# Diverse Genomes, Shared Health: Insights from a Health System Biobank

Roni Haas*[,1,2,3,4], Michael P. Margolis*[,1,5,6], Angela Wei*[,1,7-9], Takafumi N Yamaguchi[1,2,3,4], Jeffrey Feng[10], Thai Tran[6], Veronica Tozzo[9], Katelyn J. Queen[11], Mohammed Faizal Eeman Mootor[1,2,3,4], Vishakha Patil[6], Michael E. Broudy[12], Paul Tung[12], Shafiul Alam[12], Danielle B. Martinez[12], Yash Patel[1,2,3,4], Nicole Zeltser[1,2,3,4], Rupert Hugh-White[1,2,3,4], Jaron Arbet[1,2,3,4], Christa Caggiano[13], Ruhollah Shemirani[13], Mao Tian[1,2,3,4], Prapti Thapaliya[12], Lora Eloyan[12], Lawrence O. Chen[1,5,6], Maryam Ariannejad[4], Clara Lajonchere[4,6], UCLA Precision Health Data Discovery Repository Working Group, UCLA Precision Health ATLAS Working Group, UCLA Health IT HPC Team, Regeneron Genetics Center; Bogdan Pasaniuc[14], Alex Bui[3,4,10], Valerie A. Arboleda[1,3,8,9], Timothy S. Chang[6], Noah Zaitlen[1,6,9], Paul T. Spellman[†11], Paul C. Boutros[†,1,2,3,4], Daniel H. Geschwind[†,1,4,5,6]

[1] Department of Human Genetics, University of California, Los Angeles, USA
[2] Department of Urology, University of California, Los Angeles, USA
[3] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, USA
[4] Institute for Precision Health, University of California, Los Angeles, USA
[5] Department of Psychiatry, David Geffen School of Medicine, Los Angeles, USA
[6] Department of Neurology, University of California, Los Angeles, USA
[7] Interdepartmental Bioinformatics Program, University of California, Los Angeles, USA
[8] Department of Pathology and Laboratory Medicine, University of California, Los Angeles, USA
[9] Department of Computational Medicine, University of California, Los Angeles, USA
[10] Department of Radiological Sciences, University of California, Los Angeles, USA
[11] Department of Medicine, University of California, Los Angeles, USA
[12] UCLA Health Information Technology, Office of Health Informatics and Analytics
[13] Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, New York
[14] Department of Genetics Perelman School of Medicine University of Pennsylvania

[†] Corresponding authors
* Equally contributed

## Abstract

Coupling genetic profiling with electronic health records from hospital biobanks is a foundational resource for precision medicine. However, lack of ancestral heterogeneity limits discovery and generalizability. We leveraged the UCLA ATLAS Community Health Initiative, a diverse biobank with >35% non-European participants in a single health system, to inform disease prevalence and genetic risk across five continental and 36 fine-scale ancestry groups. Analyzing clinical and genetic data for 93,937 individuals, 61,797 with whole-exome sequencing (WES), we identified novel associations between genetic variants and phenotypes, including *STARD7* with asthma risk in Mexican Americans and *FN3K* with intestinal disaccharidase deficiency across Europeans and Admixed Americans. Top decile polygenic scores (PGS) predicted patient status for many common diseases (40% of patients with Type 1 diabetes); an effect markedly diminished in non-Europeans. Exploring the distribution of ACMG ClinGen rare variants across populations demonstrated European bias in curated clinical variants. Mitigating this bias using computationally predicted deleterious variants, we identified new gene-disease associations, including *EXOC1L* and blood glucose level in East Asians. We identified *PTPRU* as a modulator of semaglutide's

effects on weight loss, and additionally found variability across ancestries and a relationship with type-2-diabetes PGS. We provide an interactive web portal for accessing cross-ancestry associations at atlas-phewas.mednet.ucla.edu. Collectively, our findings support the value of ancestral diversity in advancing precision health across a broad spectrum of populations.

# Introduction

The integration of electronic health records (EHRs) with genetic data is transforming biomedical research, offering unprecedented opportunities for preventing and treating common medical conditions[1]. Broad, longitudinal sampling from deeply phenotyped individuals, linked with genetic, environmental, and lifestyle data, provides many advantages over standard cohort-driven research[2]. This ability has fueled the creation of nationwide biobanks, including the UK Biobank, All of Us[3], FinnGen[4] and the Taiwan Biobank[5], as well as several large-scale academic biobanks, such as Mt. Sinai's BioMe[6]; Vanderbilt's BioVU[7]; Geissinger's MyCode[8]; and the Michigan Genomics Initiative[9]. Integration of these biobanks has permitted innovative collaborative efforts, such as the eMERGE consortium[10], COVID-19 host genomics initiative[11], and the Global Biobank Initiative[12].

Although these existing efforts have substantially advanced genetic and biomedical discovery, their concentration on European (EUR) ancestry participants underscores the importance of increasing ancestral diversity to enhance the applicability of findings across populations findings[13–17]. As PGS are ported into clinical use, the importance of measuring diverse populations is further emphasized by recent analyses showing the continuous, linear relationship between ancestral distance from the reference population and the utility of polygenic scores[16]. Similarly, rare genetic variation is known to have substantial ancestry-specific frequency and effect size (e.g., the reduced effect of *APOE4* alleles on Alzheimer's disease risk and increased frequency of rare *PCSK9* variants in those with African ancestry). Moreover, the interpretation of clinically-relevant rare variation is hampered by the EUR bias in most current studies and genetic databases[18,19] and including non-EUR populations reveals substantial disparities in clinically-relevant rare variant frequencies[20] and increases power for discovery[21,22]. Thus, greater ancestral diversity in large biobanks with detailed medical records strengthens efforts to advance precision health[13,23,24].

Previously, we described early phases of our work developing the UCLA ATLAS Community Health Initiative, a biobank embedded within the UCLA Health System consisting of deidentified genetic data coupled with EHRs. The goal of ATLAS is to advance precision medicine efforts in diverse populations, both locally in Los Angeles County and more broadly across California[25–27]. Here, we report an important milestone - discoveries from the current data freeze consisting of Single Nucleotide Polymorphism (SNP) genotyping for 92,165 and whole-exome sequencing (WES) for 61,797 individuals. We perform phenome-wide association in 84,110 individuals for 776 biomedical conditions, identifying 19,431 genome-wide associations. We replicate known common and rare variant associations for a range of biomedical conditions, including differences in effects across five major continental ancestries within this single biobank. We further investigate disease diagnoses and genetic risk in fine-scale ancestral clusters, identifying many new ancestry-specific risk associations. These analyses further demonstrate the utility of coupling EHR and genetic data in populations of diverse ancestries, from large continental (broad-scale) populations to, perhaps more importantly, regional and sub-continental (fine-scale) European, American, African and Asian populations in a single, unified health care system.

# Results

### The demographic and clinical landscape in ATLAS

To date, the UCLA ATLAS initiative has enrolled >220,000 patients, with biomaterial availability for >130,000 patients, with data collection ongoing. ATLAS enrollment largely reflects the broader composition of patients in the UCLA Health system, primarily concentrated over a broad area on the west side of Los Angeles (**Figure 1a**) - one of the world's most ancestrally diverse metropolitan areas with a population of 9.6 million. The EHR initiated in 2013 enables continuous longitudinal stratification of individuals into disease states, with a mean and median of 8.6 and 7.9 years of participation per individual. As of November 2024, the UCLA ATLAS Biobank of genomic data included genotype by sequencing (GxS) and Whole Exome Sequencing (WES) data from 61,797 individuals, and custom array genotyping using the Illumina Global Screening Array from 92,165 individuals, which underwent extensive quality control (QC) prior to analysis (**Supplementary Figure 1; Methods**). We leveraged these data, representing ATLAS data release two to interrogate social and genetic factors that affect disease risk and health outcomes, taking advantage of the unique composition of ATLAS enriched for understudied populations.

Genotyped cohort demographics are summarized in **Figure 1** and **Table 1**. Overall, biobank participants had a higher comorbidity index compared to non-biobank adults (3.3 vs. 1.7 mean Elixhauser index; 1-year post-collection), consistent with higher number of clinical visits, which favors enrollment (**Supplementary Table 1).** Common conditions were endocrine/metabolic, cardiovascular, gastrointestinal, and neoplasms; many also had immunologic, neurologic, and neuropsychiatric conditions, reflecting global trends (**Figure 1c; Supplementary Figure 2**)[28–30]. The ATLAS EHR data consisted of a total of 71,739,582 lab tests (including complete blood count [CBC], lipid, metabolic, HbA1c, and 25-hydroxyvitamin D panels), with a mean of 540 lab test results per patient (**Supplementary Table 2**). Detailed prescription data is also available with a total of 5,952,958 prescriptions; the 50 most frequent medications are listed in **Supplementary Table 3**.

The ATLAS population is diverse relative to most other large health system biobanks and the research-oriented UK biobank[31]. Although we primarily consider genetically determined ancestry in our analyses, rather than the social constructs of race and ethnicity[32,33], we note substantial diversity based on patient self-reports. For instance, 63.4% of patients self-identify as White, 11.2% self-identified as Asian, 4.87% as Black or African American, 2.4% as Middle Eastern or North African, 0.9% as American Indian or Alaska Native and 0.3% as Pacific Islander. A substantial proportion (14.6%) report Hispanic or Latino ethnicity (**Table 1**).

### Disease burden by broad-scale genetic ancestry

Self-identified race/ancestry, a cultural-societal construct, is often used as a proxy for genetic ancestry. But, the two are not synonymous[26,32,34] and genetic ancestry must be considered to prevent confounding in genetic association studies[33,35]. We classified biobank participants into six broad-scale ancestry groups primarily corresponding to continents *via* PCA (**Methods**): European (EUR), African (AFR), South Asian (SAS), East Asian (EAS), Admixed American (AMR) and an Unclassifiable group, aligning with the 1000 Genomes Project super-populations (**Figure 1d-f; Methods**). Consistent with self-reported race/ethnicity, almost a third of the biobank participants (32%) were assigned to non-EUR genetic ancestries: AFR, SAS, EAS, or AMR. There was significant agreement between self-reported race in the medical record and continental ancestry – 99% of those who self-identified as White were assigned EUR or AMR populations

(**Supplementary Figure 3a**)**.** However, PCA across common genetic variants demonstrates more granular relationships and provides a quantitative basis for assessing the relationship between ancestry and disease[26,34]. Of participants who self-reported their race as "unknown", 46% were assigned to the AMR ancestry, and 45% to the EUR group.

We next asked how health system usage or disease burden varied by broad-scale ancestry, observing that the mean number of encounters varied significantly among ancestry groups (p value < $1\times10^{-16}$, ANCOVA, adjusted for genetic sex and age), with the highest numbers of total hospital encounters in individuals from AFR ancestry, followed by AMR, with the lowest values in EUR, SAS, and EAS (**Figure 1g**). Another way to assess health burden is by a comorbidity index score (Elixhauser Comorbidity Index[36,37]), which was similarly distributed across most ancestries, with SAS individuals having the lowest scores (SAS 1.6; all other population groups 3.0-3.1; **Figure 1h)**. Increased encounter numbers were only partially explained by elevated comorbidity index scores, indicated by modest correlation between the two (mean total encounters *vs.* Elixhauser Comorbidity Index: R = 0.32, P <$2.2\times10^{-16}$; mean hospital encounters *vs.* Elixhauser Comorbidity Index: R = 0.26, P <$2.2\times10^{-16}$; **Supplementary Figures 3b-c**).

