# Decoding Genetics, Ancestry, and Geospatial Context for Precision Health

Satoshi Koyama M.D., Ph.D.[1,2,3], Ying Wang M.D., Ph.D.[1,4,5], Kaavya Paruchuri M.D.[1,2,3], Md Mesbah Uddin Ph.D.[1,2], So Mi J. Cho Ph.D.[1,2,6], Sarah M. Urbut M.D., Ph.D.[1,2], Sara Haidermota B.S.[1,2], Whitney E. Hornsby Ph.D.[1,2], Robert C. Green M.D., M.P.H.[3,7,8], Mark J. Daly Ph.D.[4,5,9,10], Benjamin M. Neale Ph.D.[1,4,5], Patrick T. Ellinor M.D., Ph.D.[1,2,3], Jordan W. Smoller M.D., Sc.D.[3,11,12], Matthew S. Lebo Ph.D.[3,13,14], Elizabeth W. Karlson M.D., M.S.[3,13,15], Alicia R. Martin Ph.D.[1,4,5,*], and Pradeep Natarajan M.D., M.M.Sc[1,2,3,*]

1. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
2. Cardiovascular Research Center and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
3. Department of Medicine, Harvard Medical School, Boston, MA, USA
4. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
5. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
6. Integrative Research Center for Cerebrovascular and Cardiovascular Diseases, Yonsei University College of Medicine, Seoul, Republic of Korea
7. Department of Medicine (Genetics), MassGeneralBrigham, Boston, MA, USA
8. Broad Institute and Ariadne Labs, Boston, MA, USA
9. Institute for Molecular Medicine Finland (FIMM), Finland
10. University of Helsinki, Helsinki, Finland
11. Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
12. Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA
13. Mass General Brigham Personalized Medicine, Cambridge, MA, USA
14. Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA
15. Division of Rheumatology, Inflammation and Immunity, Department of Medicine, Brigham and Women's Hospital., Boston, MA, USA

* A.R.M. and P.N. jointly supervised this work.

Please address correspondence to:

Alicia R. Martin, Ph.D.
185 Cambridge Street
Boston, MA 02114
armartin@broadinstitute.org

Pradeep Natarajan, M.D., M.M.Sc
185 Cambridge Street
Boston, MA 02114
pnatarajan@mgh.harvard.edu

Figures 5

Supplemental Figures 10

Supplemental Tables 7

Supplemental Information (9 Figures)

## Abstract

Mass General Brigham, an integrated healthcare system based in the Greater Boston area of Massachusetts, annually serves 1.5 million patients. We established the Mass General Brigham Biobank (MGBB), encompassing 142,238 participants, to unravel the intricate relationships among genomic profiles, environmental context, and disease manifestations within clinical practice. In this study, we highlight the impact of ancestral diversity in the MGBB by employing population genetics, geospatial assessment, and association analyses of rare and common genetic variants. The population structures captured by the genetics mirror the sequential immigration to the Greater Boston area throughout American history, highlighting communities tied to shared genetic and environmental factors. Our investigation underscores the potency of unbiased, large-scale analyses in a healthcare-affiliated biobank, elucidating the dynamic interplay across genetics, immigration, structural geospatial factors, and health outcomes in one of the earliest American sites of European colonization.

## Introduction

Determinants of health include a complex interplay of sociodemographic, structural, genetic, and environmental factors that are also contextually dependent on time and geography. Disease risk prediction models and therapeutic paradigms are largely agnostic to many of these important features yet are intended for broad use. Such training datasets often lack the breadth and depth of information and the inherent diversity across features required for equitable applications. The United States populace is highly diverse, marked by complex migration patterns and dynamic social constructs and represents multilevel health contributors. For example, it is widely recognized that the prevalence of diseases is closely linked to individual or neighborhood social deprivation, which further varies across smaller domains and regions[1,2]. Furthermore, these determinants differentially contribute to health outcomes depending on local factors[3,4].

Contemporary healthcare-associated biobanks represent a new opportunity to discover novel determinants of health and augment translation to clinical care. Such endeavors represent a recent collaborative synergy of large-scale population-based-[5-8] and local healthcare- biobanks[9-13]. Understanding how DNA sequence variation tracks contemporary and historical population demographics can provide insights on differential disease burdens. For example, rs5742904 (c.10580G>A, p.Arg3527Gln) in *APOB*, a founder pathogenic mutation for familial hypercholesterolemia has significantly higher allele frequency in Old Order Amish people. It substantially explains the increased risk for coronary artery atherosclerosis in this population[14,15]. Important insights related to genetic variations and clinical outcomes may often require profiling diverse populations. For example, the discovery of the association between disruptive variants in *PCSK9*[16,17] and lower low-density lipoprotein cholesterol in the African population, where these variants are prevalent, facilitated drug development. As another instance, *G6PD* deficiency[18] was recognized as a prevalent hemolytic disease in Sub-Saharan Africa. Studying diverse populations across a spectrum of diseases is crucial to assess the penetrance of disease-associated monogenic alleles[19] and polygenic models[20].

Recent analyses of biobanks in the United States have uncovered the complex genetic structure of Hispanic and/or Latinx groups tracing their origins to the Americas[9,12]. In these efforts, it has been demonstrated that the fine genetic structure within biobanks can identify varying disease risks by capturing both ancestral and social structures, thereby contributing to the advancement of personalized medicine. Separately, recent advances in data size and methodology have enabled us to precisely characterize the complex population dynamics associated with multiple colonization and admixture events[21-23] However, the interplay across these features, or their interaction with large-scale genetic association studies using whole-genome imputed or sequenced data, remains understudied.

The New England region represents among the earliest European colonization of the United States with sequential ongoing migration from diverse groups. In this study, we examined the genetic variation across New England coupled with sociodemographic, clinical, and environmental/geospatial factors in the Mass General Brigham Biobank (MGBB). By applying a network-based clustering algorithm with newly generated reference dataset, we established fine genetic clusters with subcontinental resolution within MGBB. These clusters exhibited distinct genetic properties, geographic distributions, socioeconomic statuses, and disease risks. In combination with rare and common variant genetic association analyses, we gained further insights into the different disease risks among these populations. Collectively, this study highlights the power of large-scale, unbiased analyses within a healthcare-based biobank to understand the complex interplay between genotypes and phenotypes, paving the way for increasingly personalized interventions.

## Results

### Participant recruitment and Electronic Health Record (EHR) based phenotyping

Since 2010, 142,238 individuals within the Mass General Brigham (MGB) network, the largest healthcare system in Massachusetts, have consented to participate in the MGBB as of May 11, 2023 (Figure 1a, Supplemental Information Figure 1, Supplemental Table 1). Among participants, 99.5% (n = 141,519) consented to re-contact. 56.8% of participants are female (n = 80,851, Figure 1b). Median [interquartile range; IQR] age at consent was 51 [35 – 63] years for female participants and 58 [43 – 68] years for males. Self-reported races were 84.4% White, 4.5% Black, and 3.0% Asian. Self-reported ethnicities were 86.6% non-Hispanic and 2.44% Hispanic (Supplemental Information Figure 2). The participants are primarily cared for at the two flagship MGB hospitals, both located in Boston, MA, and their associated clinics – Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH, Figure 1c). The biobank data is interlinked with EHRs with phenotype data across the MGB network, as well as notable specialty centers in Boston, MA including the Mass Eye and Ear Institute (MEEI) and Dana-Farber Cancer Institute (DFCI). To generate systemically annotated prevalent/incident outcomes, we extracted International Classification of Diseases codes, Nineth (ICD9) and Tenth (ICD10) revisions, from the EHR and mapped them to PheCodes[24]. We identified 1,577 of 1,860 possible PheCodes with at least one event (Figure 1d). Participants were followed for a median [IQR] of 4.3 [2.5 – 6.0] years after inclusion to MGBB with a median of 12 [5 – 25] incident events per person.

## Fine-scale clustering of genetic ancestry in MGBB

Extending beyond traditional low-dimensional projections of continental ancestry from genome-wide data, we utilized high-dimensional principal components (PCs) to achieve greater granularity. Using genome-wide genotyping arrays, we genotyped 53,306 participants in the MGBB. By employing the top 30 genetic PCs and a network-based clustering approach[25], we identified 30 data-driven distinct ancestral clusters (Figure 2a, 2b, and 2c, Figure S1, Supplemental Table 2). The largest cluster (cluster 0, ordered by sample size) includes 11,875 (22.3%) MGBB individuals. The smallest (cluster 29) includes only one MGBB participant as well as 27/27 reference Sardinian individuals from Human Genome Diversity Project (HGDP)[26], suggesting the origin of this individual. As such, unsupervised clustering with diverse populations from the 1000 Genomes Project Phase 3 (1KG)[27] and HGDP reference panels[28] allowed us to infer the genetic similarity between these clusters and populations worldwide in an unbiased manner.

Cluster 0 was genetically similar to the Western European populations in the reference dataset (CEU [Utah residents with Western or Northern European ancestry] and GBR [British from England or Scotland] in 1KG, French and Orcadian in HGDP, Figure S2). In addition to the cluster 0, we identified eight distinct 1KG+HGDP-EUR-like clusters that cluster with Italian, Russian, Spanish, Adygei, Finnish, Basque, and Sardinian ancestries reflecting known patterns of migration to the Greater Boston Area. We also identified two distinct 1KG+HGDP-AMR-like clusters (cluster 5 enriched with PUR [Puerto Rican in Puerto Rica], cluster 8 with Colombian, Maya, PEL [Peruvian in Lima, Peru], Pima, CLM [Colombian in Medellín, Colombia], MXL [Mexican ancestry in Los Angeles, CA]), four African-like clusters (cluster 6 enriched with African Caribbean in Barbados and African Ancestry in Southwest USA, cluster 17 specific to Nigerian Africans [Esan in Nigeria, Yoruba in Ibadan, Nigeria], cluster 22 specific to Kenyan Africans [Bantu and Luhya in Webuye, Kenya], and cluster 18 with other Western Africans [Mandinka, Mende people in Sierra Leone, Gambian in Western Divisions in the Gambia]), and three East Asian-like clusters (cluster 19 specific to Japanese, cluster 20 specific to Uygur, and cluster 9 with other East Asians). We identified a single large cluster (cluster 10) enriched in South Asian reference populations, however, Hazara and Kalash populations formed distinct clusters.

Even with a diverse reference dataset, eight clusters comprising 9,874 (18.8%) MGBB individuals did not have enrichments of specific populations from the reference dataset. Among these unannotated clusters, seven clusters (4,7,11,13,14,15, and 21) exhibited genetic similarities to 1KG+HGDP-EUR populations. We calculated pairwise Fixation Index ($F_{ST}$) values among clusters, and then constructed a phylogenetic tree of the clusters. The observed population differentiation between clusters further

corroborates the ancestral relationships but notable distinctions from continental populations and residents in New England (Figure S3a).

We also conducted ADMIXTURE[29] analyses to infer continuous population structures within these genetic clusters, many of which show similar patterns of structure across increasing numbers of ancestral components (Figure S3b). Using cross validation, ten was the best fit number of components (Supplemental Information Figure 3). We identified two 1KG+HGDP-EUR-like components (distinguished by components 4 and 9). The component 4 was most enriched in the Finnish-like cluster (cluster 24), and relatively enriched in northern European-like clusters (0, 1, and 3) more than the southern European-like clusters (2, 12, and 14). In contrast, the component 9 was enriched in the southern European-like cluster. We also observed a third component included in the European ancestries (component 3). This component 3 is prominent in the Kalash (Indo-European in northwest Pakistan) and other Pakistani reference populations. While this was enriched in southern European-like clusters, it was more enriched in un-annotated European-like genetic clusters 4, 7, 11, 13, and 21 than other annotated European genetic clusters, possibly consistent with Middle Eastern origins as this group is poorly represented in reference datasets.

Cluster 4 – the 5$^{th}$ largest cluster in this study (n = 3,514) – is one of such un-annotated European clusters. By comparison of allele frequencies between gnomAD[30] ancestries and our dataset, we found that cluster 4 has allele frequencies most similar to the Ashkenazi Jewish reference population (Supplemental Table 3). We also observed strong enrichment of skin neoplasms and inflammatory bowel diseases which were previously noted to be enriched in known Ashkenazi cohorts (Figure S4). We also observed significant enrichment of Ashkenazi Jewish founder mutations (e.g., *APC* c.3920T>A, p.Ile1307Lys, *BRCA1* c.68_69del, p.Glu23fs, *BRCA1* c.5266dup, p.Gln1756fs, *BRCA2* c.5946del, p.Ser1982fs)[31,32] in this genetic cluster (Figure S5). We also observed enrichments of these founder mutations in un-annotated European clusters 11 and 14 suggesting close genetic relationships between these clusters to the Ashkenazi-like cluster 4. In addition to the European-like components, we also identified two different components (components 5 and 7) enriched in the East Asian clusters. One of these components (component 5) was also observed in the Finnish-like cluster which is consistent with previous observations[33-36].

## Effective population size of ancestry clusters in the Greater Boston area

To characterize the ancestral clusters observed in the MGBB, we estimated the historical transition of effective population size of each cluster using genome-wide genetic data (Figure S6). We conducted IBD-based estimation for effective population sizes ($N_e$). Our results were consistent with some prior results conducted outside of the U.S. For example, we replicated previously described bottleneck event in Ashkenazi-like population (cluster 4). The lowest $N_e$ was estimated to be 1,170 (95% CI = 1,100 – 1,270) individuals at 28 generations ago[37]. We observed similar bottleneck events in clusters 11 and 14 around the same generation (minimal $N_e$ was 4,570 [4,210 – 5,030] in cluster 11 and 32,600 [30,700 – 35,900] in cluster 14) consistent with the aforementioned sharing pattern of Ashkenazi founder mutations with cluster 4. The largest genetic cluster 0 indicates a population bottleneck occurring approximately 12 generations ago. This timeframe coincides with the initial colonization of the Boston area by British settlers. Intuitively, this event is not evident in the British population from the UK Biobank (UKBB) here or in previous studies[38,39] (Supplemental Information Figure 4), suggesting a unique founder event among British Americans due to colonization. We also observed a significant bottleneck event in the Admixed American populations, specifically in clusters 5 and 8, with a pronounced magnitude in the Puerto Rican-like cluster 5 (minimal $N_e$ was 11,300 [11,100 – 11,600]). However, we did not observe such bottlenecks for other clusters potentially reflective of more continuous migration.

## Genetic clusters, geographic and socioeconomic factors, and disease risks

We used geospatial scan statistics to understand the geographical structure of MGBB ancestry clusters. We observed 22 statistically significant regions of geographical enrichment among 13 genetic clusters in smaller than 4-km radius areas. We observed distributions of genetically inferred clusters recapitulating colonization histories into the Greater Boston Area (Figure 3a). One example of strong enrichment was observed in the southern area (Roslindale / Mattapan / Dorchester and separately Roxbury) by cluster 6 (ACB [African Caribbean in Barbados] and ASW [African ancestry in Southwest U.S.], expected number 105 and observed number 725, $P < 1 \times 10^{-17}$). Another strong enrichment is observed northern of Boston (Charlestown / Chelsea) by cluster 5 (PUR [Puerto Rican in Puerto Rico], expected number 180 and observed number 624, $P < 1 \times 10^{-17}$) in addition to Boston's South End extending to Roxbury / Hyde Park / Jamaica Plain. We also observed enrichment of Ashkenazi Jewish-like (cluster 4) and East Asian-like clusters (cluster 9) in areas seeded by early founding communities, such as Back Bay / Brookline / Cambridge (cluster 4) and Allston (cluster 9), respectively.

