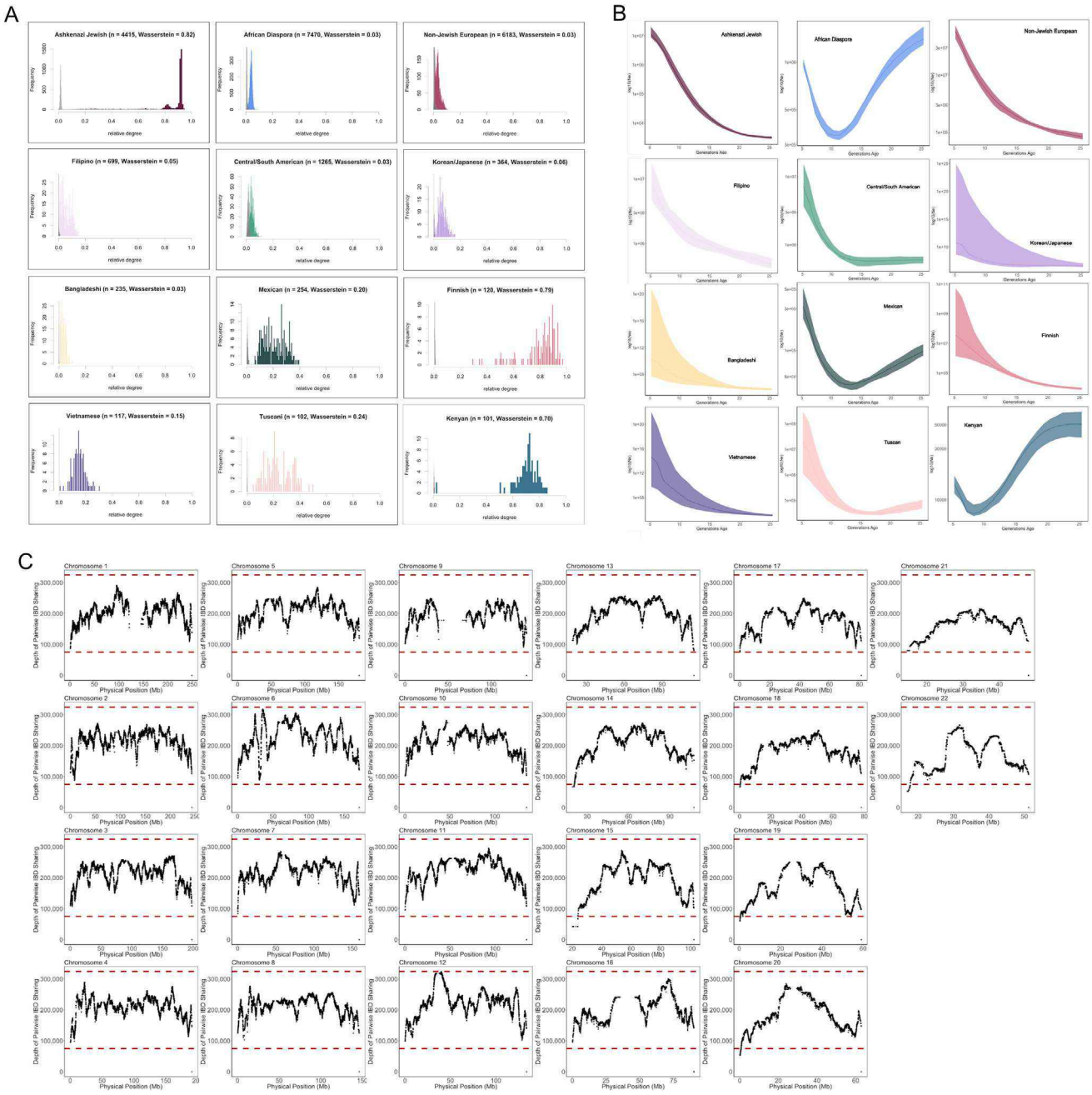


Figure S4. Related to Figure 3

(A) Distribution and properties of IBD sharing within and between IBD communities. Analysis of the distribution of degree sharing within versus between community for the Ashkenazi Jewish, African Diaspora, Non-Jewish European, Filipino, Central/South American, Korean/Japanese, Bangladeshi, Mexican, Finnish, Vietnamese, Tuscan and Kenyan communities. The intra-community degree distribution is shown in color and the inter-community degree distribution is shown in gray. The strong bimodality in the degree distributions of the Ashkenazi Jewish and Finnish communities, quantified by a Wasserstein metric value of 0.82 and 0.79 respectively, are indicative of a founder effect. The distributions related to the other communities show low bimodality, quantified by Wasserstein metric values ranging from 0.03 to 0.24 (B) Estimated effective population size for IBD communities inferred via IBD_{Ne}. Estimated effective population size of the 17 largest IBD communities. (C) Pile up of Identity-by-Descent (IBD) haplotype sharing along the genome in Bio*Me* and TGP samples (N = 34209 total). Each panel represents the number of shared IBD haplotypes covering a given site for each of the 22 autosomes. The x axis represents the physical position along the chromosome (Mb) and the y axis represents the number of haplotypes covering each position. The red dashed lines represent ± 3 standard deviations from the genome-wide mean.