The population diversity of ATLAS enables the interrogation of the combined genetic and social/environmental effects on the risk of disease diagnoses and exploration of medical comorbidity in depth. To illustrate this, we assessed variation in medical conditions and disease prevalence across population groups for several major conditions (**Supplementary Figure 4a-f; Supplementary Table 4**).

We first confirmed known associations between cancer and genetic ancestry, observing the highest risk of prostate cancer in those with AFR ancestry, stomach cancer in EAS, bladder cancer in AMR, and breast cancer in EUR[38] (**Supplementary Figure 4a**). We also replicated several known cardiovascular disease associations. For instance, AFR individuals were more affected by hypertension and myocardial infarction, while those with SAS ancestry had the lowest risk of atrial fibrillation despite the highest incidence of coronary atherosclerosis (**Supplementary Figure 4b)** – an established but seemingly contradictory risk profile[39].

Major metabolic disorders were more prevalent in AFR participants compared to other ancestries**.** This finding included diagnostic codes (9- and 10-International Classification of Diseases [ICD]) related to Type 2 diabetes, hypercholesterolemia, and hyperlipidemia, supported by laboratory measurements or vitals (**Supplementary Figure 5a)**. Type 2 diabetes was more common in all non-EUR groups relative to EUR (**Supplementary Figure 4c)**. Previous work has suggested that some of these differences reflect social and environmental, rather than genetic, factors[40]. With regard to neurological conditions, we find a higher risk of dementia in AFR participants, but a lower risk of migraines and Parkinson's Disease in AFR compared to EUR patients, consistent with prior analyses[41–43] (**Supplementary Figure 4d**). Among neuropsychiatric disorders, anxiety and major depressive disorders were the most frequent in EUR individuals, and least frequent in the continental Asian cluster, as previously reported (**Supplementary Figure 4e**)[44,45].

We also identified previously unreported associations, such as a lower risk of epilepsy in EAS compared to EUR individuals (**Figure 1i**; odds ratio [OR] = 0.46 with 95% confidence interval [0.32, 0.65], $P_{bonferroni} = 2.4\times10^{-3}$), which prior studies did not detect[46,47]. Similarly, we found significantly reduced risk for bipolar disorder in those with AMR ancestry, clarifying conflicting findings in previous smaller studies[48,49] (**Figure 1i**; OR = 0.47 [0.40, 0.56], $P_{bonferroni} = 1.5\times10^{-16}$).

We extend previous findings that sleep apnea is highest in AFR individuals but show a significantly lower risk of sleep apnea in EAS individuals relative to EUR (**Figure 1h**; OR = 0.67 [0.59, 0.75], $P_{bonferroni} = 4.05\times10^{-9}$), which has not been definitively demonstrated[50–53]. These associations illustrate the utility of ATLAS for cross-ancestry comparisons, including under-represented populations, within a single health system biobank to reduce potential confounding due to geographic effects.

**Disease prevalence varies across fine-scale clusters**

Broad-scale ancestry encompasses multiple geographically and genetically distinct fine-scale populations, which in turn reflect recent geographic or socio-demographic stratification. When a fine-scale population arises from a small group of founders, deleterious rare alleles may become disproportionately prevalent over time and drive elevated rates of disorders like epileptic encephalopathy[54] or liver disease[55]. Similarly, fine-scale populations may exhibit significantly different frequencies of common risk variants compared to large populations[35,56,57], many of which are understudied[13–15].

We have previously shown that fine-scale ancestral analysis reveals critical features of genetic and environmental contributions to health, including marked disparities in disease risk[13]. We therefore sought to apply our previous identity-by-descent (IBD) total shared length approach to construct fine-scale ancestries in a population nearly three times the size of our previous work[56]. Briefly, this approach identifies genomic regions shared between individuals due to a common ancestor and defines fine-scale ancestries, which we refer to as “clusters” (**Methods**[57,58]), based on the total amount of IBD segments shared between patients. We labeled each cluster with both a numeric identifier (e.g., IBD-01, IBD-02) and a corresponding cluster-specific name to facilitate interpretation, using reference populations and patients' self-reported demographic information, as we did previously[56]. With this approach, we identified 36 fine-scale ancestry clusters represented by at least 30 individuals (**Figure 2a-b**; **Supplementary Figures 5b-c**; **Methods**), replicating previously identified fine-scale clusters and adding new ones, such as a Native Hawaiian cluster (IBD-29; n = 64) and a Bantu cluster (IBD-35; n = 32). The largest clusters in our biobank consist of Northern Europeans (IBD-01; n = 33,675) and Southern Europeans (IBD-02; n = 14,841), making up 53.2% of our biobank. The remaining clusters represent the heterogeneity of ancestral origins in the Los Angeles community, such as Armenians (IBD-16, IBD-23, IBD-31; n = 560), Ashkenazi Jews (IBD-03; n = 14,261), Iranian Jews (IBD-11; n = 707), and Filipinos (IBD-09; n = 1,438).

For a nuanced understanding of diagnostic variation, we tested the prevalence of 1,253 phecodes[59–61] across fine-scale clusters with at least 100 patients (**Figure 2c;** 23 tested clusters**).** As an initial measure of integrity, we tested for known increases in prevalence in these fine-scale ancestries, for example replicating an increase in breast cancer, irritable bowel syndrome, and Crohn’s disease in the Ashkenazi Jewish (IBD-03) cluster (**Supplementary Table 5**). Similarly, we observed the known increase in gout in the Filipino cluster[62] and Alzheimer’s disease and dementias in the Puerto Rican cluster[63]. Consistent with the global ancestry findings and previous reports, fine-scale South Asian clusters had the highest coronary atherosclerosis risk, and the Japanese cluster had a higher prevalence of hypertension[64]. Cholelithiasis was widespread in clusters from South America and Mexico, while less frequent in the Northern European cluster (**Supplementary Table 5)**.

We additionally uncovered numerous novel associations between our fine-scale clusters and disease (**Supplementary Table 5**; **Figure 2c**). For instance, the Asian, Filipino, and combined Chinese + Korean clusters showed a decrease in risk for vitamin B-complex deficiency (Filipino: OR = 0.50 [0.36, 0.67], FDR = $2.1\times10^{-4}$; Chinese/Korean: OR = 0.6 [0.5, 0.7], FDR = $9.4\times10^{-11}$). Cholesterolosis of the gallbladder was substantially more widespread in several Asian clusters, including, Filipino, combined Chinese/Korean, Southeast Asian, and one of the West Asian clusters (Filipino: OR = 4.3 [2.9, 6.2], FDR = $5.6\times10^{-12}$; Chinese + Korean: OR = 2.3 [1.8, 3.1], $3.2\times10^{-8}$; Southeast Asian: OR = 3.8 [1.4, 8.0], $1.9\times10^{-2}$; West Asian 1: OR = 3.2 [1.3, 6.3], $3.1\times10^{-2}$), and less common in Northern EUR (OR = 0.6 [0.5 -0.7], $1.0\times10^{-5}$). Mexican and South American clusters suffered considerably more from hormones' adverse effects in therapeutic use (Mexican American clusters: OR = 2.0-2.8 [1.3, 2.3], FDR = $1.6\times10^{-2}$ - $2.1\times10^{-24}$; combined South American individuals: OR = 1.9 [1.4, 2.5], $1.3\times10^{-4}$). Glaucoma was common in the African American (IBD-06) and East Asian clusters (Chinese + Korean (IBD-05), Japanese (IBD-10)), as is already known[65], but also in the Iranian Jewish cluster (OR = 1.9 [1.5-2.4], FDR = $2.9\times10^{-6}$), which has not been reported before. Iranian and Ashkenazi Jewish clusters showed the highest risk for hyperplasia of prostate (Iranian Jewish: OR = 2.3 [1.6, 1.8], FDR = $1.1\times10^{-13}$; Ashkenazi Jewish: OR = 1.7 [1.6, 1.8], FDR = $4.6\times10^{-66}$) and bladder cancer (Iranian Jewish: OR = 1.5 [1.1, 1.9], FDR = 0.008; Ashkenazi Jewish: OR = 3.1 [1.4, 5.6], FDR = 0.02), and the lowest risk for cirrhosis of liver (Iranian Jewish: OR = 0.2 [0.03, 0.5], FDR = $3.0\times10^{-2}$; Ashkenazi Jewish: OR = 0.4 [0.3, 0.5] , FDR = $6.8\times10^{-24}$).

We next focused on cardio-metabolic diseases due to their high prevalence in ATLAS and global impact on public health, comparing disease risk for each fine-scale cluster within the same broad-scale continental ancestry **(Methods; Figure 2d)**. Strikingly, among Asian clusters, the Filipino cluster had elevated risk for all tested medical conditions, which remained after adjusting for body mass index (BMI) (essential hypertension: OR =1.6 [1.3, 1.7], FDR = $9.0\times10^{-11}$; Type 2 diabetes: OR = 1.4 [1.6, 2.1], FDR = $7.1\times10^{-6}$; Coronary atherosclerosis: OR = 1.4 [1.1, 1.7], FDR = $5.3\times10^{-3}$; abdominal aortic aneurysm: OR = 2.8. [1.4, 5.3], FDR = $7.1\times10^{-3}$ Hyperlipidemia: R = 1.2 [1.0-1.4], FDR = $2.4\times10^{-2}$. Firth's bias-reduced logistic regression adjusted for BMI, sex and age).

Across fine-scale EUR ancestry clusters within ATLAS (Ashkenazi Jewish, n = 14,262; Iranian, n = 363; Iranian Jewish, n = 707; two Armenian clusters, n = 382 and 134; Lebanese, n = 183; Egyptian Christians, n = 107), the larger Armenian cluster (IBD-16) showed an elevated risk for the range of cardiometabolic conditions including Type 2 diabetes (OR = 2.0 [1.5-2.7], FDR = $2.5\times10^{-6}$) and essential hyperlipidemia (OR = 1.6 [1.2-2.0], FDR = $8.5\times10^{-4}$). Iranian clusters, both Jewish and non-Jewish, had a relatively higher risk of coronary atherosclerosis (Iranian Jewish: OR = 1.8 [1.5, 2.2], FDR = $4.4\times10^{-8}$; Iranian cluster 1: OR =1.9 [1.4, 2.5], FDR = $2.8\times10^{-5}$), and Type 2 diabetes (Iranian Jewish: OR = 2.4 [1.9, 2.9], FDR = $6.6\times10^{-16}$; Iranian 1: OR = 2.2 [1.6, 2.8], FDR = $1.7\times10^{-6}$), but not hypertension. As is known, among AMR fine-scale clusters, those in the Mexican American cluster had a high risk of Type 2 diabetes, and combined South Americans showed higher risk for hyperlipidemia (**Supplementary Figure 5d)**.