The western European-like clusters cluster 0 and cluster 1 were similar in conventional PC space (Figure S1) and ADMIXTURE analysis (Figure S3b)[29], but well differentiated by network-based clustering (Figure 2a) as well as geospatially. The CEU/GBR-like cluster 0 was enriched in central areas of the Boston (Beacon Hill) and Cambridge (Harvard Square), representing the earliest sites of British colonization. Cluster 1 (Orcadian-like, tagging northern populations of the British Isles including those hailing from Scotland and Ireland) is enriched in two different geographical locations, including Chelsea and South Boston, where secondary colonization occurred. These different geographical enrichments of cluster 0 versus cluster 1 reflect the distinct histories of these two similar, but distinct genetic European ancestries.

Socioeconomic status was correlated with both genetic ancestry as well as ancestral geographic distributions[1]. Using geocoded location information for each participant in our study, we calculated a Social Deprivation Index (SDI, ranged from 0 to 100, higher SDI indicating greater deprivation) for each participant and associated this with healthcare outcomes (Figure S8, Supplemental Table 4). The distributions of SDI widely differed across genetic ancestries (Figure S7a). Specifically, Cluster 6 (enriched with African Caribbean in Barbados like population) exhibited the highest level of deprivation, with a median [IQR] SDI score of 83 [52 - 94]. Conversely, Cluster 4 (Ashkenazi Jewish-like) had the lowest deprivation, as indicated by a median [IQR] SDI score of 22 [8 - 45].

To systemically identify the associations between socioeconomic status and disease risk in MGBB, we associated SDI with phenome-wide outcomes captured by EHRs, adjusting for genetic ancestries. We found SDI was significantly associated with 400 out of 1,564 phenome-wide outcomes (Bonferroni $P <$

224 0.05/1,564 = $3.2 \times 10^{-5}$, Figure 3b). Greater SDI was generally associated with increased disease
225 prevalence and incidence (385 out of 400). The strongest SDI-associated PheCodes was with Tobacco
226 use disorder (odds ratio; OR [95%CI]) per one standard deviation (SD) of SDI was 1.54 [1.48 – 1.60],
227 followed by Mood disorders (OR = 1.26 [1.23 – 1.30]), and Depression (OR = 1.26 [1.23 – 1.30]). As
228 represented by these examples, we observed greater risks for Mental disorders, followed by
229 Uncharacterized Symptoms, Respiratory, and Circulatory systems per PheCodes categories,
230 respectively (Figure 3c). However, prostate (OR 0.85 [0.80 – 0.90]) and breast (OR 0.89 [0.84 – 0.95]),
231 cancer had significant/nominal negative associations with SDI (Figure 3b).

232 Using coronary artery disease (CAD) as an example of a common complex condition, we identified a
233 significant association between SDI and CAD independent from clinical and genetic risk. The
234 association remained significant even after adjustments for clinical risk score (Pooled Cohort Equation,
235 PCE)[40,41], and polygenic risk score[42] (PRS, $OR_{1SD\text{-}SDI}$ = 1.26 [1.17 – 1.35], $OR_{1SD\text{-}PCE}$ = 1.73 [1.65 –
236 1.81], $OR_{1SD\text{-}PRS}$ 1.24 [1.13 – 1.36], in the multivariate model adjusted by the first ten genetic PCs,
237 Figure S7).

## Exome Sequencing in MGBB

Using high-coverage whole exome sequencing in the same group of individuals, we systemically identified rare coding variants in MGBB. There were significant differences in variant distributions across clusters. For instance, the Ashkenazi-like cluster 4 had fewer singleton variants (median [IQR] = 138 [126 – 152] for cluster 4, and 489 [408 – 566] for others). In contrast, there were significantly more singletons in clusters 11 and 14, even though they are closely related to cluster 4 (320 [294 – 361] and 400 [372 – 436], respectively, Supplemental Information Figure 5).

We identified median 15 [12 – 18] rare (Minor allele frequency, MAF < 0.01), high-confidence predicted loss-of-function (pLOF) variants per participant (Figure 4a). The largest number of pLOF variants were observed in African-like clusters (23.5 [20.25 – 26.0] in cluster 22, 21 [19 – 22] in cluster 26, 21 [17 – 24] in cluster 17, 20 [17.0 – 23.0] in cluster 6). The Northern European-like clusters 1 and 0 had the fewest pLOFs (13 [11.0 – 16.0] in cluster 1 and 14 [12 – 17] in cluster 0 and 3). We also identified 1,425 individuals (2.8% of total population) with at least one rare autosomal pLOF homozygous genotype across 761 genes.

We next explored established pathogenic variants (Figure 4b) in MGBB. 2.6% (1,318/50,625) of participants carry a potentially actionable pathogenic/likely pathogenic variant per American College of Medical Genetics and Genomics secondary findings guideline (ACMG SF, version 3.1)[43,44]. These included 6 homozygotes (*TP53*, *LDLR, 4 MUTYH*), and 7 potential compound heterozygotes (2 *BTD*, 3 *MUTYH*, *ATP7B*, and *GAA*). Across genetic clusters, we observed substantial differences in the prevalence of these pathogenic variants (Figure 4c). The highest rate of actionable findings (> 4%) was observed in clusters 11 and 4 despite generally having the lower prevalence of very rare variants. Conversely, non-European clusters generally showed lower rates for annotated actionable variants. Considering higher number of alternate allele-counts in the non-European clusters, the genetic diagnostic rate was significantly lower in non-European populations (Figure S9). Namely, pLOF variants on ACMG SF v3.1 genes found in African and East Asian clusters have significantly lower likelihood of being annotated with a high-quality (more than equal two-stars) pathogenic/likely pathogenic annotation in comparison to the European participants ($OR_{AFR}$ 0.27 [0.18 - 0.41] and $OR_{AMR}$ and 0.48 [0.35 – 0.65], tested by Fisher's exact test), at least partly related to the underrepresentation of causative variants recurrently observed in non-European groups in the ClinVar[45,46] database.

To understand the clinical consequences of rare coding variants, we performed exome-wide and phenome-wide association study (PheWAS) across 1,454 PheCodes and 14,912 genes. We did not find substantial evidence of inflation in the test statistics (median Lambda $GC_{1.0\%}$ 0.89 [0.86– 0.93] for AFR, and 0.90 [0.87 – 0.93] for AMR, 1.02 [0.99 – 1.05] for EUR, Supplemental Information Figure 6a).

We identified 51 significant associations ($P < 1.8 \times 10^{-9}$, 0.05/28,035,307 tested phenotype-transcript pairs, Supplemental Table 5) in the burden of rare pLOF and deleterious missense variants with 14 genes, which included 8 ACMG SF v3.1 genes across 45 clinical outcomes (Figure 4d). In addition to the genes associated with known traits, we found significant associations between *PTEN* deleterious variants and increased risk for secondary hypothyroidism. This link was not described by previous rare variant targeted analysis[47,48] while *PTEN* deleterious variants have been known to be causal for hamartoma syndrome including thyroid cancers and abnormalities[49,50]. Nevertheless, we highlight numerous persistent risk signals from known Mendelian mechanisms of disease in MGBB.

## Genome-wide PheWAS in MGBB

To further explore the relationship between genotype and phenotype in MGBB, we conducted a comprehensive genome-wide PheWAS using ICD code-based PheCodes. We associated over 20 million common variants in African, European, and Admixed populations, which were imputed using the TOPMed imputation reference panel,[51] with 1,461 PheCodes. Similar to the rare variant burden test, we observed calibrated test statistics overall (median Lambda GC [IQR] were 0.98 [0.95 – 1.00] for AFR, 0.97 [0.93 – 1.00] for AMR, 1.02 [0.99 – 1.03] for EUR, Supplemental Information Figure 6b). We identified 111 associations that reached genome-wide significance ($P < 5 \times 10^{-8}/3{,}048 = 1.6 \times 10^{-11}$, Figure 5a, Supplemental Table 6), including 3 African and 1 AMR associations. We refined the prognosis of identified known low-frequency monogenic variants. For instance, we observed that the variant rs6025 (*F5,* c.1601G>A, p.Arg534Gln; Factor V Leiden) is strongly associated with Congenital deficiency of other clotting factors, including factor VII (OR [95%CI] = 11.56 [9.37 – 14.27], $P = 3.6 \times 10^{-68}$). Similarly, rs113993960 – a pathogenic variant in *CFTR* (c.1521_1523del, p.Phe508del) – is associated with Cystic fibrosis (OR 14.67 [11.66 – 18.44], $P = 2.4 \times 10^{-87}$).

Some of these variants exhibited a pronounced recessive effect on the phenotype. A prime example is the variant rs72660908, which is associated with Rhesus isoimmunization in pregnancy (Figure 5b). This medical condition exemplifies recessive inheritance resulting from the deletion of the *RHD* gene. As anticipated, the OR for heterozygotes was not significant ($OR_{Hetero} = 1.18$ [0.53 – 2.7]) compared to the strong effect observed in homozygotes ($OR_{Homo} = 24.3$ [14.6 – 43.1], Figure 5c). Recent large-scale sequencing analysis of structural variants[52] identified high linkage disequilibrium (LD) between rs72660908 and a large deletion affecting *RHD* ($R^2 > 0.99$), which we support by strong expression quantitative trait loci (eQTL) effect of rs72660908 on *RHD*[53] (Figure 5d) and low coverage by exome sequencing in the MGBB (Supplemental Information Figure 7). We also observed a significant enrichment of cases among individuals who were homozygous for rs72660908, with 127 out of 156 cases having the G/G genotype at this locus. As previously reported, individuals with the homozygous alternate allele for rs72660908 have a recessive inheritance pattern in European ancestry populations, but the frequency fluctuates among clusters (17% in cluster 0 and 9% in cluster 4, Figure 5e), and we observed very few copies in AFR/EAS populations.

Another noteworthy example is the association between rs73404549 and sickle cell anemia in the AFR population. This variant is in strong LD with rs334 (*HBB* c.20A>T, p.Glu7Val, HbS), a well-established pathogenic variant for sickle cell anemia. Despite high medical relevance, rs334 was not included in the TOPMed reference panel. We re-evaluated the impact of rs334 using exome sequencing data on sickle cell anemia and clinical red blood cell counts. rs334 showed stronger and more penetrant effect size for

sickle cell anemia than imputed rs73404549 ($\beta_{rs334}$ = 4.14 ± 0.25, $P_{rs334}$ = 5.9 × $10^{-94}$, $\beta_{rs73404549}$ = 3.05 ± 0.22, $P_{rs73404549}$ = 4.4 × $10^{-46}$) with 14 homozygotes for rs334 and a penetrance rate of 79% (Figure S10a). Additionally, we noted another missense variant rs33930165 in *HBB* (*HBB* c.19G>A, p.Glu7Lys, HbC) – previously shown to confer malarial resistance without sickle cell anemia. We found 5 participants with sickle cell anemia heterozygous for both HbS or HbC with significantly lower red blood cell counts compared to heterozygotes for either genotype (Figure S10b).

## Discussion

In this study, we conducted multidimensional investigations into the structure of a modern healthcare-based biobank based at one of the earliest sites of durable European colonization. We show how expanded immigrant communities in the U.S. often exhibit genetic similarities to contemporary continental populations and reflect common bottlenecks. However, we also observe distinct bottlenecking effects of early colonization and patterns of admixture, and identify populations not well represented in reference datasets. Using geospatial indices, genetic ancestries, and phenome-wide outcome data, we described the architecture of diseases associated with regional socioeconomic factors such as area-level poverty, education level, single parent households, living in rented housing units or overcrowded housing units, living without a care or unemployment[54]. We further leverage rich genotyping and phenotyping to clarify several clinically relevant genetic associations complementing clinical and environmental features. This work advances an overall goal of comprehensively quantifying heterogeneous health determinants that uniquely vary across diverse communities in the U.S.

Leveraging population genetics, we delineated the complex ancestral components present within the Boston area. While our findings align well with prior studies on nationwide cohorts[22,23,38,39], our research offers further granular insights into the individual-level ancestral histories of the participants, including lifestyle, genetic, and social risk factors associated with the diseases. We used genetic variation to explore the dynamic interplay of migration, expanded colonization sites, and geographic and community variation, aiming to study how social deprivation influences health, independent of the clinical and genetic risk factors. With distinctions from continental level ancestral histories, the complex history of communal- and individual-level factors can be uniquely mapped by healthcare-associated biobanks to uncover novel important drivers of health.

Area-defined SDI improved prediction performance when incorporated into existing clinical[55,56] and genetic risk stratification models[57,58] for common complex diseases. In this study, by integrating large scale EHR data and geographical information, we systemically assessed the impact of SDI on Phenome-wide scale across diverse ancestries and drew several clinical implications. First, our systemic assessment suggests that although SDI is a significant contributor to a wide range of diseases, the impacts of SDI are significantly varied across disease domains. For example, while mental and cardiopulmonary diseases are more prevalent among individuals experiencing social deprivation, cancers and congenital diseases are observed almost equally, irrespective of deprivation status. Furthermore, SDI is differentially yet ubiquitously associated with a wide array of health outcomes across various genetic ancestral groups. Finally, although the effect of SDI persisted across various

genetic clusters, the varying magnitude of association suggests an interaction between social deprivation and genetic factors as previously suggested[59,60].

In addition to enabling detailed disease modeling, healthcare biobanks are unique and powerful resources for exploring rare genetic conditions or disease outcomes. First, we identified individuals carrying actionable variants, as defined by a curated database. However, these individuals are predominantly of specific European ancestries, suggesting a bias against non-Europeans in reference datasets, potentially resulting from disparities in clinical genetic testing[19,61]. Using an unbiased genomic scan, our study uncovered several significant associations, which may further refine prognosis within healthcare settings. Furthermore, we confirmed a penetrant association between an upstream variant of the *RHD* gene and Rhesus isoimmunization during pregnancy[13,62], while also clarifying varied prevalence across diverse communities. Bringing these findings together, we highlight that healthcare biobanks, compared to general population-based biobanks, are enriched with uncommon outcomes, and associated genetic variations, thereby offering an ideal environment to study clinically pertinent scenarios.

Nevertheless, our study warrants several limitations. First, most of our enrollment occurred in tertiary hospitals. While this enabled us to include patients with rare and more severe conditions, the prevalence may not reflect the general population due to inclusion bias as previously described[63]. Second, MGBB participants are centralized in the greater Boston area of Massachusetts, which reflects the geographic location of the two main hospitals of the MGB health system. Communal and geospatial characteristics are likely to vary in other New England regions and more broadly across the U.S. Moreover, while our study provides detailed insights into European populations, the resolution for non-European populations is less robust due to limited sample sizes, reflecting the demography of the included region.

In conclusion, by utilizing a population genetics, we discerned specific ancestral clusters within the MGBB. These clusters reflect the colonization histories of the Greater Boston area and exhibit distinct genetic characteristics and disease susceptibilities. Individual-level clinical and lifestyle risk factors in combination with community context, structural factors, and genetic variation advance disease modeling toward precision medicine initiatives.

# Methods

## Patient recruitment in MGBB and study protocols

MGBB, previously known as Partners Biobank, is an integrated research initiative based in Boston, Massachusetts. It collects biological samples and health data from consenting individuals at Massachusetts General Hospital, Brigham and Women's Hospital, and local healthcare sites within the MGB network[64]. This repository of samples and data supports researchers aiming to decipher disease mechanisms, enhance personalized medicine, and innovate therapeutic solutions. Since July 1st, 2010, the MGBB has enrolled 142,238 participants, and extracted DNA from 88,665 participants' samples (62.3%). All participants provided written/electronic informed consent for broad biological and genetic research. The study protocol to analyze MGBB data was approved by the Mass General Brigham Institutional Review Board under protocol number 2018P001236. The study protocols to analyze UKBB data was approved under protocol number 2021P002228 and performed under UKBB application number 7089.