**Utility of PGS in stratifying risk for common disorders**

We next leveraged genotyping in our cohort to calculate individual PGS for a range of cancer, cardiovascular, metabolic, neuro-psychiatric and autoimmune diseases (**Methods**) and tested their relationship to disease risk. We focused on the top end of the PGS distribution (10%),

compared with the 5th decile, in EUR individuals (**Figure 3; Supplementary Table 6**). For Type 1 diabetes, the top PGS decile was the most enriched (OR = 11.7 [7.6-19.0], FDR = $1.3\times10^{-25}$), with 41% of the diagnosed patients in ATLAS assigned to the top PGS decile. The second most enriched trait was Crohn's disease, with 33% of diagnosed patients within the top PGS decile (OR = 5.5 [4.0-7.7], FDR = $2.8\times10^{-24}$), followed by gout with 27% (OR = 5.11 [4.1-6.4], FDR = $2.3\times10^{-46}$), testicular cancer with 25% (OR = 3.7 [1.2-7.3], FDR = $8.1\times10^{-5}$) and prostate cancer with 21% (OR = 3.4 [2.8-4.0], FDR = $3.64\times10^{-46}$) (**Supplementary Figure 6a-e)**. On average, the top decile of risk accounted for 18.5% of diagnosed patients across 28 tested disorders. Twenty-five of 28 (89%) showed significant enrichment of patients within the top PGS decile (mean OR for significant tests = 2.9). Performance declined when we applied PGS to non-EUR populations, as expected[55–58]. This was due to both reduced sample sizes and the model fit (**Supplementary Figure 6f-i**), identifying only 16.7, 27.3, 36.4, and 38.5% significant associations of diseases with the top PGS decile, for SAS, AFR, AMR, and EAS, respectively. Concordantly, the top PGS decile accounted for fewer cases (on average, 12.6%, 15.0%, 15.5% and 15.9%, for SAS, AFR, AMR, and EAS). This further supports the need for larger, more ancestrally heterogeneous cohorts for clinical development of PGS[55–58].

**Common variant associations across broad- and fine-scale ancestries**

We next sought to understand how genetic risk for common, complex traits was distributed across broad- and fine-scale ancestry. Through phenome-wide association analysis (PheWAS) of 84,110 genotyped individuals using the Regenie framework[70] (**Methods**), with analyses run separately for cases/controls within each cohort, we refined numerous known associations and identified novel associations that represent cohort-specific disease risk. After LD pruning, we identified a total of 19,431 unique variant-phenotype associations passing a genome-wide significance threshold of $5\times10^{-8}$ (**Supplementary Table 7**), with 5,772 passing the most conservative Bonferroni threshold of $6.4\times10^{-11}$ ($5\times10^{-8}$ conditioned on 776 tested phenotypes). We provide a web portal at atlas-phewas.mednet.ucla.edu to facilitate browsing of these associations.

First, we examined the distribution of *APOE* alleles across fine-scale clusters, highlighting the increased frequency of *ε4* risk alleles in the African American (IBD-06) and Bantu (IBD-35) clusters (**Figure 4a**). We replicate the finding that African American individuals with *ε4/ε4* haplotype have a lower risk of Alzheimer's disease compared to other ancestral cohorts (**Supplementary Figure 8**). We additionally replicated two well-known genetic associations in the African American (IBD-06) cluster, between *HBB* rs334-A (**Figure 4b**; $MAF_{IBD\text{-}06}$ = 5.21%) and a diagnosis of sickle cell anemia ($P_{IBD\text{-}06} = 1.96\times10^{-78}$; $OR_{IBD\text{-}06}$ = 15.08) and between the Duffy null *ACKR1* rs2814778-C polymorphism (**Figure 4b**; $MAF_{IBD\text{-}06}$ = 77.06%) and a decrease in neutrophil count ($P_{IBD\text{-}06}$ $4.00\times10^{-31}$; $OR_{IBD\text{-}06}$ 0.67). Using quantitative blood work findings from ATLAS, we also profiled the impact of rs334-A on mean corpuscular hemoglobin concentration ($P_{IBD\text{-}06} = 5.21\times10^{-15}$; $\beta_{IBD\text{-}06}$ = 0.41), nucleated red blood cell count ($P_{IBD\text{-}06} = 1.59\times10^{-10}$; $\beta_{IBD\text{-}06}$ = 0.16), and mean corpuscular volume (MCV) ($P_{IBD\text{-}06} = 3.27\times10^{-8}$; $\beta_{IBD\text{-}06}$ = -0.29). Both risk variants have been concluded to be under strong selection for resistance to malaria[71].

Across multiple Asian clusters, we confirmed the association between a variant in high LD with the $\text{--}^{SEA}$ deletion, which causes inherited alpha-thalassemia[72], and microcytic anemia. We observed a significant decrease in MCV in response to *LUC7L* rs372755452-A in the broad-scale EAS ancestry ($P_{EAS} = 6.10\times10^{-88}$; $\beta_{EAS}$ = -1.64; $MAF_{EAS}$ = 0.91%), and the fine scales Chinese +

Korean ($P_{IBD-05} = 2.71\times10^{-52}$; $\beta_{IBD-05} = -1.58$; $MAF_{IBD-05} = 0.79\%$) and Filipino ($P_{IBD-09} = 2.15\times10^{-24}$; $\beta_{IBD-09} = -1.77$; $MAF_{IBD-09} = 1.21\%$) clusters. We also replicated the finding that AMR individuals are twice as likely to carry a *PNPLA3* rs738409-G missense variant (**Figure 4b**) that greatly increases the risk for non-alcoholic fatty liver disease, a major cause of cirrhosis that often necessitates liver transplant[73].

We extended these findings by identifying significant associations between rs738409-G and non-alcoholic cirrhosis of liver in two Mexican American clusters (**Figure 4c**), IBD-04 ($P_{IBD-04} = 2.66\times10^{-14}$; $OR_{IBD-04} = 1.90$) and IBD-07 ($P_{IBD-07} = 1.46\times10^{-8}$; $\beta_{IBD-07} = 1.95$), with differing allele frequencies across these cohorts ($MAF_{IBD-04} = 45.00\%$; $MAF_{IBD-07} = 52.93\%$). In the Northern European cluster, this variant has a smaller effect size ($P_{IBD-01} = 3.62\times10^{-09}$; $OR_{IBD-01} = 1.53$) and is half as frequent ($MAF_{IBD-01} = 22.88\%$). We identified a previously unreported association between *STARD7* rs17419569-C, an intronic variant, and asthma in a Mexican American cluster (**Figure 4d**; $P_{IBD-04} = 1.94\times10^{-9}$; $OR_{IBD-04} = 2.62$; $MAF_{IBD-04} = 2.49\%$). Prior work suggests that decreased *STARD7* expression is associated with enhanced allergic responses in the human lung and significant increases in airway hyperresponsiveness in haploinsufficient *Stard7* mice, supporting this discovery[74]. In the same cluster, we identified a novel association between gastrointestinal reflux disease (GERD) and rs74744741-C (**Figure 4e**; $P_{IBD-04} = 5.79\times10^{-10}$; $OR_{IBD-04} = 1.63$; $MAF_{IBD-04} = 8.92\%$), nominating *GPX7* and *SHISAL2A* as risk genes. *GPX7* has previously been associated with carcinogenesis in the context of GERD-associated Barrett's esophagus[75], while SHISAL2A has unknown function, but is highly expressed in the small intestine and in lymphoid tissues[76].

We further identified two low frequency variants associated with chronic renal failure in the AMR cohort (**Figure 4f**), rs112680741-C ($P_{AMR} = 2.94\times10^{-11}$; $OR_{AMR} = 3.78$; $MAF_{AMR} = 1.24\%$) and rs2744548-C ($P_{AMR} = 9.59\times10^{-11}$; $OR_{AMR} = 3.17$; $MAF_{AMR} = 1.72\%$), nominating *GPLD1*, *ALDH5A1*, and *KIAA0319* as potential risk genes. None have previously been associated with kidney disease, although *ALDH5A1* and *KIAA0319* are highly expressed in the kidney[76].

Lastly, we identify a new common variant association for intestinal disaccharidase deficiency (**Figure 4g**), the inability to completely digest sugars such as lactose and sucrose, which is a common cause of IBS-like symptoms such as persistent abdominal pain and diarrhea[77]. Across four EUR cohorts ($P_{EUR} = 3.80\times10^{-41}$; $OR_{EUR} = 1.26$; $MAF_{EUR} = 33.26\%$) and the AMR cohort ($P_{AMR} = 4.90\times10^{-11}$; $OR_{AMR} = 1.31$; $MAF_{AMR} = 42.33\%$), we associate a high-frequency *FN3K* rs7208565-T variant with increased risk for intestinal disaccharidase deficiency. *FN3K* phosphorylates glycated proteins, preventing the formation of advanced glycation end-products (AGEs), and has been implicated in a wide range of conditions[78].

**Frequency of known clinically relevant rare variants vary across broad- and fine-scale ancestries**

The ATLAS WES catalog comprises more than 12 million autosomal variants, with a median of 9,821 missense and 149 loss of function (LOF) variants per individual (including both rare and common variants), closely matching expectations from other studies[79]. Most WES variants were rare (MAF < 1% in any broad-scale ancestry; n = 11,808,484), including 2,724,021 rare missense and 155,006 rare LOF variants (**Table 2**), with a median of 15 rare LOF and 403 rare missense variants per participant (**Supplementary Table 8**). These values varied significantly across

ancestry groups (**Supplementary Table 9**), with AFR individuals showing the most rare or common LOF and synonymous variants (180 and 11,550), matching prior results[80].

We selected several clinically relevant rare variants with known elevated frequencies in specific ancestral clusters and confirmed that these patterns were replicated in ATLAS. For instance, Ashkenazi Jewish (IBD-03) had the primary risk for carrying either pathogenic or likely pathogenic (P/LP) variants in *BRCA1* or *BRCA2* (*BRCA1*: OR = 47.143 [20.6, 133.0]; *BRCA2*: OR = 48.2 [23.6, 114.2]) (**Figure 5a**). Similar to past reports[81], 4.6% of breast cancer and 11.8% of ovarian cancer patients carry either P/LP variants in *BRCA1* or *BRCA2*. The three *BRCA* Ashkenazi Jewish founder alleles were the most prevalent in EUR, and particularly the Ashkenazi Jewish cluster (**Supplementary Figure 9a-b**). Additionally, we identified 753 carriers of variants associated with risk for Familial Mediterranean Fever (FMF), with the highest carrier frequencies in the Armenian clusters[82,83], and the next highest carrier frequencies in West Asian, Lebanese, and Egyptian Christian clusters (**Supplementary Figure 9c**). Carriers of these variants had a higher risk of the amyloidosis phecode (OR = 1.3 [0.8-1.8]; **Supplementary Figure 9d**). The HBB:p.E7V variant, which is responsible for most sickle cell anemia cases, was carried by 273 participants, with elevated frequency in the African American (IBD-06) cluster (OR = 51.4 [39.7, 67.0]; **Supplementary Figure 9e**), in line with previous findings[84]. Finally, we identified 49 carriers of loss-of-function variants in the *PCSK9* gene causing lowered LDL levels, based on curated protective variants[85] within the African American (IBD-06) cluster (OR = 9.1 [4.5-17.4]) (**Supplementary Figure 9f**).

Given the ability to replicate known findings, we next tested ATLAS' unique ancestral heterogeneity to detect new enrichments of known rare clinically relevant variants across fine-scale clusters. We queried the full set of the curated rare P/LP ClinGen[86] variants, which have strong clinical and genetic evidence (n = 643), identifying 5,223 unrelated carriers for ClinGen P/LP variants. Within the Filipino (IBD-09) cluster, we detected an elevated carrier frequency for variants within the Tier 1 gene, *LDLR*, that causes familial hypercholesterolemia (OR = 3.9 [1.4-8.8], FDR = 0.03; **Figure 5a**). This finding has interesting implications given the high prevalence of dyslipidemia in Filipinos[87] (**Figure 2d**). Additionally, we identify an elevated carrier frequency of variants that cause non-syndromic genetic deafness within two different genes in two separate clusters: (1) the *MYO15A* gene in our Central American 1 (Mexican) (IBD-04) cluster (OR = 6.9 [2.6-17.4], FDR = $8.44 \times 10^{-4}$; **Figure 5a**) and (2) the *CDH23* gene in our East Asian 1 (Han Chinese/Korean) (IBD-05) cluster (OR = 6.0 [2.0-15.6], FDR = $6.04 \times 10^{-3}$; **Figure 5a**) and East Asian 2 (Japanese) (IBD-10) cluster (OR = 28.1 [8.9-77.1], FDR = $7.38 \times 10^{-6}$) that have not been previously reported. We also identified previously unreported elevated carrier frequencies in the Ashkenazi Jewish (IBD-03) cluster for Pendred syndrome variants in the *SLC264A* gene (OR = 3.8 [2.9-5.0], FDR = $8.61 \times 10^{-19}$) and recombinase activating gene 2 deficiency in the *RAG2* gene (OR = 2.9 [1.4-5.8], FDR = 0.02). This illustrates the utility of fine-scale ancestry clustering in identifying populations at higher risk of rare, monogenic disease.

Finally, we focused on the clinically actionable American College of Medical Genetics and Genomics (ACMG) genes secondary findings (SF) v3.2[88]. We asked if the total allele frequencies of ACMG ClinGen P/LP variants vary across broad- and fine-scale ancestries. Overall, 17 ACMG genes had at least one P/LP variant in ClinGen. For a more nuanced biological understanding, we divided the ACMG variants into two groups of rare LOF (n = 53) and rare P/LP missense (n = 131) variants (we defined 'LOF' for variants ranked as high-confidence LOF by loss-of-function transcript effect estimator [LOFTEE][89], which accounts for biological context beyond truncating

variant identification, and 'missense' based on the VEP[90] "missense_variant" annotation; see **Methods**). We observed the highest frequency of rare P/LP LOF variants in EUR individuals among broad-scale groups, and in the Ashkenazi Jewish cluster among fine-scale clusters (**Supplementary Figure 9g**). The numbers of total rare P/LP LOF alleles (as defined based on broad-scale groups) in these populations were by far the highest compared to all others ($OR_{EUR}$ = 3.7 $CI_{EUR}$ = 2.6-5.4, $P_{Bonferroni} = 6.3\times10^{-17}$; $OR_{Ashkenazi\ Jewish}$ = 6.5, $CI_{Ashkenazi\ Jewish}$ = 5.3 - 8.1, $P_{Bonferroni} = 3.5\times10^{-62}$; **Figure 5b**). No differences in rare P/LP missense total counts were identified at the level of broad-scale ancestries, but Northern EUR individuals had significantly more at the level of fine-scale clusters ($OR_{Northern\ EUR}$ = 1.4, $CI_{Northern\ EUR}$ = 1.15 - 1.8, $P_{Bonferroni}$ = 0.02). This replicates a bias toward EUR ancestry for rare curated clinical variants, as was recently shown in All of Us[20], but we demonstrate this in a single health system.

**The distribution of ACMG damaging variants varies across broad- and fine-scale ancestries**

As clinical variant databases are primarily ascertained from EUR cohorts, we took a conservative approach to call putative P/LP missense and LOF rare variants, using a consensus of a majority (5 of 9; **Methods**) of state-of-the-art computational methods based on the Critical Assessment of Genome Interpretation (CAGI) project[91] for pathogenicity prediction (**Methods**). Keeping our focus on ACMG reportable genes[88], this yielded 1,602 rare LOF variants and 6,487 rare predicted damaging missense variants within this ACMG gene set. We emphasize that these are distinct from the approach taken with ClinVar variants, as variants identified in this manner do not necessarily have evidence supporting their association with disease but have the advantage that they are not biased by oversampling EUR cohorts.

We then tested differences in the overall abundance of these variants across populations. Among broad-scale populations, EAS individuals carried significantly more predicted damaging rare missense variants compared to all other groups. AFR individuals had substantially more rare LOF variants compared to other ancestries (**Figure 5c-d**); and EUR clusters had the lowest rare predicted damaging missense, which is consistent with published data[92]. However, among fine-scale ancestries (n>100), both the fine-scale EUR Iranian Jewish and AFR clusters had significantly higher numbers of rare LOF (as defined based on broad-scale groups). Rare predicted damaging missense alleles were also significantly more common in Iranian Jewish, along with the EAS fine-scale Filipino and Chinese + Korean clusters (**Figure 5c**). In contrast, the Japanese cluster carried the smallest number of rare predicted damaging missense variants in this ACMG gene list (**Figure 5c-d**). These observations contrast in comparison with rare variants identified in ClinGen (**Figure 5b**), consistent with the interpretation that ClinGen variant frequencies are biased because they are based primarily on a majority of EUR patients, as is the case for most curated clinical data sets.

**Rare variant associations**

To examine the influence of predicted deleterious rare LOF (dLOF) and missense coding variants (dMIS) on ATLAS phenotypes (**Methods**), we performed exome-wide association studies (ExWAS) across 17,537 protein-coding genes. Using Regenie's unified gene burden association strategy ("GENE_P" test[70]), we identified 1,099 unique gene-trait associations at a significance threshold of $2.85\times10^{-6}$ (0.05/17,537 protein-coding genes). Within the EUR cohort, we replicated numerous known associations – 45 of the top 50 associations have been previously reported[79] (**Supplementary Table 10**). These include *PKD1* with cystic kidney disease (**Figure 5e**; $P_{EUR,LOF}$

= $9.41\times10^{-42}$, $OR_{EUR,LOF}$ = 107.14 [58.14, 197.43]), *TTN* with primary intrinsic cardiomyopathy ($P_{EUR,LOF}$ = $2.95\times10^{-14}$, $OR_{EUR,LOF}$ = 4.52 [3.18, 6.41]), and *JAK2* with polycythemia vera ($P_{EUR,MIS}$ = $1.14\times10^{-45}$, $OR_{EUR,MIS}$ = 27.42 [18.25, 41.18]).

In the Northern European (IBD-01) cluster, we detected a novel gene-level association between *HNRNPA1L2* and acquired absence of the breast/breast cancer (GENE_$P_{IBD-01}$ = $1.00\times10^{-25}$; GENE_$P_{IBD-01}$ = $2.21\times10^{-8}$; **Supp Table 10**). Though *HNRNPA1L2* has not been previously associated with breast cancer, it is highly expressed in breast invasive carcinoma[76] and a paralog *HNRNPA1* was previously implicated in breast cancer progression[93]. In the Ashkenazi Jewish (IBD-03) cluster, we also find novel associations between *EPG5* and HDL cholesterol level ($P_{LOF_MIS}$ = $3.55\times10^{-10}$; $\beta_{LOF_MIS}$ = 1.77 [1.22, 2.33]) as well as triglyceride level ($P_{LOF_MIS}$ = $1.33\times10^{-8}$; $\beta_{LOF_MIS}$ = -1.81 [-2.44, -1.18]). *EPG5* is an autophagy tethering factor classically associated with Vici syndrome[94], a severe developmental disorder, though our findings suggest an additional role in lipophagy and lipid metabolism[95].

In the AMR population, we identified a novel association between dLOF and dMIS variation in *CLN3* and cystic kidney disease (GENE_$P_{AMR}$ = $8.67\times10^{-9}$), supported by experimental evidence that *CLN3* is highly expressed in medullary collecting duct principal cells and plays a role in osmoregulation[96]. We additionally identified novel associations between *PPARG* and viral pneumonia ($P_{AMR,MIS}$ = $1.06\times10^{-6}$, $OR_{AMR,MIS}$ = 52.96 [13.45, 208.48]) as well as *NADSYN1* and abnormal lung examination findings ($P_{AMR,MIS}$ = $1.24\times10^{-5}$, $OR_{AMR,MIS}$ = 3.73 [2.13, 6.51]). *PPARG* is known to regulate macrophage response to pulmonary inflammation[97], while *NADSYN1* knockout in mouse models was shown to cause abnormal lung development[98], though neither has been previously associated with respiratory traits in patients.

Within the AFR cohort and African American (IBD-06) cluster, we uncovered novel associations between *DDHD2* and dysphagia ($P_{IBD-06,LOF}$ = $1.24\times10^{-6}$, $OR_{IBD-06,LOF}$ = 8.29 [3.68, 18.71]) as well as *EFCAB13* and essential hypertension ($P_{AMR,MIS}$ = $9.31\times10^{-7}$, $OR_{AMR,MIS}$ = 13.07 [4.53, 37.67]). These findings are consistent with prior studies – mutations in *DDHD2* have been shown to cause hereditary spastic paraplegia and symptoms of dysphagia[99], though not in African Americans, while methylation studies have prioritized *EFCAB13* as a risk factor for heart failure[100]. We detected an additional association between *ANKZF1* and peripheral vascular disease ($P_{AMR,LOF_MIS}$ = $1.56\times10^{-6}$, $OR_{AMR,LOF_MIS}$ = 58.83 [12.58, 275.12]), supported by a recent finding that *ANKZF1* acts downstream of *HIF-1α* to promote angiogenesis in the human mesenchymal stem cells under hypoxic conditions[103].

Next, in the EAS cohort, we observed a novel association between *EXOC1L* and glucose mass volume in plasma ($P_{EAS,LOF}$ = $3.76\times10^{-7}$, $OR_{EAS,LOF}$ = 6.45 [3.14, 13.22]). The exocyst complex, including *EXOC1* through *EXOC8*, has previously been identified as a regulator of insulin-stimulated glucose uptake in skeletal muscle cells[101], although *EXOC1L* has not been previously associated with insulin resistance or blood glucose. We further identified an association between *KIF2B* and nephritis/nephropathy in the Chinese/Korean (IBD-05) cluster ($P_{IBD-05,MIS}$ = $7.18\times10^{-7}$, $OR_{IBD-05,MIS}$ = 26.18 [8.97, 76.41]); though *KIF2B* has not been previously associated with renal

traits, other genes in the kinesin (KIF) superfamily have been implicated in renal pathology, such as *KIF2C* with clear cell renal cell carcinoma[102] or *KIF3A* with polycystic kidney disease[103].

Finally, we examined known risk associations to identify ancestry-specific patterns of deleterious rare variation. We examined two known *GBA1* rare LOF mutations that increase risk for Parkinson's disease[104], p.Glu365Lys and p.Thr408Met, in the AMR and EUR cohorts. Despite similar allele frequencies within each cohort, p.Glu365Lys was only significantly associated with increased risk in EUR ($P_{EUR,E365K}$ = 0.014; $P_{AMR,E365K}$ = 0.75), while p.Thr408Met was only associated with increased risk in AMR ($P_{EUR,T408M}$ = 0.212; $P_{AMR,T408M}$ = 0.0029), displaying evidence of ancestry-specific risk stratification (**Figure 5f**). We also found that the AMR cohort exhibited greater decreases in LDL cholesterol level, in response to additive LOF and deleterious missense risk variation in *PCSK9* ($OR_{EUR,LOF}$ = 0.51 [0.35, 0.73]; $OR_{AMR,LOF}$ = 0.33 [0.19, 0.57]), compared to the EUR cohort (**Figure 5g**). Further, LDL variants increase LDL in an ancestry indifferent manner, as is known[105] (**Supplementary Figure 9h**). These results underscore the importance of interrogating rare variants across different ancestries to capture cohort-specific genetic risk, with potential for guiding future precision medicine-based clinical interventions.