## Genotype quality control and imputation

53,306 individuals were genotyped by Illumina Global Screening Array (Illumina, CA) in four batches (13,140 in the 1st batch, 11,649 in the 2nd batch, 5,976 in the 3rd batch, and 22,541 in the 4th batch). Genotypes were called using the Z-call software[65]. After genotype calling, we conducted quality control with the following steps. We re-aligned genotyping probes to the GRCh38 reference genome using the blast software[66] and extracted probes with perfect- unique match. We removed indels and multiallelic sites and removed variants with high missingness (> 2%) and low minor allele counts (≤ 2). After genotype quality control, we estimated continental level ancestry using the 1KG dataset. We extracted common, high-quality SNPs (missingness < 1%, MAF > 1%) across MGBB and the 1KG dataset. After pruning SNPs, we computed SNP weights for genetic principal component using the 1KG dataset. Then, we projected MGBB participants into the same principal component space using 10 PCs. Using genetic PCs in 1KG dataset as a feature matrix, we trained a K-nearest neighbor model for population assignment (AFR, AMR, EAS, EUR, and SAS) to assign population labels to MGBB participants. With these inferred labels, we calculated Hardy Weinberg disequilibrium for each population and removed variants with $P < 1 \times 10^{-6}$. Finally, we compared the allele frequency in these populations with gnomAD allele frequency, then removed variants with deviation from ancestry specific gnomAD allele frequency (Chi-square value > 300). These quality control procedures were done by genotyping batch. We took the intersection of the variants in these four batches and generated dataset for imputation. Using the same set of variants, we imputed the genotypes by TOPMed imputation server[67]. We used TOPMed multi-ancestry imputation reference panel (TOPMed r2 panel) including 97,256 reference samples and

308,107,085 variants. Pre-phasing was carried out by Eagle software[68], and imputation was conducted by Minimac4 software[67,69]. After the imputation, we merged all the four batches by vcftools[70] and converted to bgen file by PLINK2 software (9 Jan 2023)[71] for the downstream analysis.

**Exome sequencing and quality control**

Exome sequences were performed by on Illumina NovaSeq instruments (Illumina, CA) with a custom exome capture kit (Human Core Exome, Twist Bioscience, CA), with a target of at least 20x coverage at > 85% of target sites. Alignment, processing, and joint calling of variants were performed using the Genome Analysis ToolKit (GATK, version v4.1)[72] following GATK best practices. The joint called dataset containing all 53,420 individuals processed by Hail framework[73] for further 1) genotype, 2) variant, and 3) sample quality controls. First, we split the multi-allelic site into biallelic by split_multi_hts function. Following this process, we removed low-quality genotypes and genotypes called by unbalanced allele balance. We consider genotypes that meet the following criteria as missing: For reference homozygotes: total depth (DP) < 10, or Genotype Quality (GQ) < 20. For heterozygotes: DP < 10, Genotype Likelihood for reference homozygote (PL) < 20, (Reference depth + A1 depth)/DP < 0.8, or (A1 depth)/DP < 0.2. For alternate homozygotes: DP < 10, PL < 20, or (A1 depth)/DP < 0.8. Following genotype quality control, we conducted variant-level quality control. First, we filtered variants in the low complexity region or outside of the target region (broad_custom_exome_v1.Homo_sapiens_assembly38.bed) with 50bp flanks. We excluded i) monomorphic variants and, ii) variants with high missing rate (>10%). Using a quality-controlled variant set, we conducted sample-level quality control. We collected sample QC metric by Hail's sample_qc function. We implemented five hard filters (percent chimeric reads, percent contamination, call rate, mean depth, and mean genotyping quality, Supplemental Information Figure 8) and four soft filters (number of singletons, Ts/Tv ratio, Het/Hom variant ratio, and Insertion/Deletion ratio, Supplemental Information Figure 9). For soft filters, we obtained residuals of metrics regressing by the first ten genetic PCs and excluded +/- 4SD outliers. Finally, using only unrelated quality-controlled samples, we computed Hardy-Weinberg P-values by continental ancestry estimated from genotyping data. Hardy-Weinberg P-values in chromosome X was computed only for Female participants. We excluded variants with ancestry-wise Hardy-Weinberg *P*-values < $1 \times 10^{-6}$ or monomorphic variant. After quality control steps, 7,895,027 variants in 22 autosomes and chromosome X were found in 50,625 individuals remained.

**Relatedness inference**

We utilized the pc_relate[74] function from Hail to adjust for the presence of an admixed population within the MGBB, using 91,615 pruned, common (MAF > 1%) variants that are located outside the major

444 histocompatibility complex (chromosome 6 24,000,000 – 37,000,000 base pair). Among 53,306
445 individuals, we identified 3,147 pairs with a kinship greater than 0.0884.

446 **Derivation of genetic principal components**

447 To obtain insights utilizing reference populations, first we combined array genotypes from unrelated
448 MGBB participants with recently generated whole genome sequence datasets from ancestrally diverse
449 populations including 3,381 individuals from 1KG and HGDP[28]. We intersected 495,213 autosomal,
450 non-palindromic variants outside the high LD region with minor allele counts ≥ 10. After merging two
451 datasets, we pruned variants by PLINK2 software[71] with –indep-pairwise option 1000 100 0.2 resulting
452 in 257,754 variants. Using these genotypes, we derived the weight for each variant for PCs excluding
453 related samples. Using derived weights, we calculated 30 PCs for all the individuals from MGBB, 1KG,
454 and HGDP which were used in subsequent analysis.

455 **Fixation index**

456 Pairwise Fixation indices ($F_{ST}$) were computed among in MGBB-, 1KG- and HGDP- populations using
457 PLINK2 software. The phylogenetic tree was constructed neighbor-joining method[75] implemented by
458 ape R package[76].

459 **ADMIXTURE analysis**

460 Using PCs derived above, we conducted admixed component analysis using ADMIXTURE software
461 (version 1.3.0)[29]. We optimized the number of admix component K from 1 to 20 and found that K = 10
462 showed the least cross-validation error (Supplemental Information Figure 3).

463 **Genetic ancestral clustering**

464 To derive fine-scale genetic clusters in the population, we conducted Graph-based clustering which is
465 frequently used in single-cell RNA-seq clustering analysis implemented in Seurat software (version
466 4.1.0)[25]. Though Seurat is primarily tailored for single-cell RNA seq data analyses, we leveraged its
467 robust clustering capabilities for genetic ancestry clustering. Using the first 30 PCs derived above, we
468 constructed a nearest-neighbor graph and classified individuals into distinct clusters using the Louvain
469 algorithm, a default clustering approach in Seurat with resolution parameter of 0.2. As Seurat identified
470 the clusters in an unsupervised mode, we used individuals from the 1KG or HGDP as a “spike in”
471 positive controls (true labels).

472 **Effective population size estimation**

473 To estimate the effective population size using haplotype sharing information, we used IBDNe in
474 combination with the hap-ibd. First, we phased the genotypes of unrelated MGBB participants with

SHAPEIT software (version 4.2). Then, using hap-ibd software (version 1.0, 15Jun23.92f)[77], we calculated IBD sharing, and this output was fed into IBDNe software (version 23Apr20)[78] to determine the effective population size for each ancestry. To compare the effective population size trajectories in British population in UK and MGBB, we computed effective population size in down-sampled, unrelated UKBB European-like population to the same sample size as MGBB British-like population (n=11,508), using microarray-based genotypes.

**Variant annotation**

We annotated WES data using the VEP software (version 107)[79], supplemented with the Loftee[30] and dbNFSP[80] plugins. The "—pick" option was enabled to prioritize the canonical transcript. Additionally, in-silico predictions from dbNFSP (version 4.2) were employed to prioritize missense variants.

**Pathogenic variant annotation**

We downloaded ClinVar database[45,46] on Aug 16, 2022, and annotated all the variants identified by exome sequence using snpEff software (version 5.0e) [81]. We identified 536,729 variants registered in the ClinVar Database overall. To identify the carriers of pathogenic/likely pathogenic variants in the ACMG SF v3.1 actionable genes[43,44], we only used variants with review status "reviewed_by_expert_panel", "criteria_provided,_multiple_submitters,_no_conflicts"

**Polygenic risk score**

Using the PRS-CS software[82], we determined posterior weights for 1.2 million hapmap3 SNPs from a previous CAD GWAS[42], which did not include the MGBB/UKBB population. Given our study's predominant European population, we utilized the European reference panel provided by the PRS-CS authors. Our model training and derivation of posterior weights incorporated parameters phi ranging from $1 \times 10^{-1}$, $1 \times 10^{-2}$, $1 \times 10^{-3}$, $1 \times 10^{-4}$, and $1 \times 10^{-5}$. With these weights, we determined the CAD-PRS for both UKBB and MGBB populations. From the UKBB results, we identified the optimal parameter phi as $1 \times 1 \times 10^{-3}$ and applied this PRS in the MGBB analysis.

**Disease Phenotyping**

We obtained patient data from the electronic health record system within the MGB network. We specifically extracted ICD9 and ICD10 codes assigned to each patient. To enhance the interpretability and powered analysis of the disease outcome, we employed the PheWAS R package (version 1.2)[83] to map these codes to corresponding PheCodes[24]. The PheWAS package utilizes a comprehensive catalog of PheCodes (https://phewascatalog.org/phecodes). This mapping process facilitated a more standardized and consistent representation of the patient's conditions for subsequent analyses. To

506 determine the prevalence or incidence of diseases, we considered the date of blood draw for
507 genotyping as the reference date. By aligning with the corresponding date of PheCodes occurrences,
508 we identified the prevalent or incident outcomes related to the date of enrollment in MGBB.

509 **Clinical risk, genetic risk, and social risk for CAD**

510 For the CAD analysis, we calculated the 10-year Atherosclerotic Cardiovascular Disease (ASCVD) risk
511 scores based on the PCE using the *PooledCohort* R package[39,40,84]. The PCE accounts for sex, race,
512 age, total and HDL cholesterol, systolic blood pressure, antihypertensives prescription, current smoking,
513 and prevalence of diabetes mellitus. For missing values, we performed multiple imputation by chained
514 equations using the *mice* R package[85], using enrollment age, sex, and race as predictors. Among
515 participants without prior CAD, the first post-enrollment CAD incidence was ascertained based on
516 relevant ICD-9 and ICD10 codes from in-hospital records. We assessed the individual association of
517 10-year ASCVD risk, CAD-PRS, and SDI with incident CAD based on logistic regression.

518 **Exome-wide burden test**

519 We conducted a rare variant aggregation burden test implemented in Regenie software (version
520 3.2.5)[86]. We generated masks comprised of predicted loss of function (high confidence by Loftee
521 software[30]) and damaging missense variants predicted by > 90% of 29 in-silico prediction programs
522 included in dbNFSP (version 4.2)[80] with MAF < 0.01.

523 **Genome-wide, phenome-wide association study**

524 We conducted single variant PheWAS using imputed genotypes and 1,461 PheCodes. We applied
525 mixed model approach implemented in Regenie software (version 3.2.5)[86]. Null model was fit using
526 pruned common variants derived from microarray-derived genotypes (MAF > 1%, pruned by PLINK2
527 software[71] with option –indep-pairwise 1000 100 0.9). The analyses were conducted ancestry wise. For
528 sex specific endpoint, only male or female were included in the analysis. The genome-wide significant
529 threshold was set at $P = 1.6 \times 10^{-11}$ dividing conventional genome-wide significant threshold $5 \times 10^{-8}$ by
530 3,048 tested phenotypes across three ancestries. To define the associated loci, we added the flanking
531 region (± 500,000 base-pairs) for all the variants with genome-wide significance ($P < 1.6 \times 10^{-11}$) and
532 merged all overlapping regions.

533 **PheWAS Inflation statistics**

534 We estimated Lambda GC (observed chi-squared value divided by expected value) i) at top 1.0
535 percentile of the test statistics for rare variant burden PheWAS, and ii) using HapMap3 SNPs for

common variant PheWAS. For common variant PheWAS, LD score regression[87] was additionally performed to estimate intercept using ldscr R package (https://github.com/mglev1n/ldscr).

**Geocoding**

The participants' current address data was geocoded using the DeGAUSS framework[88], a collection of geospatial tools designed for cleaning and formatting geographic data. This process converts the address information into standardized spatial data, specifically latitude and longitude coordinates. Our analysis focused on participants residing in Massachusetts. We excluded 1) Participants whose addresses were located outside of the state of Massachusetts, 2) Participants for whom the geocoding process failed. We successfully geocoded 48,369 individuals with genotype data.

**Area-based deprivation score index**

For individuals whose addresses were successfully geocoded, we proceeded with the following steps: A) We assigned each individual's address to a corresponding U.S. Census tract. Census tracts are small, relatively stable geographic areas that are defined by the United States Census Bureau. They are designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions. B) We then merged this Census tract-level data with SDI (2018 SDI, downloaded from https://www.graham-center.org/maps-data-tools/social-deprivation-index.html)[54]. SDI is a composite measure of area level deprivation based on seven demographic characteristics collected in the American Community Survey (ACS) and used to quantify the socio-economic variation in health outcomes. The final SDI is a composite measure of seven demographic characteristics collected in the ACS: percent living in poverty, percent with less than 12 years of education, percent single-parent households, the percentage living in rented housing units, the percentage living in the overcrowded housing unit, percent of households without a car, and percentage non-employed adults under 65 years of age. This approach allows for a detailed, regional census tract-level analysis of the social conditions experienced by the study participants.

**Spatial enrichment analysis**

We utilized the Bernoulli model in SaTScan[89]. Under this model, individuals belonging to a specific genetic cluster were treated as "cases," while all other individuals were treated as "controls." This model compares the rates of cases in different areas to determine if the rate of cases inside the potential cluster area is significantly different from outside. To avoid detecting overly large and potentially less meaningful clusters, we limited our scan by setting the maximal diameter of the spatial cluster window. Specifically, we restricted this to a maximum radius of 4 kilometers.

**Data availability**

Genotyping and exome sequencing data for 13,500 participants from the MGBB are available in dbGAP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002018.v1.p1). Additional MGBB data were accessed under institutional review board protocol for this current study and are not publicly available due to restrictions on the data. The summary statistics for phenome-wide common/rare variant association analysis and the allele frequencies of genetic clusters will be publically available upon the publication.

## Acknowledgements

### Funding sources

S.K. is supported by Japan Society for the Promotion of Science (202160643), Uehara Memorial Foundation, and National Institute of Health (NIH), National Heart Lung and Blood Institute (NHLBI, K99HL169733). K.P is supported by the MGH Executive Committee for Research Fund for Medical Discovery. S.J.C. is supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (HI19C1330). S.M.U. is supported by NIH National Human Genetics Research Institute (NHGRI, T32HG010464). R.C.G. is supported by NIH (HG009922, OD026553, HL143295, TR003201). B.M.N. is supported by R37MH107649. P.T.E. is supported by grants from the NIH (1R01HL092577, 1R01HL157635, 5R01HL139731), from the American Heart Association (18SFRN34230127, 961045), and from the European Union (MAESTRIA 965286). J.W.S. is supported by OT2OD026553, U01HG008685, MH118233. M.L. is supported by grants from the NIH (OT2OD002750, R01HL143295, U01HG008685, U01TR003201, and U24HG006834). E.K. is supported by grants from the NIH (OT2OD026553, U01HG008685, P30 AR070253, OT2HL161841). A.R.M. and Y.W. are supported by NIH, NHGRI (K99/R00MH117229 to A.R.M.) and by European Union's Horizon 2020 research and innovation program under grant agreement 101016775. A.R.M. is also supported by U01HG011719. P.N. is supported by grants from the NHLBI (R01HL127564), and NHGRI (U01HG011719).