**Utilizing longitudinal EHR records to investigate GLP-RAs efficacy**

One of the main advantages of ATLAS is the consideration of dynamic changes over time. To illustrate this, we integrated all four study components (demographics, genetic ancestry, common and rare genetic variants) to study the efficacy of GLP-1 receptor agonists (GLP-1 RAs) for weight loss. We focused on semaglutide, which had the greatest number of prescriptions in the database (**Supplementary Figure 10a-b**), with 7,214 individuals having recorded a prescription (**Figure 6a**). As a general tendency, we observed a steady decrease in weight up to ~60 weeks of semaglutide treatment, consistent with previous findings[106–108] (**Supplementary Figure 10c**). We tested if the medication dose, route, sex, age and initial weight affect semaglutide efficacy, corroborating that dose and subcutaneous delivery versus oral delivery were positively correlated with weight loss as has been well established (**Figure 6b**; linear mixed model: $P_{Bonferroni} = 3.2\times10^{-62}$, rsq (explained variance) = 0.71, effect size = -0.6 ; route oral *vs.* subcutaneous: $P = 2.5\times10^{-20}$, rsq = 0.65, effect size = 1.9). We did not detect a significant effect of patient age and sex on efficacy.

Next, we asked if semaglutide's effects varied by broad-scale ancestry over the first 60 weeks of treatment. Ancestry and its interaction with time significantly influenced weight loss (**Figure 6c;** ANOVA on a linear mixed-effects model; ancestry: $P_{Bonferroni}$ = 0.002, Sum Sq = 120.5; ancestry×time: $P_{Bonferroni}$ = 0.0005, Sum Sq = 185.20). Further comparisons between EUR and the other populations revealed less weight loss in AMR, and a slower rate of weight loss in AMR and EAS (**Figure 6c**; linear mixed model; $ancestry_{AMR}$: $P_{Bonferroni} = 2.6\times10^{-2}$, estimate = 0.8; $ancestry_{AMR\times}$time: $P_{Bonferroni} = 4.6\times10^{-3}$ ,estimate = 45.6; $ancestry_{EAS\times}$time: $P_{Bonferroni} = 3.2\times10^{-3}$, estimate = 76.4).

Given limited evidence on the role of inherited factors in GLP1-RAs effectiveness[109], we used ATLAS to explore the contribution of common genetic variation to semaglutide-related weight loss. We first tested whether genetic scores related to BMI or Type 2 diabetes mellitus (DM2) could influence semaglutide efficacy and found no correlation between BMI PGS and semaglutide efficacy. However, we observed that weight loss was negatively correlated with DM2 PGS (**Figure**

**6e**; linear mixed model, $PGS_{High\ vs.\ Low}$: $P_{Bonferroni} = 2.8\times10^{-4}$, effect size = 0.96; $PGS_{Med\ vs.\ Low}$: $P_{Bonferroni} = 2.8\times10^{-3}$, effect size = 0.73). The same relationship was observed when simplifying the model using methods similar to previous approaches[109], using the maximum weight loss record for each participant (linear regression, $P_{Bonferroni} = 1.3\times10^{-2}$, beta = 0.31; **Supplementary Figure 10d**). As a second step, we conducted a genome-wide association (GWAS) meta-analysis to combine data across ancestries, but did not identify any significant loci (**Methods**; **Supplementary Figure 10e**).

Finally, we used WES to test if we could identify rare variant associations of semaglutide response within our cohort (Regenie[70]; **Methods**). Given the modest sample size, we focused on the EUR cluster and limited our test to only proteins whose plasma abundance was recently shown to be altered by semaglutide[110]. We identified a Bonferonni-corrected significant association of weight loss on semaglutide with one gene, *PTPRU* ($P_{Bonferroni} = 0.0047$, beta = -0.834) (**Figure 6f; Supplementary Figure 10f).** The negative effect size indicates that *PTPRU* activity is negatively associated with weight loss on semaglutide (variants affecting the function of *PTPRU* contribute to weight loss on semaglutide), concordant with the finding of decreased PTPRU levels as a response to semaglutide treatment[110]. This association involves 37 variants, of which one is common (rs2235937; nominally associated with weight loss ($P = 9.7\times10^{-3}$, beta = -0.06, SAIGE[111] in EUR), and the rest are rare (**Supplementary Table 11).** The overall frequency of rare variants in *PTPRU* varied across ancestries **(Supplementary Figure 10g**). This gene has no previous known functional relationship to weight loss or related metabolic functions. However, combined with the strong data from serum proteomics[110], our genetic analysis nominates this protein kinase and its pathways as novel candidates worth further investigation.

# Discussion

Understanding of the causes of a wide variety of biomedical conditions to improve healthcare outcomes and reduce the burden of disease has been made possible by large scale population studies, starting with the Framingham Heart study[112,113]. Health system biobanks continue to advance this line of research by providing cost effective, longitudinal data in large population cohorts[6,7,9,31,57,79,114].

Despite the known sources of errors and incomplete phenotyping in health system EHRs, multiple studies have shown that many of these factors can be mitigated, and robust analyses can be conducted using these data[7,8,31,115] Indeed, we were able to validate previously identified genetic associations tested in the UCLA ATLAS Community Health Initiative biobank based on EHR-derived phenotypes. This includes the application of polygenic scores, which we show can have high utility in identifying patients at high risk for common disorders, such as Type 1 Diabetes Mellitus, where > 40% of our cases lie in the top PGS decile. However, this predictive power, which is derived from EUR PGS, decreases dramatically in those with non-EUR ancestry across phenotypes, further emphasizing the well-articulated need for ancestral diversity in genetic studies of human disease[66–69]. We also leverage the heterogeneity of genetic ancestries in our population, > 30% of which are non-European, to validate and extend our knowledge of disease burden and genetic risk factors in ancestrally distinct populations.

Prior work has mostly relied on utilizing broad-scale genetic ancestry to understand how ancestry contributes to health risks[3,27,114]. Genetic ancestry can be measured on a continuum. While broad-scale ancestry discerns more ancient genetic variation via PCA, fine-scale ancestry discerns more

recent genetic variation via shared IBD[116]. Both resolutions of genetic ancestry are useful for identifying groups of individuals with differing disease risk, but different aspects of disease risk are identified depending on the resolution of genetic ancestry leveraged[56,57]. Identifying fine-scale ancestry clusters within our biobank has leveraged its diversity to enable findings that will benefit populations that are underrepresented in genetics and medical research, with the potential to improve ancestry-targeted screening for diseases[117]. For example, to our knowledge UCLA ATLAS consists of the largest known genetic data set of ancestrally Filipino individuals (n = 1,438), with the next largest genetic data set consisting of 1,028 individuals[118]. We identified novel findings within our Filipino cluster, which can be used to improve preventative and precision care for this population.

Through PheWAS across broad- and fine-scale ancestries, we identify numerous novel genetic associations, including rs17419569-C with asthma risk; rs74744741-C with GERD risk in the Mexican American (IBD-04) cohort; and rs7208565-T with intestinal disaccharidase deficiency across multiple EUR and AMR cohorts. By conducting ExWAS using burden masks of rare, predicted deleterious LOF and missense variants, we prioritize novel risk genes, including *CLN3* for cystic kidney disease in AMR, *EXOC1L* for blood glucose level in EAS, and *HNRNPA1L2* for breast cancer in the Northern European (IBD-01) cohort. Using our fine-scale ancestral clusters, we reveal differences in risk frequency – for instance, the combined South American (IBD-07) cohort carries the non-alcoholic cirrhosis risk variant *PNPLA3* rs738409-G at a higher frequency than the Mexican American cohort where this association was first discovered. Discerning common and rare variant risk frequency, as well as effect size, across fine-scale ancestral cohorts poses a powerful tool for precision medicine – offering opportunities to optimize risk stratification, alter screening guidelines, and tailor therapeutic interventions to individuals based on their ancestry-specific risk factors.

It has been well demonstrated that diagnostic misclassification due to differences in variant frequencies across ancestries is a serious risk in homogeneous populations[19]. We emphasize that having such diversity within a single regional biobank can help mitigate against confounders, not only caused by ancestral genetic differences, but non-genetic variables that differ across countries or geographically distinct biobanks. We demonstrated the application of longitudinal EHR to gain deeper insights into health outcomes over time. We showed that weight loss on semaglutide varies by broad-scale ancestry and genetically aligns with DM2 PGS. Leveraging omics data, we display the influence of genetic variants in a semaglutide-affected candidate protein on weight loss. These were enabled by integrating dynamic EHR changes with detailed prescription information, including the medication dose and route. Lastly, our analysis of ACMG clinical variants further supports that preponderance of EUR individuals in clinical genetic databases reduces power to detect true associations due to allele frequency differences, which can lead to misclassification or loss of power[19,20]. Our findings highlight new fine-scale populations with generally low or high ACMG putative damaging variants for further investigation.

While the findings are valuable, certain limitations should be considered. First, due to the nature of EHR, our study captures a partial view of patient care, potentially affecting the completeness of our findings. Second, our analysis focuses on the risk of receiving a diagnosis for a disease, which is related but not equivalent to the underlying risk of disease development. This may influence how our findings should be interpreted in the context of true disease incidence. It is our

expectation that ATLAS will continue to expand, improving its power and utility, but will also become an engine for clinical intervention, allowing researchers to rapidly identify individuals for genetic (and non-genetic) health studies and the implementation of precision medicine as the field continues to move forward.

# Methods

### Cohort description

ATLAS enrollment reflects a modest over-representation of females by self-reported sex (female 55.4%, male 44.6%), especially among patients aged 20-60. The participants are older than the population as a whole (median ATLAS age: male, 61; female, 56, **Figure 1b; Table 1**). Medical morbidity was also notably more widespread in the biobank population compared to all adult UCLA patients. This point was evidenced by a significantly larger number of diagnoses in the biobank population compared to all UCLA patients (16 vs. 10.5 mean ICD codes per patient, one year after collection; **Supplementary Table 1**). The biobank population experienced a substantially higher number of diagnoses across the most prevalent medical condition categories, including endocrine, cardiovascular, musculoskeletal, gastrointestinal, neoplasms, neurological, respiratory, mental, genitourinary and infections (**Supplementary Figure 2**). For instance, endocrine/metabolic and cardiovascular diseases were diagnosed in 52.0% and 44% of the biobank participants versus 38% and 36% of non-biobank adults, respectively. Biobank participants also experienced substantially more neoplasms (30% vs. 19%) (**Supplementary Figure 2c-d**). This likely led to significantly more medical encounters in the biobank population (119 vs. 64 mean total medical encounters per patient, and 16.2 vs. 12.3 mean inpatient days), which is consistent with expectations, as greater numbers of clinical visits facilitate the passive blood collection that powers our biobank. The most common diagnoses included hypertension (30% of participants), hyperlipidemia (25%), GERD (18%), anxiety disorder (17%), and depression (15%) (**Supplementary Figure 2a**).