### Declaration of interests

K.P. reports research grants, paid to her institution, from Allelica, Apple, Amgen, AstraZeneca, Boston Scientific, Genentech / Roche, and Ionis. R.C.G. has received compensation for advising the following companies: Allelica, Atria, Fabric, Genome Web, Genomic Life and Juniper Genomics; and is co-founder of Genome Medical and Nurture Genomics. M.J.D is a founder of Maze Therapeutics and is a member of the scientific advisory board for Neumora Therapeutics, Inc. (formerly known as RBNC Therapeutics). B.M.N. is a member of the scientific advisory board at Deep Genomics and Neumora Therapeutics, Inc. J.W.S. is a member of the Scientific Advisory Board of Sensorium Therapeutics (with equity), and has received grant support from Biogen, Inc. P.T.E. receives sponsored research support from Bayer AG, IBM Research, Bristol Myers Squibb, Pfizer and Novo Nordisk; he has also served on advisory boards or consulted for MyoKardia and Bayer AG. J.W.S. is PI of a collaborative study of the genetics of depression and bipolar disorder sponsored by 23andMe for which 23andMe provides analysis time as in-kind support but no payments. P.N. reports research grants from Allelica, Apple, Amgen, Boston Scientific, Genentech / Roche, and Novartis, personal fees from Allelica, Apple,

AstraZeneca, Blackstone Life Sciences, Eli Lilly & Co, Foresite Labs, Genentech / Roche, GV, HeartFlow, Magnet Biomedicine, and Novartis, scientific advisory board membership of Esperion Therapeutics, Preciseli, and TenSixteen Bio, scientific co-founder of TenSixteen Bio, equity in MyOme, Preciseli, and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work.

## Author contributions

R.C.G., M.J.D., B.M.N., P.T.E., E.W.K., A.R.M., and P.N. conceptualized this project. S.K., K.P., S.J.C., and E.W.K. curated phenotype data. S.K., and M.S.L. curated genotype data. S.K., M.U., and S.M.U. analyzed data. S.K., Y.W., P.T.E., E.W.K., A.R.M., and P.N. interpreted data. S.K., A.R.M., and P.N. prepared the initial draft. Y.W., K.P., M.U., S.J.C., W.E.H., R.C.G., M.J.D., B.M.N., P.T.E., J.W.S., E.W.K., A.R.M., and P.N. provided critical review and edits for the manuscript. S.H., W.E.H., R.C.G., M.J.D., B.M.N., P.T.E., E.W.K., A.R.M., and P.N. supervised the project. S.H., W.E.H., P.T.E., J.W.S., M.S.L., and E.W.K. managed the project administration. P.T.E., J.W.S., M.S.L., E.W.K., and P.N. obtained funding for the project.

## Figure Legends

### Figure 1 | The cohort characteristics of MGBB

**a,** The columns represent the cumulative number of individuals who have consented to the MGBB. Colors indicate the vital status of participants as of May 2023. **b,** Gender/Age at Consent Distribution: The columns represent the distribution of the participants based on gender and age at the time of consent. The individuals older than 100 years at the time of consent are not included in the displayed numbers. **c,** Encounter patterns in the MGB Network. The number of encounters that participants have had within the MGB network. Please note that these encounters include sites where recruitment did not take place. **d,** Distribution of PheCodes based outcomes. The columns indicate the number of outcomes in PheCode-category. Colors distinguish between incident and prevalent cases. MGH, Massachusetts General Hospital; BWH, Brigham's and Women's Hospital, MEEI Mass Eye and Ear Institute; FH Faulkner Hospital; NWH, Newton-Wellesley Hospital; DFCI, Dana-Farber Cancer Institute.

### Figure 2 | Fine-scale genetic clusters within MGBB

**a,** and **b,** UMAP representation of genetic clusters in MGBB. Each dot represents a participant, with colors indicating distinct genetic clusters identified through graph-based clustering from genetic principal components (Methods). The numbers indicate cluster identification. The color legend and detailed cluster information will be found in Supplemental Table 2 and Figure S2. **c,** Population differentiation in MGBB revealed by ADMIXTURE analysis. The heatmap displays the proportions of ADMIXTURE components (K = 10) within each genetic cluster. The columns at the top of the heatmap represent the number of MGBB participants in each cluster. UMAP, Uniform Manifold Approximation and Projection; MGBB, Mass General Brigham Biobank

### Figure 3 | Geospatial distribution, socioeconomic status, and disease risks in MGBB

**a,** Geographical enrichment of genetic ancestries in the greater Boston area. The circle indicates area of significant enrichment of corresponding genetic ancestries. **b,** We tested the association between the socioeconomic deprivation index (SDI) and 1,564 PheCodes based outcomes (Prevalence + Incidence) in 47,070 MGBB participants. The model was adjusted for age, sex, and the first ten genetic principal components. An association was considered statistically significant if *P*-value was less $3.2 \times 10^{-5}$ (0.05/1,564). The color of bars indicated the direction of the effect of SDI (higher SDI suggests higher deprivation). **c,** The disease frequency (prevalence + incidence) by deprivation status and genetic ancestry. The color of bars indicated deprivation status (higher or lower than the median SDI). Complication of Birth, Other and unspecified complications of birth; puerperium affecting management of mother (PheCode 654).

### Figure 4 | Rare variant identification in the MGBB

**a,** Distribution of the number of protein-truncating variants across genetic ancestries. Each dot represents a participant. The horizontal axis represents the genetic ancestries in the MGBB, while the vertical axis represents the number of protein-truncating alleles in each participant. **b,** Carrier counts of pathogenic/likely pathogenic variants in ACMG actionable genes. The colors of the bars indicate the mode of inheritance of the genes. **c,** Carrier frequency of individuals with pathogenic variants in ACMG actionable genes, categorized by ancestry. The colors of the bars correspond to the continental ancestries. The dotted line represents the average frequency in the MGBB. **d,** Summary of phenome-wide gene burden testing in the MGBB. We conducted exome-wide phenome wide association analysis

across 1,454 PheCodes based outcomes in 14,912 genes. The columns indicate the number of significant associations ($P < 1.8 \times 10^{-9} = 5 \times 10^{-2}$/28,035,307 phenotype-transcript pairs) for designated genes. The color of each column corresponds to the associated PheCode-category. MGBB, Mass General Brigham Biobank; ACMG, American College of Medical Genetics and Genomics; pLOF, predicted loss of function; AFR, African; AMR, Admixed-American; EAS, East Asian; EUR, European; SAS, South Asian.

**Figure 5 | Common variant association study in the MGBB**

**a,** Number of associations in the common variant phenome-wide association analysis in MGBB. We conducted associations between common genetic variants (Minor allele counts ≥ 40) and 1,461 PheCodes (Case counts ≥ 60), categorized by continental ancestries (AFR, n = 2,846; AMR, n = 3,756; EUR, n = 44,163). The columns represent the number of significant associations ($P < 1.6 \times 10^{-11}$) on each chromosome. The color indicates the ancestry in which the association was observed. We annotated chromosomes with more than 10 associations, indicating the representative locus in the chromosome. **b,** Manhattan plot of GWAS for Rhesus isoimmunization during pregnancy in women (n = 23,959). The horizontal axis displays the genomic coordinates from chromosome 1 to chromosome X. The vertical axis represents the strength of association in negative $\log_{10}$ *P*-value. The significantly associated variants in the *RHD* locus is highlighted. **c,** Odds ratio for Rhesus isoimmunization during pregnancy by rs72660908 genotypes. The dots and error bars represent the estimated odds ratios and 95% confidence intervals compared to the reference homozygotes ([C/C]). **d,** *RHD* read counts from Whole Blood RNA sequence data obtained from the GTEx dataset. The horizontal axis displays the number of reads aligned to the *RHD* gene, categorized by rs72660908 genotypes. **e,** Frequencies of rs72660908 homozygotes across genetic clusters in MGBB. The horizontal axis corresponds to the genetic ancestries in MGBB, while the vertical axis represents the ancestral frequency of rs72660908 homozygotes ([G/G]). The colors of the columns correspond to continental ancestries. MGBB, Mass General Brigham Biobank; ACMG, American College of Medical Genetics and Genomics; pLOF, predicted loss of function; AFR, African; AMR, Admixed-American; EAS, East Asian; EUR, European; SAS, South Asian; GWAS, Genome-Wide Association Study; GTEx, Genotype-Tissue Expression.

# Supplemental Figure Legends

## Figure S1 | Projection of the fine genetic clusters in the continental ancestry space

**a,** The distribution of the continental ancestries in the UMAP space. Each point on the plot represents an individual participant of MGBB. To infer the continental ancestry of each participant, we utilized the K-nearest neighbor algorithm trained on the 1000 Genomes Project dataset. The colors assigned to the points represent the inferred continental ancestries. The horizontal axis corresponds to the first axis of the UMAP projection, and the vertical axis represents the second axis. This two-dimensional representation allows us to visualize the clustering and distribution patterns of different continental ancestries within the MGBB population. **b,** The distribution of the fine-scale genetic clusters in the MGBB on conventional PC space. Each point on the plot represents an individual in MGBB. The colors assigned to the points represent genetic clusters inferred by the network-based clustering method (Methods). The horizontal axis corresponds to the first genetic PC, and the vertical axis represents the second genetic PC. UMAP, Uniform Manifold Approximation and Projection; MGBB, Mass General Brigham Biobank; PC, principal component.

## Figure S2 | Enrichment of the reference groups in the genetic clusters inferred in MGBB

The horizontal axis shows reference populations from 1000 Genomes Project and Human Genome Diversity Project. The vertical axis shows genetic clusters inferred from MGBB. In the left panel, the heights of bars show the number of MGBB participants included in the genetic clusters. In the right panel, the size and transparency of rectangle shows intersection size between reference population and genetic clusters. The number of individuals in the clusters and the color legend are found in Supplemental Table 2. Abbreviations for reference populations will be find in Supplemental Table 7. MGBB, Mass General Brigham Biobank.

## Figure S3 | ADMIXTURE analysis for reference population (1KG + HGDP) and genetic clusters

**a,** Phylogenetic tree generated using *F*st Values. The phylogenic tree was constructed using the Neighbor-Joining method with pairwise *F*st values serving as a measure of genetic distance between populations. A higher *F*st value indicates a greater genetic differentiation between populations. The numeric numbers indicate genetic cluster in MGBB. Ancestry names indicated reference populations in 1KG or HGDP. The number indicate genetic clusters in MGBB. **b,** In each panel, a stacked column of color segments represents an individual participant. The color of each segment corresponds to one of the K = 10 ancestral components, as determined by ADMIXTURE software. The length of each colored segment within a participant's column indicates the estimated proportion of their genome attributed to that specific ancestral population. The choice of K = 10 was informed by cross-validation results, with the aim of minimizing prediction error (Supplemental Information Figure 3). The abbreviations and detailed descriptions for the ancestral populations corresponding to each color are provided in Supplemental Table 7. 1KG, 1000 Genomes Project; HGDP, Human Genome Diversity Project; MGBB, Mass General Brigham Biobank.

## Figure S4 | Phenome wide association analysis for cluster 4

Each dot represents the association results based on PheCodes. We examined the relationship between membership in genetic cluster 4 of the MGBB biobank and PheCodes outcomes. This association was assessed using a logistic regression model, adjusted for age and sex. MGBB, Mass General Brigham Biobank.

**Figure S5 | Shared founder mutations across genetic clusters 4, 11, and 14**

The left panels are showing the distribution of genetic cluster 4, 11, and 14 in the UMAP space. The right panel shows the distribution of Ashkenazi founder mutations described in the previous literatures. The horizontal axes show the genetic clusters identified in MGBB, and the vertical axes show the allele frequencies determined by whole exome sequencing of MGBB by the genetic clusters. UMAP, Uniform Manifold Approximation and Projection; MGBB, Mass General Brigham Biobank.

**Figure S6 | Effective population sizes in the genetic clusters in MGBB**

The horizontal axes show generations ago. The vertical axes show the estimated population size. Numbers on the top of panels show the cluster identification. The black lines show the estimates and gray lines show 95% confidence interval.

**Figure S7 | Geographical, socioeconomical, distributions of genetic clusters**

Distributions of socioeconomic (**a**), clinical (**b**), and genetic risks (**c**). Each dot indicates MGBB participants. The horizontal axes show genetic cluster. SDI, social deprivation index; PCE, pooled cohort equation, CAD PRS, polygenic risk score for coronary artery disease.

**Figure S8 | Social deprivation in Massachusetts and Greater Boston Area**

The color of each grid represents the median Social Deprivation Index (SDI) of the participants from MGBB. A darker shade denotes a higher level of socioeconomic deprivation. Grids with fewer than five participants have been excluded.

**Figure S9 | Different annotation rate for the functional variants in the pathogenic genes**

**a,** The bar heights depict the proportion of pLOF variants in ACMG genes within specific genetic clusters identified in MGBB with pathogenic/likely pathogenic annotations as determined by multiple expert reviews in ClinVar. **b,** The odds ratio indicates the likelihood of pLOF variants in ACMG genes having pathogenic/likely pathogenic annotations based on multiple expert reviews. The results are presented in the continental ancestries and in reference to the European population. Error bars indicate 95% confidence intervals. OR, Odds Ratio; AFR, African; AMR, Admixed American; EAS, East Asian; EUR, European; SAS, South Asian. ACMG, American College of Medical Genetics and Genomics; MGBB, Mass General Brigham Biobank; pLOF predicted loss of function.

**Figure S10 | Penetrant association of rs334 for sickle cell anemia**

**a,** The prevalence of Sickle cell anemia by rs334 genotypes. **b,** The RBC measurements in the MGB participants by combined genotype of rs33930165 and rs334. The box plot depicts the first and third quartiles, with the line inside the box indicating the median value. RBC, Red Blood Cell count.

## References

1. Townsend, P., Phillimore, P., and Beattie, A. (1988). Health and Deprivation: Inequality and the North (Routledge).

2. Li, X., Memarian, E., Sundquist, J., Zöller, B., and Sundquist, K. (2014). Neighbourhood Deprivation, Individual-Level Familial and Socio-Demographic Factors and Diagnosed Childhood Obesity: A Nationwide Multilevel Study from Sweden. Obesity Facts *7*, 253-263. 10.1159/000365955.

3. Bann, D., Johnson, W., Li, L., Kuh, D., and Hardy, R. (2018). Socioeconomic inequalities in childhood and adolescent body-mass index, weight, and height from 1953 to 2015: an analysis of four longitudinal, observational, British birth cohort studies. Lancet Public Health *3*, e194-e203. 10.1016/S2468-2667(18)30045-8.

4. Fan, J.X., Wen, M., and Li, K. (2020). Associations between obesity and neighborhood socioeconomic status: Variations by gender and family income status. SSM Popul Health *10*, 100529. 10.1016/j.ssmph.2019.100529.

5. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat Genet *49*, 1458-1467. 10.1038/ng.3951.

6. 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. 10.1038/s41586-018-0579-z.

7. Zhou, W., Kanai, M., Wu, K.H., Rasheed, H., Tsuo, K., Hirbo, J.B., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., et al. (2022). Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genom *2*, 100192. 10.1016/j.xgen.2022.100192.

8. Kurki, M.I., Karjalainen, J., Palta, P., Sipila, T.P., Kristiansson, K., Donner, K.M., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., et al. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. Nature *613*, 508-518. 10.1038/s41586-022-05473-8.

9. Belbin, G.M., Cullina, S., Wenric, S., Soper, E.R., Glicksberg, B.S., Torre, D., Moscati, A., Wojcik, G.L., Shemirani, R., Beckmann, N.D., et al. (2021). Toward a fine-scale population health monitoring system. Cell *184*, 2068-2083 e2011. 10.1016/j.cell.2021.03.034.

10. Verma, A., Damrauer, S.M., Naseer, N., Weaver, J., Kripke, C.M., Guare, L., Sirugo, G., Kember, R.L., Drivas, T.G., Dudek, S.M., et al. (2022). The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. J Pers Med *12*. 10.3390/jpm12121974.

11. Brumpton, B.M., Graham, S., Surakka, I., Skogholt, A.H., Loset, M., Fritsche, L.G., Wolford, B., Zhou, W., Nielsen, J.B., Holmen, O.L., et al. (2022). The HUNT study: A population-based cohort for genetic research. Cell Genom *2*, 100193. 10.1016/j.xgen.2022.100193.

12. Johnson, R., Ding, Y., Bhattacharya, A., Knyazev, S., Chiu, A., Lajonchere, C., Geschwind, D.H., and Pasaniuc, B. (2023). The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank. Cell Genom *3*, 100243. 10.1016/j.xgen.2022.100243.

13. Zawistowski, M., Fritsche, L.G., Pandit, A., Vanderwerff, B., Patil, S., Schmidt, E.M., VandeHaar, P., Willer, C.J., Brummett, C.M., Kheterpal, S., et al. (2023). The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. Cell Genom *3*, 100257. 10.1016/j.xgen.2023.100257.

14. Soria, L.F., Ludwig, E.H., Clarke, H.R., Vega, G.L., Grundy, S.M., and McCarthy, B.J. (1989). Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. Proceedings of the National Academy of Sciences *86*, 587-591. 10.1073/pnas.86.2.587.

15. Shen, H., Damcott, C.M., Rampersaud, E., Pollin, T.I., Horenstein, R.B., McArdle, P.F., Peyser, P.A., Bielak, L.F., Post, W.S., Chang, Y.-P.C., et al. (2010). Familial Defective Apolipoprotein B-100 and Increased Low-Density Lipoprotein Cholesterol and Coronary Artery Calcification in the Old Order Amish. Archives of Internal Medicine *170*. 10.1001/archinternmed.2010.384.

16. Cohen, J.C., Boerwinkle, E., Mosley, T.H., and Hobbs, H.H. (2006). Sequence Variations in*PCSK9,* Low LDL, and Protection against Coronary Heart Disease. New England Journal of Medicine *354*, 1264-1272. 10.1056/nejmoa054013.

17. Cohen, J., Pertsemlidis, A., Kotowski, I.K., Graham, R., Garcia, C.K., and Hobbs, H.H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nature Genetics *37*, 161-165. 10.1038/ng1509.

18. Luzzatto, L., Ally, M., and Notaro, R. (2020). Glucose-6-phosphate dehydrogenase deficiency. Blood *136*, 1225-1240. 10.1182/blood.2019000944.

19. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic Misdiagnoses and the Potential for Health Disparities. New England Journal of Medicine *375*, 655-665. 10.1056/nejmsa1507092.

20. Aragam, K.G., Dobbyn, A., Judy, R., Chaffin, M., Chaudhary, K., Hindy, G., Cagan, A., Finneran, P., Weng, L.C., Loos, R.J.F., et al. (2020). Limitations of Contemporary Guidelines for Managing Patients at High Genetic Risk of Coronary Artery Disease. J Am Coll Cardiol *75*, 2769-2780. 10.1016/j.jacc.2020.04.027.

21. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J., Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., and Gravel, S. (2016). The Great Migration and African-American Genomic Diversity. PLOS Genetics *12*, e1006059. 10.1371/journal.pgen.1006059.

22. 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. 10.1038/ncomms14238.

23. Dai, C.L., Vazifeh, M.M., Yeang, C.-H., Tachet, R., Wells, R.S., Vilar, M.G., Daly, M.J., Ratti, C., and Martin, A.R. (2020). Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis. The American Journal of Human Genetics *106*, 371-388. 10.1016/j.ajhg.2020.02.002.

24. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol *31*, 1102-1110. 10.1038/nbt.2749.

25. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587 e3529. 10.1016/j.cell.2021.04.048.

26. Bergstrom, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*. 10.1126/science.aay5012.

27. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. Nature *526*, 68-74. 10.1038/nature15393.

28. Koenig, Z., Yohannes, M.T., Nkambule, L.L., Goodrich, J.K., Kim, H.A., Zhao, X., Wilson, M.W., Tiao, G., Hao, S.P., Sahakian, N., et al. (2023). A harmonized public resource of deeply sequenced diverse human genomes. Cold Spring Harbor Laboratory.

29. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res *19*, 1655-1664. 10.1101/gr.094052.109.

30. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443. 10.1038/s41586-020-2308-7.

31. Laken, S.J., Petersen, G.M., Gruber, S.B., Oddoux, C., Ostrer, H., Giardiello, F.M., Hamilton, S.R., Hampel, H., Markowitz, A., Klimstra, D., et al. (1997). Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. Nature Genetics *17*, 79-83. 10.1038/ng0997-79.

32. Levy-Lahad, E., Catane, R., Eisenberg, S., Kaufman, B., Hornreich, G., Lishinsky, E., Shohat, M., Weber, B.L., Beller, U., Lahad, A., and Halle, D. (1997). Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families. Am J Hum Genet *60*, 1059-1067.

33. Salmela, E., Lappalainen, T., Fransson, I., Andersen, P.M., Dahlman-Wright, K., Fiebig, A., Sistonen, P., Savontaus, M.L., Schreiber, S., Kere, J., and Lahermo, P. (2008). Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. PLoS One *3*, e3519. 10.1371/journal.pone.0003519.

34. Huyghe, J.R., Fransen, E., Hannula, S., Van Laer, L., Van Eyken, E., Maki-Torkko, E., Aikio, P., Sorri, M., Huentelman, M.J., and Van Camp, G. (2011). A genome-wide analysis of population structure in the Finnish Saami with implications for genetic association studies. Eur J Hum Genet *19*, 347-352. 10.1038/ejhg.2010.179.

35. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409-413. 10.1038/nature13673.

36. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207-211. 10.1038/nature14317.

37. Carmi, S., Hui, K.Y., Kochav, E., Liu, X., Xue, J., Grady, F., Guha, S., Upadhyay, K., Ben-Avraham, D., Mukherjee, S., et al. (2014). Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nature Communications *5*, 4835. 10.1038/ncomms5835.

38. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet *81*, 1084-1097. 10.1086/521987.

39. 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. 10.1371/journal.pgen.1007385.

40. Goff, D.C., Lloyd-Jones, D.M., Bennett, G., Coady, S., D'Agostino, R.B., Gibbons, R., Greenland, P., Lackland, D.T., Levy, D., O'Donnell, C.J., et al. (2014). 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Journal of the American College of Cardiology *63*, 2935-2959. https://doi.org/10.1016/j.jacc.2013.11.005.

41. Goff, D.C., Lloyd-Jones, D.M., Bennett, G., Coady, S., D'Agostino, R.B., Gibbons, R., Greenland, P., Lackland, D.T., Levy, D., O'Donnell, C.J., et al. (2014). 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. Circulation *129*, S49-S73. 10.1161/01.cir.0000437741.48606.98.

42. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. (2015). Nature Genetics *47*, 1121-1130. 10.1038/ng.3396.

43. Miller, D.T., Lee, K., Abul-Husn, N.S., Amendola, L.M., Brothers, K., Chung, W.K., Gollob, M.H., Gordon, A.S., Harrison, S.M., Hershberger, R.E., et al. (2022). ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Genetics in Medicine *24*, 1407-1414. 10.1016/j.gim.2022.04.006.

44. Miller, D.T., Lee, K., Chung, W.K., Gordon, A.S., Herman, G.E., Klein, T.E., Stewart, D.R., Amendola, L.M., Adelman, K., Bale, S.J., et al. (2021). ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med *23*, 1381-1390. 10.1038/s41436-021-01172-3.

45. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res *46*, D1062-D1067. 10.1093/nar/gkx1153.

46. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res *42*, D980-985. 10.1093/nar/gkt1113.

47. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature *599*, 628-634. 10.1038/s41586-021-04103-z.

48. Jurgens, S.J., Choi, S.H., Morrill, V.N., Chaffin, M., Pirruccello, J.P., Halford, J.L., Weng, L.C., Nauffal, V., Roselli, C., Hall, A.W., et al. (2022). Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. Nat Genet *54*, 240-250. 10.1038/s41588-021-01011-w.

49. Pilarski, R. (2019). PTEN Hamartoma Tumor Syndrome: A Clinical Overview. Cancers (Basel) *11*. 10.3390/cancers11060844.

50. Tischkowitz, M., Colas, C., Pouwels, S., Hoogerbrugge, N., Group, P.G.D., and European Reference Network, G. (2020). Cancer Surveillance Guideline for individuals with PTEN hamartoma tumour syndrome. Eur J Hum Genet *28*, 1387-1393. 10.1038/s41431-020-0651-7.

51. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290-299. 10.1038/s41586-021-03205-y.

52. Huang, L., Rosen, J.D., Sun, Q., Chen, J., Wheeler, M.M., Zhou, Y., Min, Y.-I., Kooperberg, C., Conomos, M.P., Stilp, A.M., et al. (2022). TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome sequence data. The American Journal of Human Genetics *109*, 1175-1181. 10.1016/j.ajhg.2022.04.006.

53. Genetic effects on gene expression across human tissues. (2017). Nature *550*, 204-213. 10.1038/nature24277.

54. Butler, D.C., Petterson, S., Phillips, R.L., and Bazemore, A.W. (2013). Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. Health Serv Res *48*, 539-559. 10.1111/j.1475-6773.2012.01449.x.

55. Kimenai, D.M., Pirondini, L., Gregson, J., Prieto, D., Pocock, S.J., Perel, P., Hamilton, T., Welsh, P., Campbell, A., Porteous, D.J., et al. (2022). Socioeconomic Deprivation: An Important, Largely Unrecognized Risk Factor in Primary Prevention of Cardiovascular Disease. Circulation *146*, 240-248. 10.1161/CIRCULATIONAHA.122.060042.

56. Schultz, W.M., Kelli, H.M., Lisko, J.C., Varghese, T., Shen, J., Sandesara, P., Quyyumi, A.A., Taylor, H.A., Gulati, M., Harold, J.G., et al. (2018). Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. Circulation *137*, 2166-2178. 10.1161/CIRCULATIONAHA.117.029652.

57. Bann, D., Wright, L., Hardy, R., Williams, D.M., and Davies, N.M. (2022). Polygenic and socioeconomic risk for high body mass index: 69 years of follow-up across life. PLoS Genet *18*, e1010233. 10.1371/journal.pgen.1010233.

58. Cromer, S.J., Lakhani, C.M., Mercader, J.M., Majarian, T.D., Schroeder, P., Cole, J.B., Florez, J.C., Patel, C.J., Manning, A.K., Burnett-Bowie, S.M., Merino, J., and Udler, M.S. (2023). Association and Interaction of Genetics and Area-Level Socioeconomic Factors on the Prevalence of Type 2 Diabetes and Obesity. Diabetes Care *46*, 944-952. 10.2337/dc22-1954.

59. Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. Elife *9*. 10.7554/eLife.48376.

60. He, Y., Qian, D.C., Diao, J.A., Cho, M.H., Silverman, E.K., Gusev, A., Manrai, A.K., Martin, A.R., and Patel, C.J. (2023). Prediction and stratification of longitudinal risk for chronic obstructive pulmonary disease across smoking behaviors. medRxiv. 10.1101/2023.04.04.23288086.

61. 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. 10.1038/s41588-019-0379-x.

62. Flegel, W.A. (2007). The genetics of the Rhesus blood group system. Blood Transfus *5*, 50-57. 10.2450/2007.0011-07.

63. Lee, Y.H., Thaweethai, T., Sheu, Y.H., Feng, Y.A., Karlson, E.W., Ge, T., Kraft, P., and Smoller, J.W. (2023). Impact of selection bias on polygenic risk score estimates in healthcare settings. Psychol Med, 1-11. 10.1017/s0033291723001186.

64. Boutin, N., Mathieu, K., Hoffnagle, A., Allen, N., Castro, V., Morash, M., O'Rourke, P., Hohmann, E., Herring, N., Bry, L., et al. (2016). Implementation of Electronic Consent at a Biobank: An Opportunity for Precision Medicine Research. Journal of Personalized Medicine *6*, 17. 10.3390/jpm6020017.

65. Goldstein, J.I., Crenshaw, A., Carey, J., Grant, G.B., Maguire, J., Fromer, M., O'Dushlaine, C., Moran, J.L., Chambert, K., Stevens, C., et al. (2012). zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinformatics *28*, 2543-2545. 10.1093/bioinformatics/bts479.

66. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410. 10.1016/s0022-2836(05)80360-2.

67. Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat Genet *48*, 1284-1287. 10.1038/ng.3656.

68. Loh, P.R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet *48*, 811-816. 10.1038/ng.3571.

69. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nature Genetics *48*, 1284-1287. 10.1038/ng.3656.

70. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158. 10.1093/bioinformatics/btr330.

71. 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. 10.1186/s13742-015-0047-8.

72. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a

MapReduce framework for analyzing next-generation DNA sequencing data. Genome research *20*, 1297-1303. 10.1101/gr.107524.110.

73. The Hail Team. Hail. https://github.com/hail-is/hail.

74. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free Estimation of Recent Genetic Relatedness. Am J Hum Genet *98*, 127-148. 10.1016/j.ajhg.2015.11.022.

75. Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution *4*, 406-425. 10.1093/oxfordjournals.molbev.a040454.

76. Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526-528. 10.1093/bioinformatics/bty633.

77. Zhou, Y., Browning, S.R., and Browning, B.L. (2020). A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. The American Journal of Human Genetics *106*, 426-437. 10.1016/j.ajhg.2020.02.010.

78. Browning, S.R., and Browning, B.L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. The American Journal of Human Genetics *97*, 404-418. 10.1016/j.ajhg.2015.07.012.

79. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol *17*, 122. 10.1186/s13059-016-0974-4.

80. Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med *12*, 103. 10.1186/s13073-020-00803-9.

81. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92. 10.4161/fly.19695.

82. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun *10*, 1776. 10.1038/s41467-019-09718-5.

83. Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics *30*, 2375-2376. 10.1093/bioinformatics/btu197.

84. Yadlowsky, S., Hayward, R.A., Sussman, J.B., McClelland, R.L., Min, Y.-I., and Basu, S. (2018). Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. Annals of Internal Medicine *169*, 20-29. 10.7326/M17-3011.

85. van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software *45*, 1 - 67. 10.18637/jss.v045.i03.