As time progresses, more diagnoses have an opportunity to be added, and the rate of diagnoses is consistent across a wide spectrum of organ systems (1-2 years after collection) (**Supplementary Figure 1b**). We retrieved the latest results of the 39 most common laboratory tests: blood count, lipid panel, metabolic panel, HbA1c, and vitamin D,25-Hydroxy to explore test frequencies and variation in results. The largest variability in adults was observed for bilirubin (median = 0.4, IQR = 0.3), followed by triglycerides (median = 90, IQR = 66), alanine aminotransferase (median = 22, IQR = 14), neutrophils (median = 3.73, IQR = 2.07), and LDL cholesterol (median = 97, IQR = 50). Similarly, we retrieved information on the 50 most abundant filled prescriptions to learn about treatment and prescribing tendencies. (**Supplementary Table 2**). As hospital visits and surgeries were the most common types of encounters in the biobank population, it is not surprising that among the most prescribed generic medications were Acetaminophen (n = 738,061 prescriptions), Ondansetron HCl (n = 636,165), Propofol (n = 633,830), and Lidocaine HCl (n = 492,838) (**Supplementary Table 3**).

### Array data and imputation

Array genotypes were obtained using the Global Screening Array. All data was mapped to GRCh38 and dbSNP, build 147[119]. Common haplotypes and variants were imputed using the TOPMed Freeze 5 panel[26,27] using 668,127 observed SNPs, resulting in a total of 50,757,223 high-quality called genotypes variants following imputation, an average of 2,048,050 per

individual. Details regarding QC and imputation procedure were described before[26]. Minimal QC was applied to the imputed genotypes. We retained only non-duplicated, bi-allelic variants with high imputation quality (R2 > 0.7), a minor allele frequency (MAF) > 0.1%, and a missing rate < 5%. Genotypes with a missing rate > 5% were excluded from further analysis. Concordance between genotypes determined by observed array variants and targeted Illumina sequencing was approximately 99.6%. The *vcf-compare* command in VCFtools (v0.1.16) was used to compare genotypes between the two platforms[120].

**Retrieving phenotype data**

Demographic information, vitals, lab tests, ICD codes, and medication prescriptions were retrieved from the UCLA Data Discovery Repository (DDR), established on March 2, 2013[112], containing deidentified patient EHR data from our health system.

Demographic, vital, and lab data was up-to-date as of Aug 24, 2024. For vitals and encounter counts, only records from in-person encounters (hospital visits, appointments, surgeries, office visits, and walk-ins) were considered. Both the latest records and median values (medians across all encounters, the latest three years, and the latest year) were calculated to reduce possible recorded typos.

Yearly encounter numbers were retrieved in 2024. We included only complete yearly data up to 2023 to ensure consistency and avoid partial data from 2024. Hospital encounters during the COVID pandemic years (2019-2022) were excluded.

For associations involving clinical phenotypes, both ICD-9 and 10 were extracted. The ICD-9 codes were mapped to phecodes using phecode Map v1.2[59], while ICD-10 codes were mapped using phecode Map v1.2b1[60], both of which were obtained through the PheWAS catalog[61]. In total, there were 1,253 phecodes with over 100 cases in ATLAS. For each phecode, individuals were classified as cases if they had at least two occurrences of the same phecode, with occurrences spaced at least 30 days apart. Controls were defined as individuals who had no record of the phecode in their EHR data and at least two recorded encounters in the system. Undecidable individuals were excluded.

**ATLAS EHR baseline characteristics and its comparison to non-biobank UCLA patients**

As all samples were collected incidentally, there was interest in characterizing the population, especially in comparison to non-biobank UCLA patients. To characterize the non-biobank UCLA patients while mitigating time-dependent confounding, we included a subpopulation with an encounter within one year of ATLAS launch date. Encounters could be either inpatient or outpatient, and we summarized the relative proportion of patients with only outpatient encounters or at least one inpatient encounter (there were no patients with only inpatient encounters).

To identify clinical phenotype patterns in ATLAS and to compare these to patterns of non-biobank UCLA patients, we used all ICD codes from each patient to encapsulate past and present conditions, subject to practical challenges[121]. For the ATLAS population, we captured a snapshot using their encounter closest to their biobank sample collection date. For the non-biobank patients, we used their closest encounter to the ATLAS launch date. These encounters represented the baseline encounters for both populations. The ICD codes at these encounters were converted to phecodes and phecode groups[59–61,122] to represent meaningful categories of disease. Phecodes were extracted using pandas v2.2.2 with Python v3.9.19. We reported the unique phecodes with a prevalence ≥5% in the UCLA ATLAS population. ICD codes at the

baseline encounters were also used to calculate Charlson and Elixhauser Comorbidity Indices [36,37,123] both measures that predict mortality. The scores were derived using the comorbidity R function v1.0.7[124] with R v4.1.2.

In addition to characterizing patient populations at a baseline time, we also described differences in how they changed over time - an important consideration when assessing relative disease burden across populations. Patients were enrolled in UCLA ATLAS and were encountered within the health system at different times; therefore, we controlled the interval over which we measured change. We identified patients with at least one encounter between one and two years after their baseline encounter. From this subpopulation we obtained patients' first encounter within this window and summarized the cumulative encounters that occurred between the two times. We also used the ICD codes at the second encounter to derive new comorbidity scores, as well as extract updated phecodes. The change in comorbidity scores and increases in phecode prevalence provided insight into the evolving disposition of the UCLA ATLAS population over time.

**Broad scale genetic ancestry**

The genetic ancestry of individuals in the ATLAS dataset was estimated by assessing their proximity to the centroids of 1000 Genomes superpopulations in principal component (PC) space. The top 20 PCs were calculated using the bed_projectPCA bigsnpr[119] R function v1.12.2 with default parameters. For every individual, the Euclidean distance to the centroids of the five broad-scale populations (AMR, AFR, EUR, EAS, SAS) was computed. Individuals were assigned AMR or AFR ancestry if the nearest centroid corresponded to one of these populations, as these groups are well-separated in PC space. For EUR, EAS, and SAS ancestries, which exhibit more genetic overlap, a stricter distance threshold was enforced to minimize misclassification. Specifically, an individual was assigned to one of these ancestries if their distance to the nearest centroid was less than a scaled threshold, calculated as: Threshold=max(dist)×min(FST)/max(FST)×0.5 where max(dist) is the largest squared distance among centroids and min(FST)/max(FST) accounts for genetic differentiation. Individuals who could not be assigned to any ancestry cluster under these criteria were labeled as "admixed/unknown." Visualization was performed using Boutros Plotting General (BPG) R package v.7.1.0[125].

**Broad-scale ancestry and clinical phenotype associations**

To test differences in disease diagnosis across ancestries, phecodes (retrieved as described in **Retrieving Phenotype Data**) were associated with genetic ancestry groups using logistic regression, adjusting for age (age at diagnosis for cases and the latest age for controls) and genetic sex if applicable. Differences in encounter numbers and comorbidity index were tested using ANCOVA, adjusted for genetic sex and age. For all parts, patients under 18 years old, with ambiguous sex or with "unclassified" genetic ancestry were excluded. Visualization was performed using the BPG R package v.7.1.0[125]. The following disease-phecodes pairs were used to define disease diagnosis: Type 2 diabetes-Type 2 diabetes, Sleep apnea-Sleep apnea, Essential hypertension-Essential hypertension, Hyperlipidemia-Hyperlipidemia, Anxiety disorders-Anxiety disorders&Anxiety disorder&Generalized anxiety disorder, Asthma-Asthma, Parkinson's disease-Parkinson's disease, Type 1 diabetes-Type 1 diabetes, Schizophrenia-Schizophrenia, Crohn's disease-Regional enteritis, Chronic kidney disease-Chronic kidney disease, Stage I or II, Multiple sclerosis-Multiple sclerosis, Major depressive disorder-Major depressive disorder, Cerebrovascular disease-Cerebrovascular disease, Atrial fibrillation-Atrial fibrillation, Hypercholesterolemia-Hypercholesterolemia, Hyperlipidemia-Hyperlipidemia,

Coronary atherosclerosis-Coronary atherosclerosis, Hypertrophic obstructive cardiomyopathy-Hypertrophic obstructive cardiomyopathy, Myocardial infarction-Myocardial infarction, Systemic lupus erythematosus- Systemic lupus erythematosus , Gout-Gout, Bipolar-Bipolar, Psoriatic arthropathy-Psoriatic arthropathy, Epilepsy-Epilepsy, Neurofibromatosis-Neurofibromatosis, Dementias-Dementias, Obesity-Obesity, Obsessive-compulsive disorders-Obsessive-compulsive disorders, Autism- Autism, Migraine- Migraine, Alzheimer's disease- Alzheimer's disease, Coronary atherosclerosis-Coronary atherosclerosis, Posttraumatic stress disorder-Posttraumatic stress disorder.

**PGS analysis**

PGS were calculated using array data after imputation in EUR individuals. Related individuals based on their genetic similarity were excluded (defined using PLINK v2.0a[109] with the relatedness coefficient --king-cutoff 0.05). PGS were calculated using pgsc_calc[126] with the default setting but setting the –min_thres to 0.65. Logistic regression was used to associate every PGS with the corresponding trait based on phecodes (see Retrieving phenotype data)**.** The following PGS model IDs from the PGS catalog[127] and their phecode pairs were tested: PGS002250-Malignant neoplasm of ovary, PGS003766-Cancer of prostate, PGS000004-Malignant neoplasm of female breast&Breast cancer, PGS001794-Thyroid cancer, PGS000079-"Melanomas of skin, PGS002264-Pancreatic cancer, PGS003395-Colorectal cancer&Colon cancer, PGS000729-Type 2 diabetes, PGS002025-Type 1 diabetes, PGS004254-Regional enteritis, PGS004699-Multiple sclerosis, PGS000134-Schizophrenia, PGS004760-Major depressive disorder, PGS001806-Malignant neoplasm of testis, PGS004687-Malignant neoplasm of bladder & Cancer of bladder, PGS000039-Cerebrovascular disease, PGS004526-Essential hypertension, PGS004706-Atrial fibrillation, PGS004784-Hypercholesterolemia, PGS002029-Hyperlipidemia, PGS003726-Coronary atherosclerosis, PGS000739-Hypertrophic obstructive cardiomyopathy, PGS004528-Myocardial infarction, PGS000803-Systemic lupus erythematosus, PGS001789-Gout, PGS002786-Bipolar, PGS000198-Psoriatic arthropathy. For prostate and testicular cancer, only males were included, and for breast and ovarian cancer, only females. An adjustment was made for age at diagnosis for cases and the latest age for controls, genetic sex when both sexes were included, and the first ten genetic principal components (PCs). For Figure 3, the top end bottom PGS deciles compared with the 5$^{th}$ decile were considered, testing only EUR individuals. FDR was used for multiple testing correction. The same process was repeated for other ancestries as presented in the supplementary material. Visualization was generated using BPG R package v.7.1.0[125].

**Fine-scale ancestry pre-processing and quality control**

***Data***. ATLAS array data was merged with genotyping data from the 1000 Genomes Project[128], the Simons Genome Diversity Project[129], and the Human Genome Diversity Project[130]. BCFtools[131] annotate was used to harmonize variant reference SNP ID (RSIDs), and BCFtools[131] norm with a GRCh38 genome reference was used to standardize the genotyping data. Sites or individuals with more than 1% missing were removed using PLINK[132] --mind and --geno. Only SNPs with MAF > 1% across all individuals were kept.

***Phasing.*** SHAPEIT5[133] with default parameters and the distributed GRCh38 map files were used to phase genotyping data, one chromosome at a time.