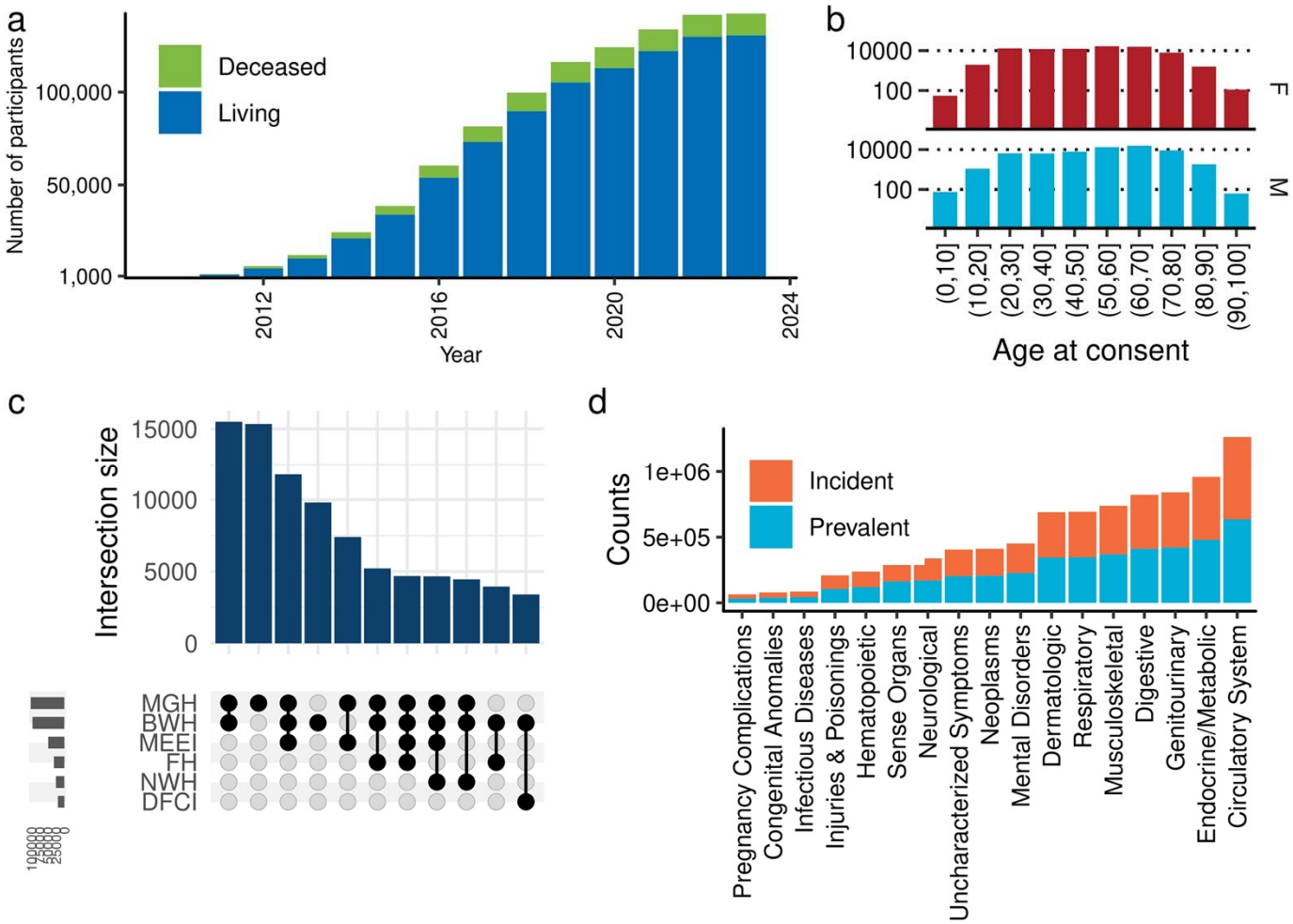
86. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet *53*, 1097-1103. 10.1038/s41588-021-00870-7.

87. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics *47*, 291-295. 10.1038/ng.3211.

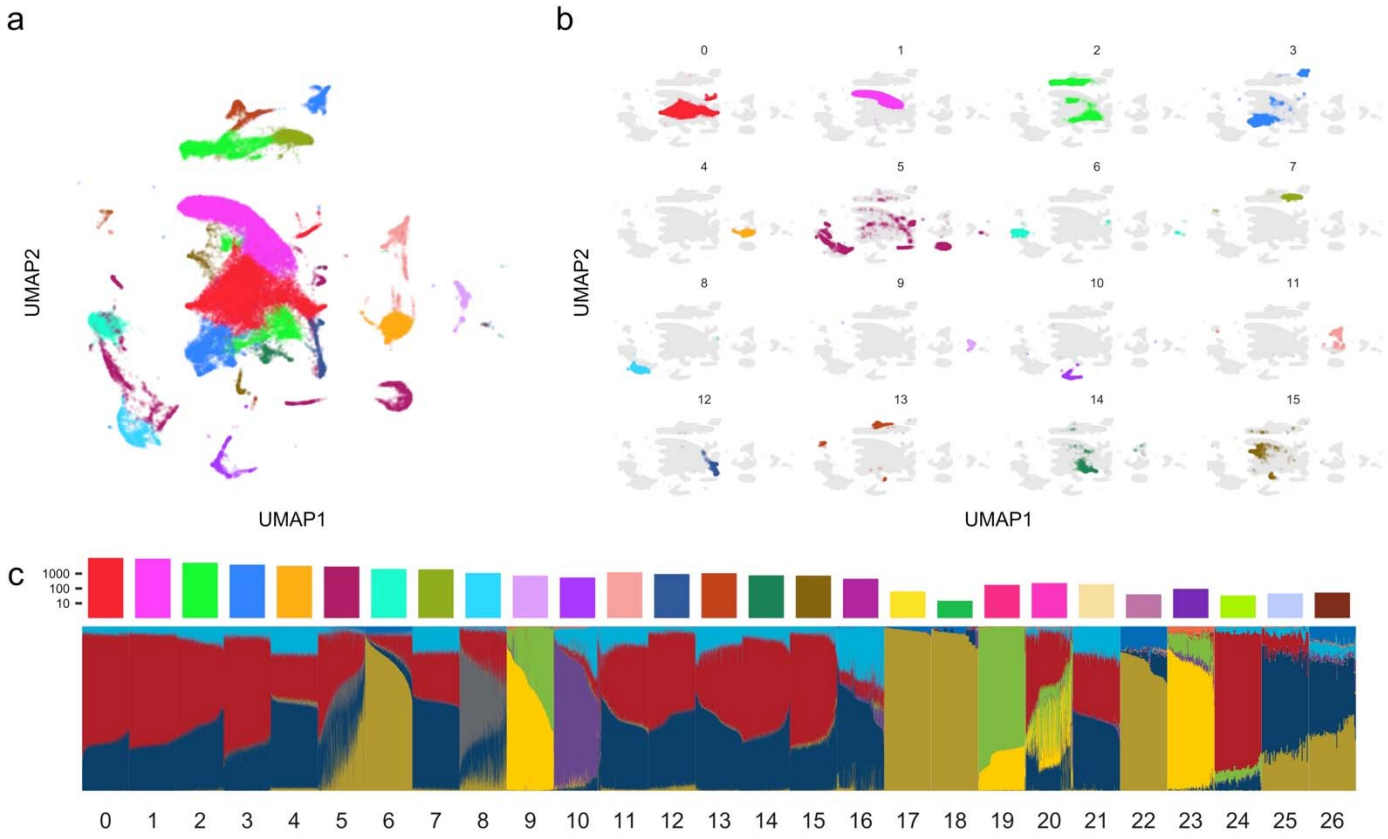
88. Brokamp, C., Wolfe, C., Lingren, T., Harley, J., and Ryan, P. (2018). Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies. J Am Med Inform Assoc *25*, 309-314. 10.1093/jamia/ocx128.

89. Kulldorff, M. (1997). A spatial scan statistic. Communications in Statistics - Theory and Methods *26*, 1481-1496. 10.1080/03610929708831995.

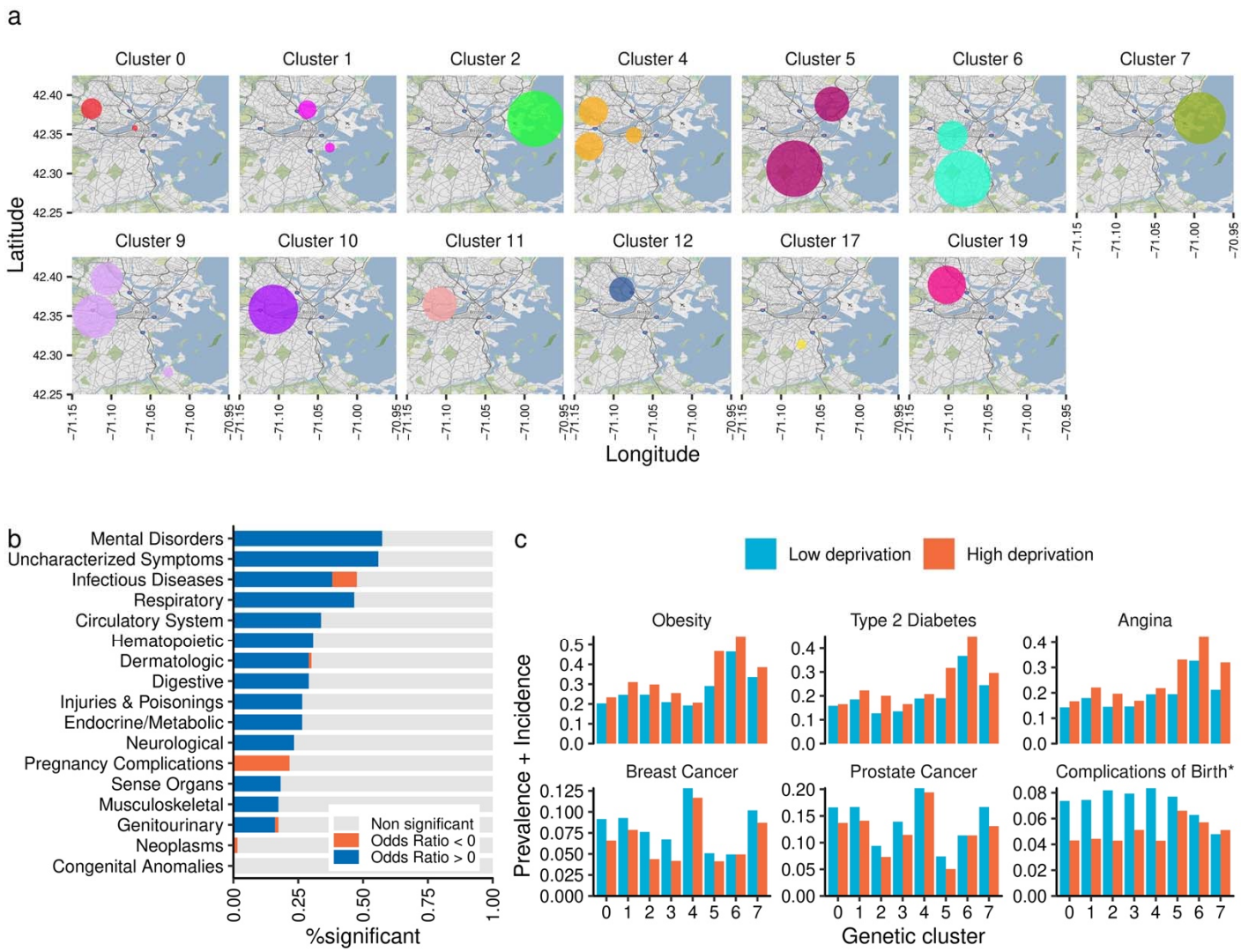
Figure 1



# Figure 2



# Figure 3



# Figure 4

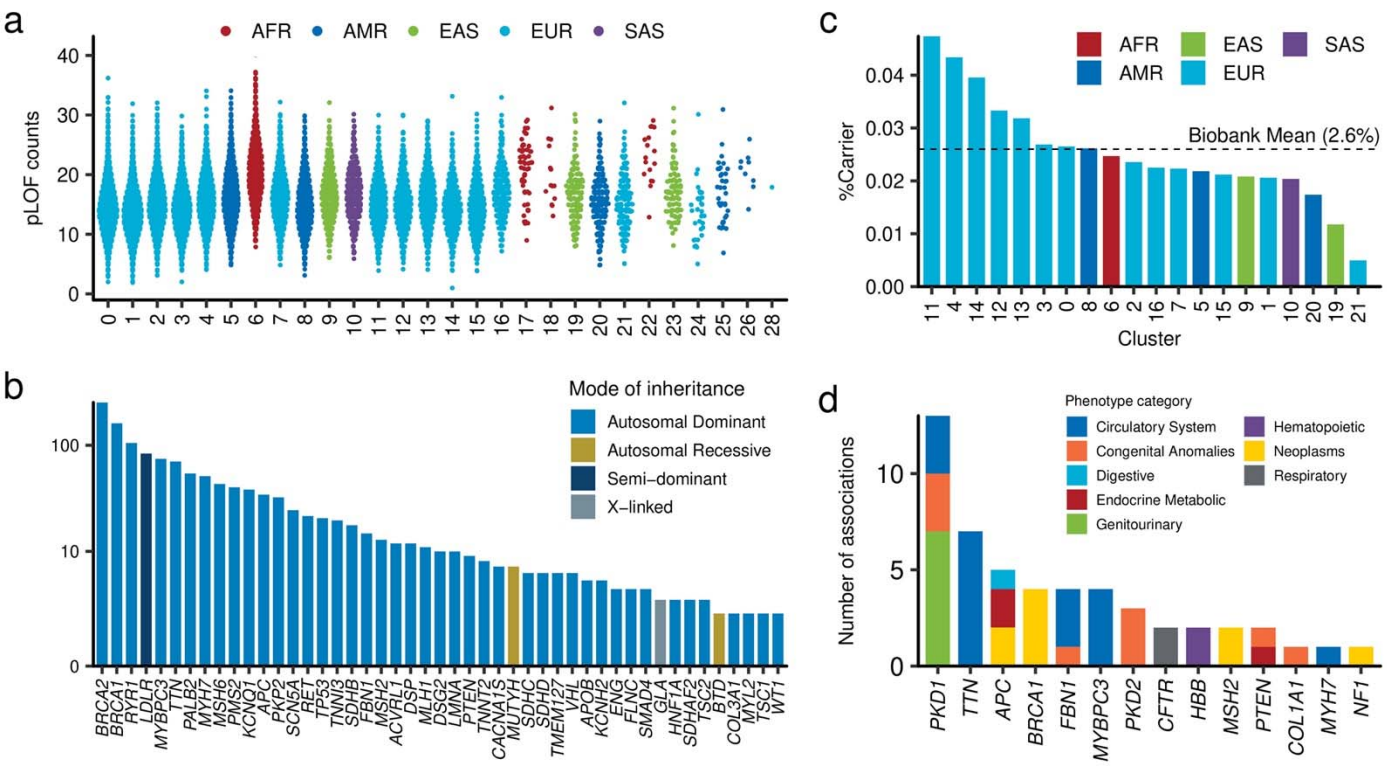


Figure 5

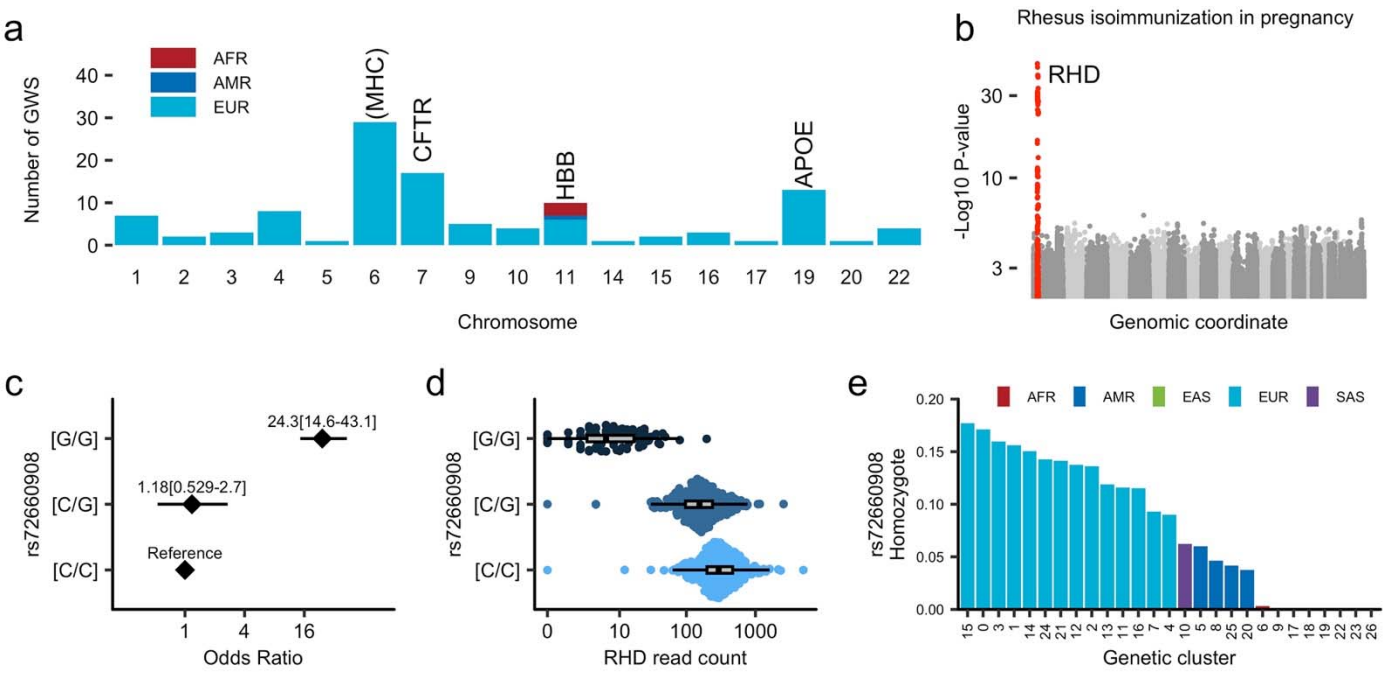
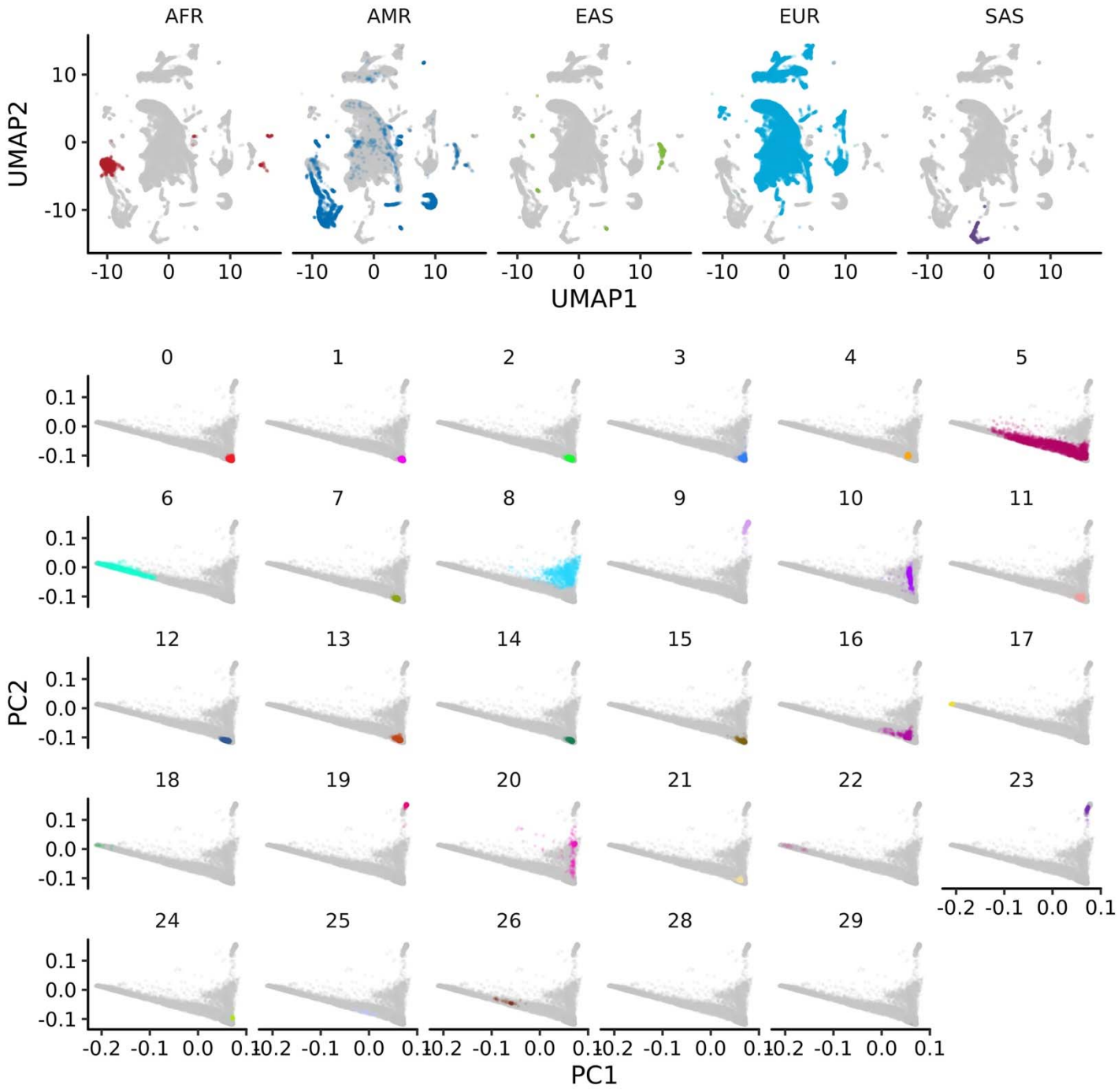


Figure S1



# Figure S2

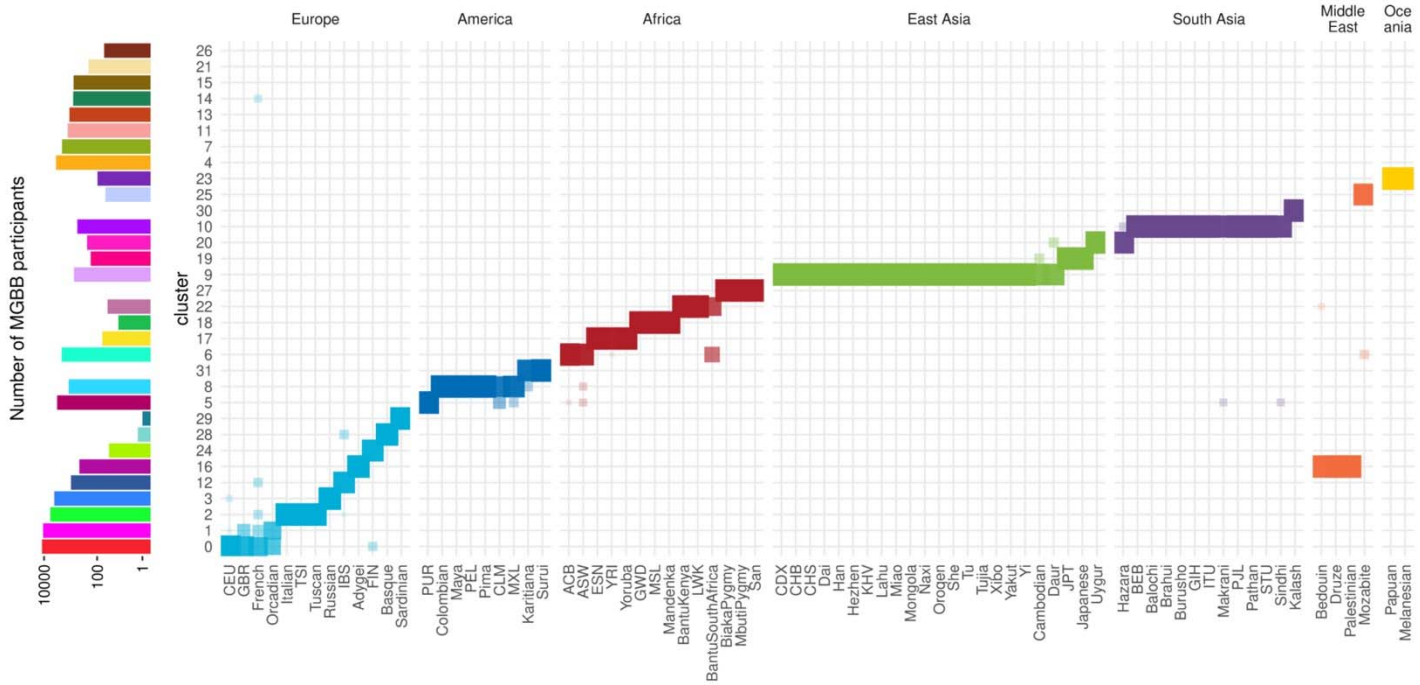
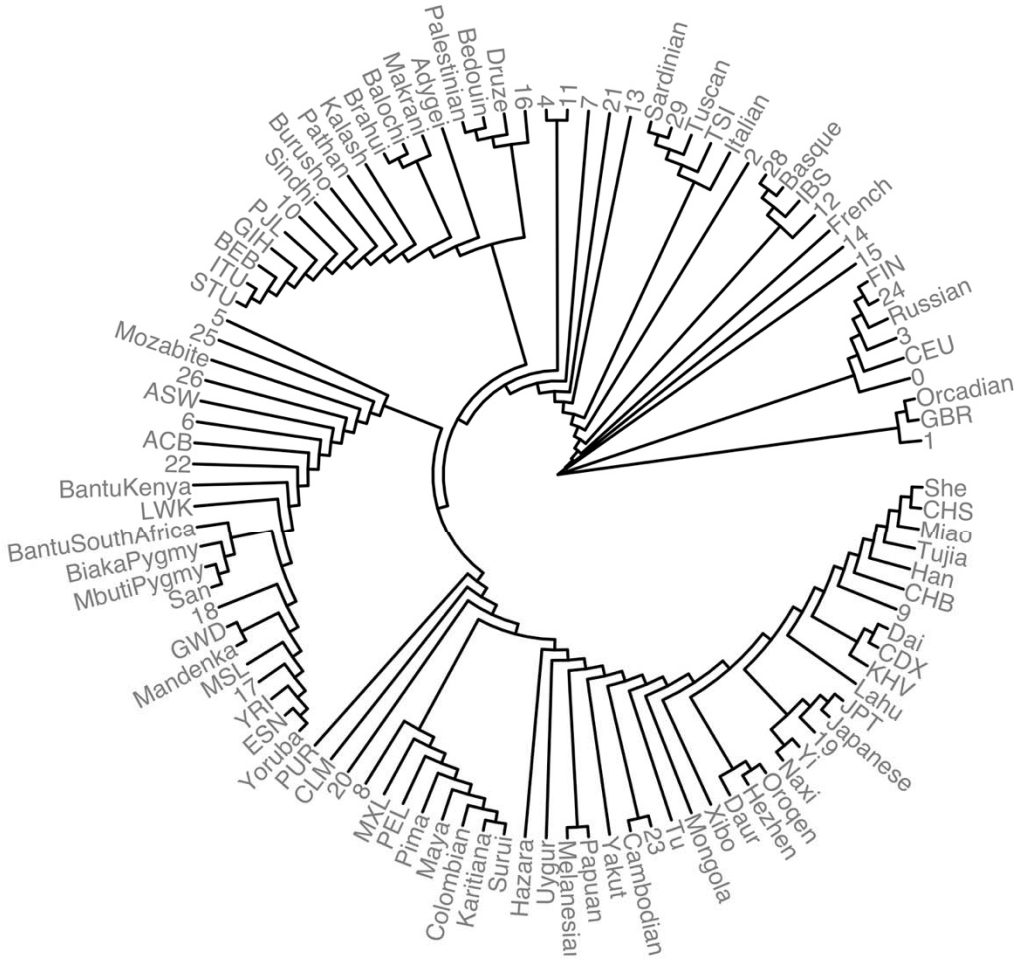