***Identity-by-descent calling and processing***. A custom Python script that converts PLINK bed files to PLINK ped/map[132] files while conserving phasing information was used to convert

genotyping data. Centimorgan data for the map files were generated using the same genetic map data in SHAPEIT5.

Identity-by-descent segments were called using iLASH[134] with the following parameters: slice_size 350, step_size 350, perm_count 20, shingle_size 15, shingle_overlap 0, bucket_count 5, max_thread 20, match_threshold 0.99, interest_threshold 0.70, min_length 2.9, auto_slice 1, slice_length 2.9, cm_overlap 1 and minhash_threshold 55. Identity-by-descent was called one chromosome at a time.

***Identity-by-descent quality control.*** Identity-by-descent segment outliers were removed as described in Belbin et al.[57] and Caggiano et al.[56]. Segments overlapping centromeres or the human leukocyte antigen (HLA) region were removed. Regions that may have false positive identity by descent were identified using the following process and removed: total identity by descent per each SNP was identified by summing across all identity-by-descent segments that overlapped each SNP; SNPs with a total identity by descent greater than or less than three standard deviations from the genome-wide mean were removed.

***Cluster identification.*** To identify clusters, we followed the approach of Caggiano et al.[56] and Dai et al.[58] and applied Louvain clustering[135]. An undirected network is generated based on pairs of individuals who share identity-by-descent segments: nodes are the individuals, edges are the total, genome-wide identity-by-descent shared as the edges. We used the Python package, NetworkIt[136], to iteratively run Louvain clustering four times to detect fine-scale clusters.

***Cluster merging.*** To avoid redundancy and maximize sample size, clusters were merged in two stages to produce a final set of consensus fine-scale cohorts. First, following Caggiano et al.[56]and Dai et al.[58], we computed pairwise Hudson's $F_{ST}$ using PLINK v2.0a[137] across 378 clusters identified from the fourth layer of Louvain clustering. After removing clusters with fewer than 10 individuals, to avoid unreliable $F_{ST}$ estimates, we merged the remaining 356 clusters into 67 clusters if the pairwise $F_{ST}$ was less than 0.001.

In the second stage, we refined clusters using IBD sharing. For each cluster pair, we examined all inter-cluster individual pairs to calculate: (1) $IBD_{mean}$, the average cM shared, and (2) $IBD_{prop}$, the proportion of pairs sharing at least 3cM of IBD as detected by iLash. We defined a composite metric for cluster merging, $IBD_{weighted} = IBD_{mean} \times IBD_{prop}$, that captures both the extent and prevalence of genomic IBD sharing. Next, clusters were sorted from smallest to largest number of ATLAS participants. For each cluster, we computed the $IBD_{weighted}$ score with all larger clusters and calculated the mean of these pairwise values. The cluster was then merged into the most similar larger cluster – i.e., the one with the highest $IBD_{weighted}$ score-only if that score exceeded the mean. Otherwise, the smaller cluster was retained independently. This process merged 14 small clusters into larger parent clusters, resulting in a total of 36 fine-scale clusters with ≥ 30 individuals each for downstream analyses[56]. These fine-scale clusters were assigned unique identifiers (IBD-01 through IBD-36) and manually annotated with labels to ease interpretation.

***Cluster labeling.*** We primarily relied on reference individuals, described in "Fine-scale Ancestry Pre-processing and Quality Control", to add labels to clusters. When clusters did not contain reference individuals or were heterogeneous, we utilized patient-reported race, ethnicity, preferred language, and religious affiliation (in order of priority) to inform our cluster labeling. These aspects are not caused by identity-by-descent segment sharing but can be indicative of a shared culture for individuals within a cluster; these shared practices can influence a group's

demography, environment, and disease risk[138]. Our labels are not definitive and are our best attempt to generate informative assignments for each cluster.

**Fine-scale ancestry and clinical phenotype associations**

We used logistic regression to model the association between fine-scale ancestry assignment and phecode prevalence, estimating OR with 95% confidence intervals using the logistf R package v1.26.0 for Firth's bias-Reduced penalized-likelihood logistic regression[139]. Differences were tested between every fine-scale ancestry cluster with at least 100 participants and all other ATLAS individuals. Adjustment was made for age at diagnosis for cases, and the latest age for controls, and for sex when applicable. Phecodes with diagnoses in at least 100 patients were tested, resulting in 1,253 phecodes.

The same model was used to test differences in cardio-metabolic diseases using appropriate phecodes, for each fine-scale cluster within the same broad-scale continental ancestry. For this goal, clusters were assigned to broad-scale ancestries based on the predominant ancestry match among individuals within each cluster. For this, the largest population cluster within each broad-scale ancestry was used as the reference level for associations. FDR was used for Multiple testing correction. The data were visualized using BPG R package v.7.1.0[125].

**PheWAS**

PheWAS were conducted using the Regenie v4.0 framework[70] for all five broad-scale cohorts (AMR, AFR, EAS, EUR, SAS) and fifteen fine-scale cohorts with at least 400 individuals (IBD-01 through IBD-15). Association testing was performed separately within each cohort for 1,437 binary traits and 41 quantitative traits, including ICD-derived diagnoses and clinical laboratory measurements (see “Phenotype Definitions”). Samples were restricted to those in predefined inclusion lists (–keep), and trait-specific covariates were provided via –covarFile, including age, sex (modeled categorically with –catCovarList sex), BMI, and the top 10 genetic principal components. Quantitative traits were divided into two analysis groups based on missingness patterns, following Regenie’s recommendation that traits with similar levels of missing data are modeled together in Step 1 (**Supplementary Figure 11**).

***Ridge regression (step 1).*** For each cohort, unimputed array genotypes in bed format were filtered using PLINK v2.00a[137] to produce a minimal set of variants suitable for estimating genome-wide polygenic effects through ridge regression in step 1 of Regenie. We included autosomal variants with call rate ≥99% (--geno 0.01), minor allele frequency ≥1% (--maf 0.01), and Hardy-Weinberg equilibrium $p > 1\times10^{-}15$ (--hwe 1e-15). Additional linkage disequilibrium (LD) pruning was applied using a sliding window of 1000 SNPs, advanced by 100 SNPs, with an r2 threshold of 0.9 (--indep-pairwise 1000 100 0.9), producing an average of 389,209 SNPs per cohort. For binary traits, step 1 was run with a minimum case count of 50 (--minCaseCount 50). For quantitative traits, step 1 was run with Rank Inverse Normal Transformation (--apply-rint) enabled to stabilize variance across lab values with differing distributions. For all traits, a block size of 1000 (--bsize 1000) was used, leave-one out cross validation (--loocv) was enabled, and sex was marked as a categorical covariate (--catCovarList sex).

***Association testing (step 2).*** Prediction files from step 1 for each trait and imputed genotypes in BGEN format were used to perform association testing in step 2 of Regenie. For binary traits, a minimum allele count of 20 (--minMAC 20) and minimum case count of 50 (--minCaseCount 50) were required, and Firth approximation was performed (--firth --approx) using a p-value threshold of 0.01 (--pThresh 0.01). For quantitative traits, a minimum allele count of 20 (--minMAC 20) was

required and Rank Inverse Normal Transformation (--apply-rint) was applied. For all traits, a block size of 500 (--bsize 500) was used and sex was marked as a categorical covariate (--catCovarList sex).

**WES sample preparation**

Genomic DNA libraries were created by enzymatically shearing high molecular weight genomic DNA to a mean fragment size of 200 base pairs. Multiplexity of exome capture and sequencing was achieved by adding unique asymmetric 10-bp barcodes to the DNA fragments of single samples during library amplifications. Equal molar amounts of DNA samples were pooled for exome capture using a slightly modified version probe library of xGen exome research panel from Integrated DNA Technology (IDT). After PCR amplification and quantification of the captured DNA, samples were multiplexed and loaded to Illumina sequencing machines for sequencing to generate 75 base pair paired end reads. The samples in this study were sequenced using the Illumina sequencing machines including NovaSeq 6000 with S2 or S4 flow cells and the NovaSeqX with 25B flow cells.

**WES read alignment and variant detection**

Sequencing reads in FASTQ format were generated from Illumina image data using bcl2fastq program (v2.20, Illumina). Whole exome reads alignment and germline small variant detection were conducted using the Original Quality Functionally Equivalent (OQFE) protocol described in Krasheninina et al., 2020[140]. Briefly, raw read files (FASTQ) were mapped to the GRCh38 reference obtained from https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/ using BWA-MEM v0.7.17-r1188 in an alt-aware manner[141]. Duplicate reads were then marked with Picard v2.21.2[142]. The final CRAM files were compressed with SAMtools v1.2[131]. Germline variant detection was performed on each CRAM using a Parabricks accelerated version of DeepVariant v0.10.0 with a custom WES model [143], resulting in a sample-level gVCF (genomic VCF). Per-sample gVCFs were merged with GLnexus v1.4.3[144] into joint-genotyped multi-sample project-level VCF (pVCF). Variant prediction was restricted to the exome capture region and the 100 base-pairs buffer on each side of the target regions. The WES was of high quality and reached an average coverage of 41.6-fold with a minimum of 24.2-fold, 90% having ≥ 34.7-fold in targeted regions, consistent with a previous publication[145]. The pVCF was converted to a plink file format using PLINK 1.9[132] for downstream analyses.

These data underwent extensive quality control to ensure the absence of contamination, duplication, and other technical errors, as well as sufficient read depth to guarantee reliability and accuracy. Genetic duplicates were defined based on the aggregated genotype data of all sequenced samples. The gender was predicted based on the ratio of read coverage on chromosome Y over the whole exome read coverage. VerifyBamID v1.1.3[146] was used to estimate sample contamination. Samples were excluded if they showed gender discordance, duplication, cross-Indvidual contamination > 5%, or coverage < 20X in more than 20% of target regions.

Alignment quality metrics were generated using Picard CollectHsMetrics v2.27.4[142] and SAMtools stats v1.15.1 with default settings[114]. MultiQC v1.27.1[147] was used to systematically aggregate sample-level quality metrics. RTGtools v 3.12.1[148] was used to assess variant QC metrics, including the number of variants for SNPs, small insertions and deletions; genotype counts; heterozygous-to-homozygous ratios for each variant type and transition/transversion (Ti/Tv) ratio.

Genotypes were further confirmed using an alternative approach[149] using 500 random samples yielding a high-concordance on-target sites (median concordance of 99.5% for SNPs and indels).

**WES summary statistics**

For WES-based analyses, we used samples that also had array genotyping data, which was necessary to consistently define genetic ancestry. We annotated variants using Ensembl Variant Effect Predictor (VEP) v.112[90] for 58,387 samples assigned to the EUR, AFR, SAS, EAS, or AMR ancestry class (*i.e.* excluding unclassified genetic ancestry samples). Annotations were performed using the GRCh38 cache and corresponding reference FASTA file. We limited our analysis to autosomal chromosomes to avoid technical artifacts in variant calling caused by the differences in ploidy between males and females, as well as the high-sequence similarity between the X and Y chromosomes in certain regions[150]. Counts for single nucleotide variants, indels, multi-allelic, synonymous, missense, and LOF variants were restricted to whole exome sequencing (WES)-targeted regions. Consistent with a previous UK Biobank WES study[151], we classified LOF variants as those with the following consequences: stop_gained, start_lost, splice_donor, splice_acceptor, stop_lost, and frameshift. To increase reliability, we further restricted LOF variants to those flagged as high confidence by LOFTEE[89]. For multi-allelic variants, the predicted function for each alternate allele was determined using the --pick-allele option based on the default ordered set of criteria defined by VEP[90]. To determine the number of variants with MAF < 1% while accounting for ancestry-specific allele frequency differences, we retained variants with MAF < 1% in at least one ancestry group.