Figure S3

a




b

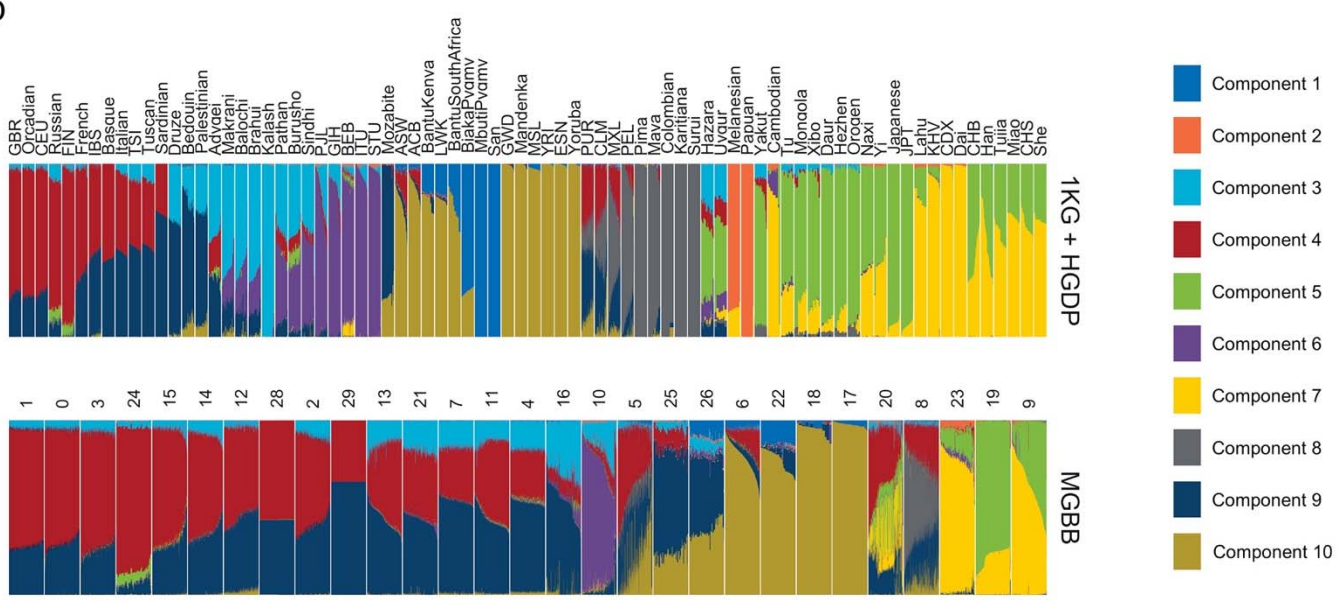


Figure S4

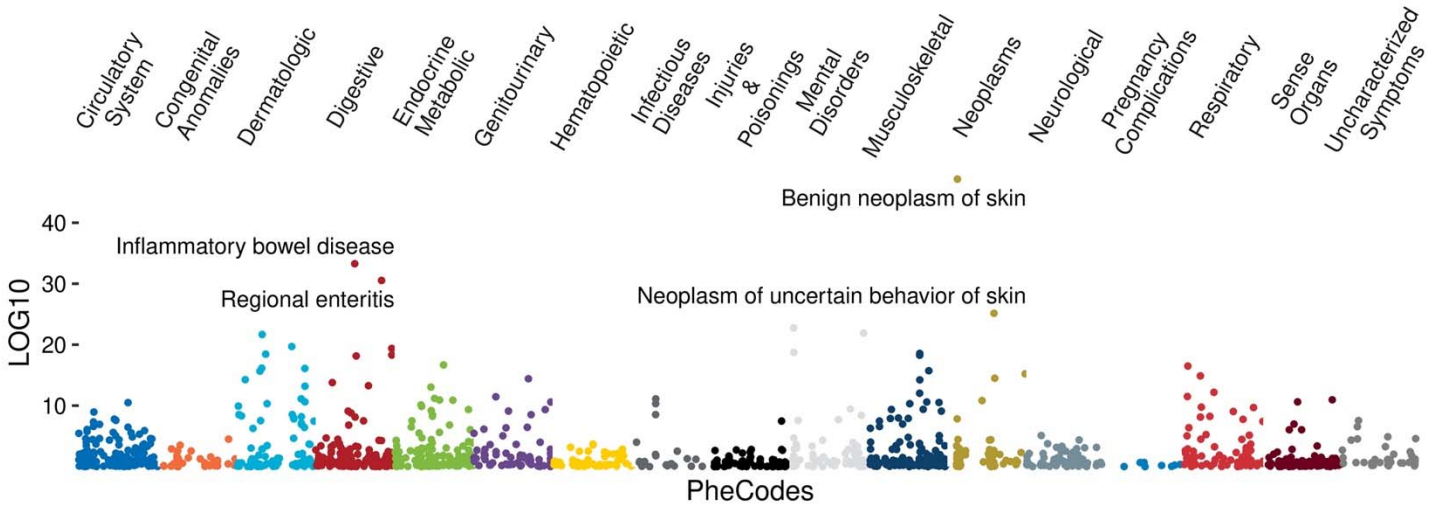


Figure S5

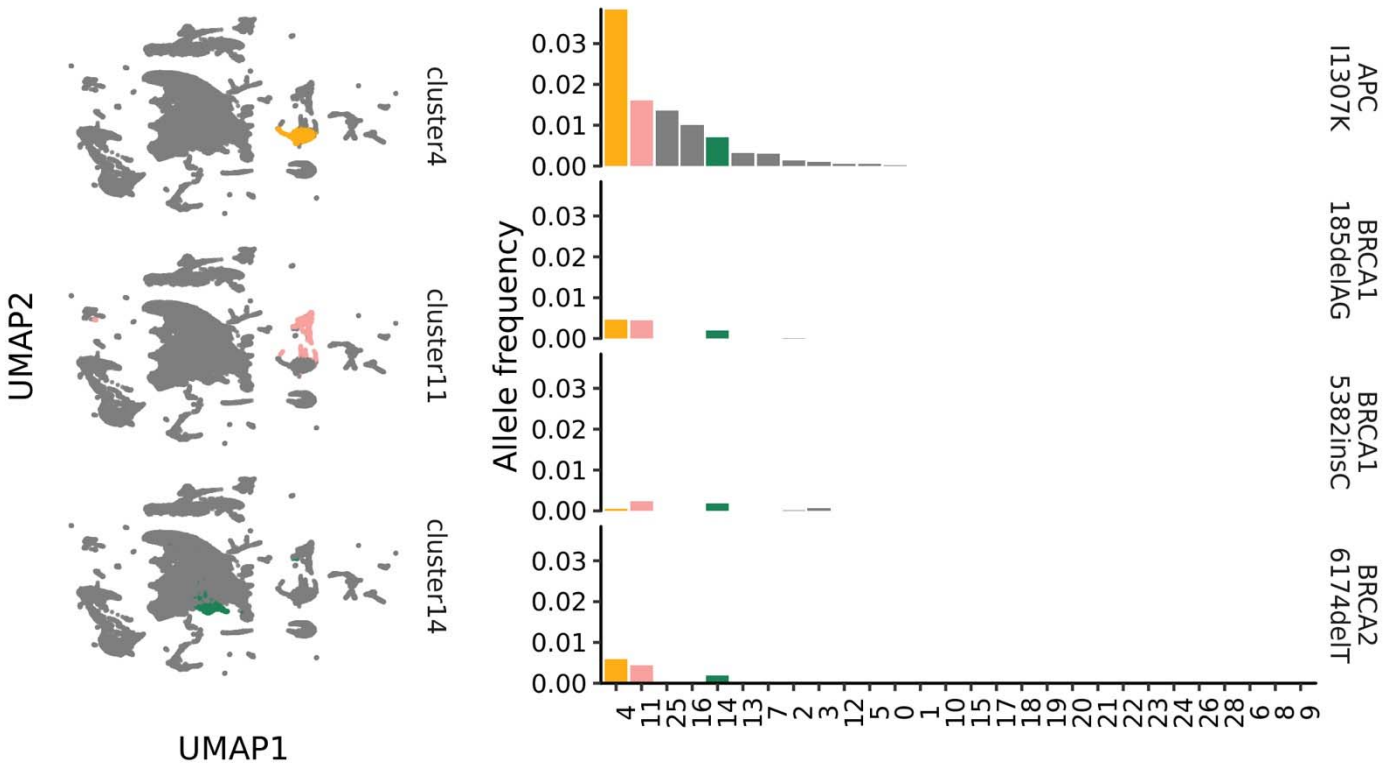
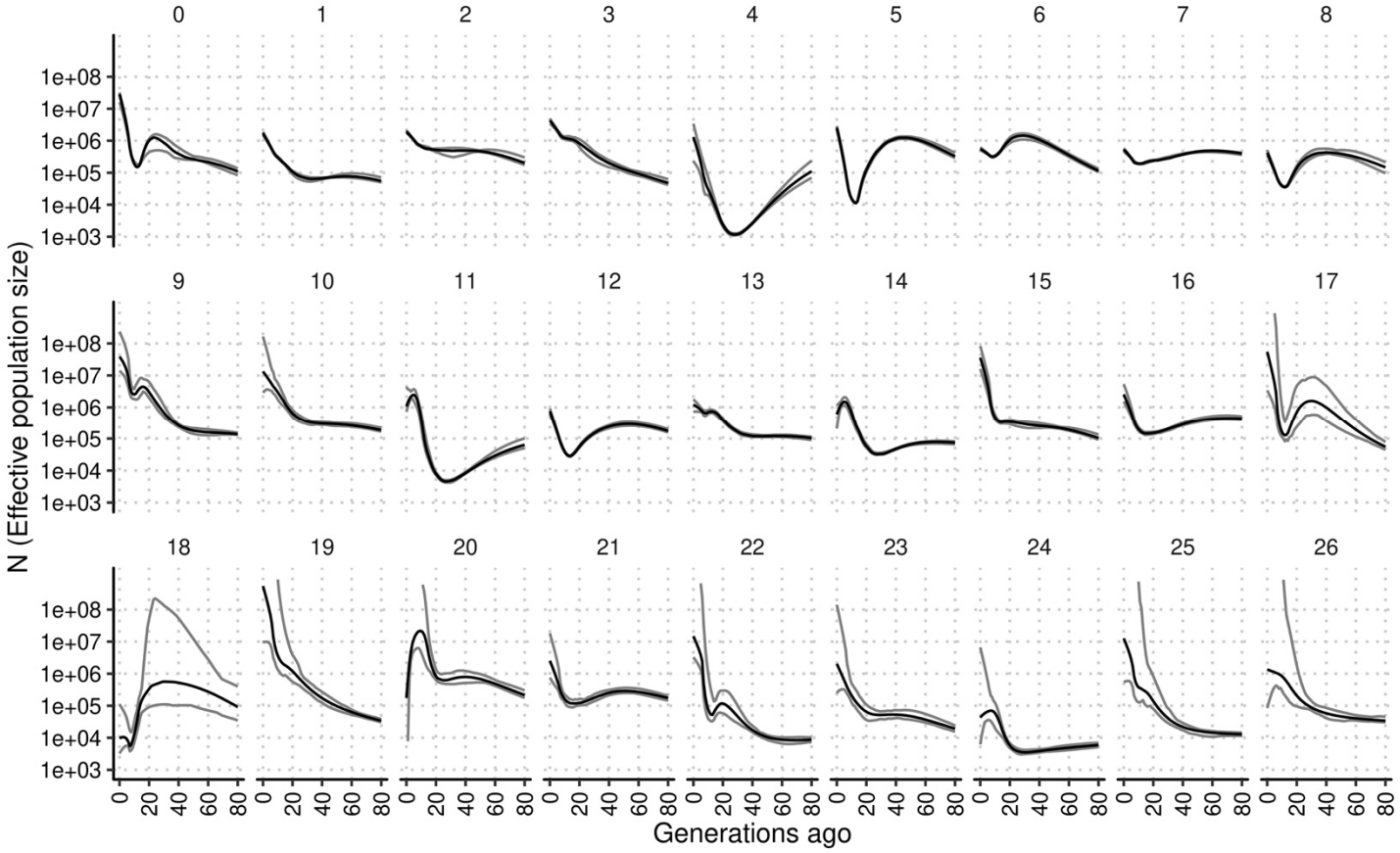


Figure S6



# Figure S7

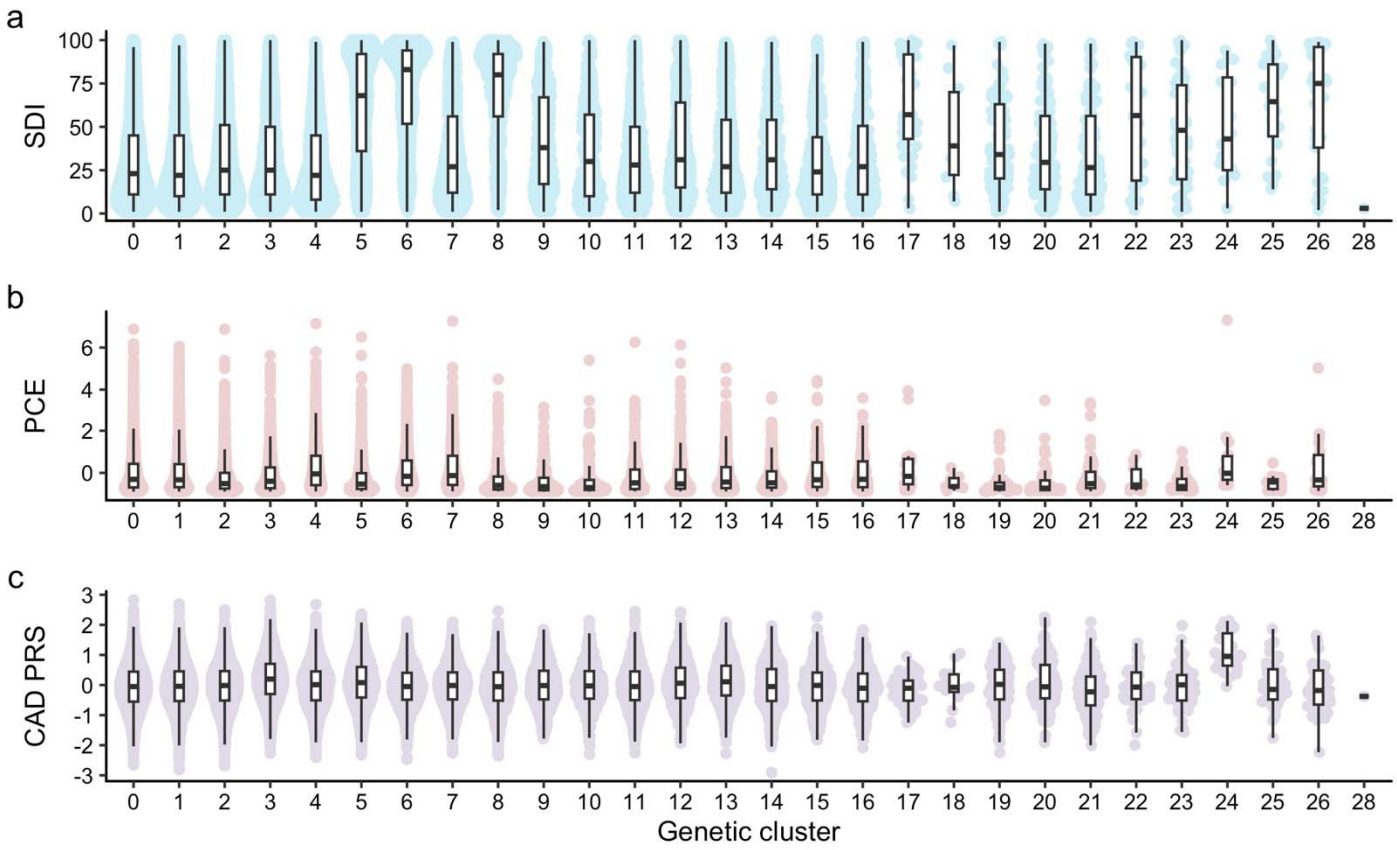


Figure S8

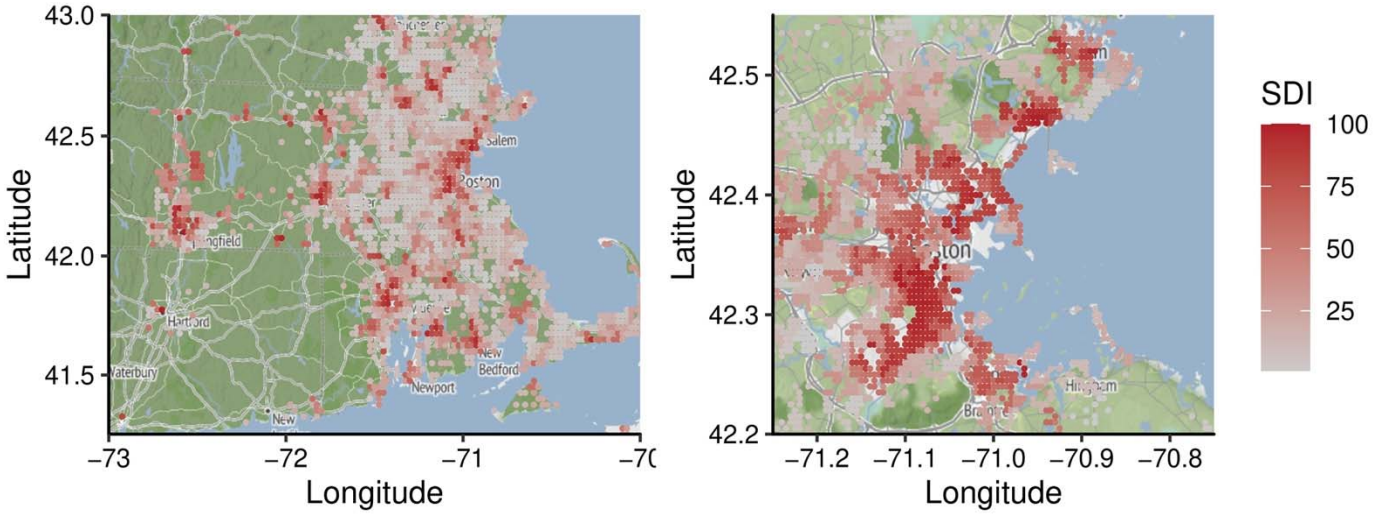


Figure S9

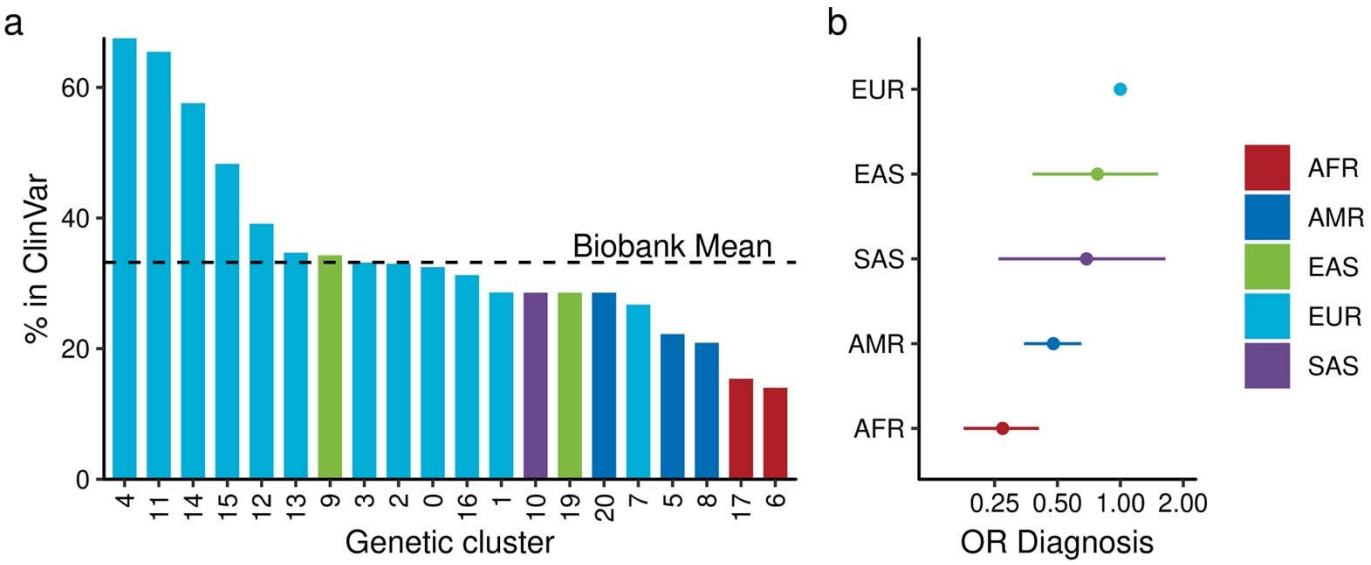


Figure S10