For multi-allelic variants, we defined the minor allele as the second most common allele (including the reference) and classified the variant as rare if the cumulative allele frequency of all alternate alleles was < 1%. For rare multi-allelic variants, we incremented a sample's count only if it carried an alternate allele with MAF < 1%. In multi-allelic cases where the minor allele was the reference or the alternate alleles had AF ≥ 1%, we incremented a sample's count if it carried any alternate allele. For rare functional variants (e.g., missense), we incremented a sample's count only if it carried an alternate allele with MAF < 1% corresponding to that functional category.

**Carrier frequency of single ClinGen alleles by fine-scale ancestries (cluster membership)**

To identify clusters where individuals were enriched for carriers of monogenic variants associated with disease, we identified pathogenic variants based on different sets of criteria. First, as a positive control, we identified variants, *via* literature search, where variants are known to be enriched within certain populations. Next, we used pathogenic variants that were curated by experts within the field and underwent stringent review to be considered pathogenic from ClinGen[86]. ClinGen variants filtered to identify P/LP variants with autosomal dominant inheritance, autosomal recessive inheritance, or semidominant inheritance, resulting in 3,521 variants associated with 204 conditions overall according to ClinGen. We removed one *HNF4A* ClinGen variants due to high allele frequencies (MAF>0.01).

To test enrichment of single alleles across clusters, “carriers” are identified as individuals who carry at least one allele of a pathogenic variant within a gene associated with a monogenic condition. We then calculated carrier frequency as the number of identified carriers over the total number of individuals with WES available and were unrelated using a kinship coefficient of 0.05 based on PLINK v2.0a[137] king-cutoff. To identify clusters with a significantly different carrier

frequency of variants, we applied a Fisher's Exact test to the carrier frequency of each cluster against the carrier frequency of the other clusters grouped together. We then used FDR corrections for p-values generated from all Fisher's exact tests, and selected significant clusters where at least 5 carriers were identified and FDR < 0.05.

**Differences in total ClinGen allele frequency across broad- and fine-scale ancestries**

The list of all ClinGen[86] P/LP variants identified in any ACMG[88] genes was extracted as described above. Variants were labeled as missense or LOF, based on VEP v.112[90] annotations. Missense variants were defined for 'missense_variant' variants according to the 'consequence' VEP output column. LOF was defined for high-confidence "HC" LOF variants based on LOFTEE[89]. All variants were rare across broad-scale ancestries. A WES plink file with all individuals was filtered to include only the listed variants. Then, this file was broken into ancestry groups using PLINK v2.0a[137]. Related individuals were removed. The frequency of each allele in every population group was calculated with the PLINK v2.0a –freq command, and total frequencies were summed-up for rare missense and LOF variants separately. To test differences in the allele counts across populations, allele dosages were calculated with PLINK v2.0a[137] --recode A option. Dosages were summed to calculate the total alternative (ALT) allele count within each population. The number of reference (REF) alleles was defined as twice the number of individuals in the group minus the number of ALT alleles. Fisher's Exact tests were applied to test differences between the number of REF and ALT alleles in each population compared to all other individuals not assigned to that specific group. Bonferroni correction was applied for multiple testing correction. Plotting was done using BPG R package v.7.1.0[125].

**Variant annotation using a consensus of computational tools**

Variants from exome sequencing were annotated using VEP v.112[90] with the dbNSFP v4.9a[153] and LOFTEE[89] plugins installed. The LOFTEE high-confidence "HC" flag was used to select for LOF variants predicted to have deleterious effects.

Missense variants were assigned a 9-point deleteriousness score based on a consensus of nine missense deleteriousness prediction toolkits, similar to methods described in prior biobank-scale rare variant studies[154]. We used the Critical Assessment of Genome Interpretation (CAGI) project[91] to prioritize well-performing tools not trained on the same features. We selected five meta-predictors – ClinPred[155], MetaRNN[156], BayesDel_addAF[157], VARITY_R[158], REVEL[159] – and four stand-alone predictors – AlphaMissense[160], MutPred2[161], VEST4[162], ESM-1b[163] – for use in our analysis. We assigned each variant a binary score per tool based on dbNSFP rank scores: 1 if the variant's score exceeded the threshold score for being more likely a deleterious ClinGen or ClinVar variant than a background variant (**Supplementary Figure 12)**, and 0 otherwise. Summing these binary scores produced a deleteriousness score ranging from 0 to 9, with predicted damaging missense variants scoring ≥ 5 retained for downstream analysis.

**ACMG putative damaging variant distribution across ancestries**

A WES PLINK file was filtered to include computationally predicted LOF and predicted damaging missense variants (see above) in ACMG genes[88], providing a more comprehensive evaluation than the one based solely on ClinGen P/LP variants, which included only 17 genes. A WES plink file with all individuals was filtered to include only the ACMG putative damaging variants, and the file was split into ancestry groups (broad- and fine-scale ancestries) using PLINK v2.0a[137]. Related individuals were removed. Allele frequencies were calculated with PLINK v2.0a, and only rare variants (MAF < 1% in all broad-scale ancestries) were kept. Allele dosages per individual

were extracted using PLINK v2.0a, and the total rare LOF and predicted damaging missense alleles were counted separately per patient. First, the differences between the total numbers of REF and ALT alleles across ancestries were evaluated with a Fisher's Exact test as described above (ClinGen ACMG analysis). Second, a Mann-Whitney U test with a Bonferroni correction was applied to test the difference in the distribution of rare LOF and predicted damaging missense counts per individual between any broad- or fine-scale group compared to all others. **Figure 5d** presents the Mann-Whitney U test results, with statistically significant differences indicated by an asterisk (*). Visualization was done with BPG[125].

**ExWAS**

ExWAS were conducted using the Regenie v4.0 framework[70] for all five broad-scale cohorts (AMR, AFR, EAS, EUR, SAS) and fifteen fine-scale cohorts with at least 400 individuals (IBD-01 through IBD-15). Within each cohort, imputed genotype dosages in BGEN format (--bgen) and prediction scores from PheWAS step 1 (--pred) were used to conduct gene-based association testing for binary and quantitative traits across 17,676 genes. Samples were restricted to those in predefined inclusion lists (--keep), and trait-specific covariates were provided via --covarFile, including age, sex (modeled categorically with --catCovarList sex), BMI, and the top 10 genetic principal components.

***Genotype QC (step 2).*** Whole-exome genotypes in bed format were filtered using PLINK v2.0a, retaining variants with a call rate ≥ 90% (--geno 0.1), minor allele count ≥ 1 (--mac 1), and Hardy-Weinberg equilibrium p-value > $1\times10^{-15}$ (--hwe 1e-15). In addition, 784,548 variants overlapping low-complexity regions (LCRs) were excluded prior to analysis. Briefly, SNP positions were extracted from the .bim file, intersected using bedtools v2.29.1[164] with annotated LCRs from the Genome in a Bottle Consortium[165], and filtered from the genotype files, resulting in 12,326,160 exome variants retained for burden testing.

***Gene-based association testing (Step 2).*** Filtered whole-exome genotypes in BGEN format and prediction files from PheWAS step 1 were used to perform gene-based association testing. Regenie-style annotation, set, and mask files for predicted deleterious LOF and missense variants (see “Variant Annotation”) were created programmatically using Python v3.11.9 with the polars v1.2.1 package. For binary traits (--bt), a minimum case count of 50 (--minCaseCount 50), minimum minor allele count of 5 (--minMAC 5), and block size of 1000 (--bsize 1000) were enforced. Firth logistic regression with saddlepoint approximation (--firth --approx) was applied for variants with $p < 0.01$ (--pThresh 0.01). For quantitative traits (--qt), rank inverse normal transformation (--apply-rint), a minimum minor allele count of 5 (--minMAC 5), and block size of 500 (--bsize 500) were enforced. Regenie’s implementation of the RGC gene-based p-value test (–rgc-gene-p) was enabled for all analyses. Additionally, variants were binned by minor allele frequency using 1% bins (--aaf-bins 0.01), and SNP-level membership for each burden mask was recorded (--write-mask-snplist).

***Variant-based association testing (step 2).*** In addition to gene-based analyses, single variant association testing was performed for selected predicted deleterious LOF and missense variants using filtered whole-exome genotype dosages in BGEN format and prediction scores from PheWAS step 1. For both binary and quantitative traits, tests incorporated the same set of covariates (age, sex modeled categorically, BMI, and the top 10 genetic principal components) and enforced a minimum minor allele count of 5 (--minMAC 5). For binary traits, Firth logistic regression with saddlepoint approximation was applied for variants with $p < 0.01$, along with a

block size of 1000, while quantitative traits were evaluated using rank inverse normal transformed phenotypes with a block size of 500.

**Semaglutide investigation**

***EHR data curation***. ATLAS GLP1-RAs prescriptions, including medication names, start and end dates, discrete dose, usage instruction, strength and route (oral or subcutaneous), were retrieved and grouped based on the following simple generic names: dulaglutide, semaglutide, liraglutide, exenatide, albiglutide and lixisenatide. For all statistical analyses, only semaglutide users were considered. Usage start dates were defined based on the earliest prescription start date for each patient. In the case of a missing start date, the prescription ordering date was used instead (in most cases, these two fields were identical). Overlapping medication periods were handled such that when a new prescription started, the previous one was considered to have ended. Similarly, missing end dates were determined using the start date of the next prescription when available. In the case of completely overlapping prescriptions with different instructions, the combination of both routes and the weighted average dose was considered. If discrete dose information was missing, the medication dose was extracted from the instruction's free text field. If the instructions also omitted the dose information, it was imputed for a given medication type, based on the weighted dose average from the entire cohort on the corresponding type. Initial weight and BMI were defined as the median of all available weight or BMI measurements recorded from in-person visits, within 6 months prior to the first prescription start date. Not relying solely on a single data point helps minimize the likelihood of typos and errors in the EHR. The percentage of weight change was obtained for every weight measurement recorded on in-person visits within the period of active prescriptions between 4-60 weeks in total on semaglutide. The medication dose at each time point was defined as the weighted sum of all medication doses (doses multiplied by number of prescription weeks) by the measurement date. In case both oral and subcutaneous medications were used within a period, a combined route category was defined.

***Statistical analysis***. To identify the overall weight loss patterns across time, the FPCA R function from the fdapace package v.0.6.0[166,167], which is suited to plot smoothed longitudinal data with repeated measurement, was used. This identified a consistent weight loss pattern up to 60 weeks, and sparse data points beyond ~150 weeks. Thus, we restricted all analyses to this period.

For all analysis parts that involved longitudinal data with repeated measurements, a linear mixed model with the bobyqa optimizer and an increased function evaluation limit (maxfun = 10000) was used (lmer R function; the lmerTest R package v.3.1.3[168]). In each case, ANOVA was used to identify differences between models to define the best way to model a potential nonlinear relationship between weeks and weight loss, when treating weeks as a fixed effect, based on the Akaike Information Criterion (AIC). In some cases, the best model was achieved using a restricted cubic spline (RCS) with the rcs R function from the rms package (v.7.0.0), applied to weeks. In other cases, a polynomial function of weeks provided a better fit. To plot smoothed longitudinal data with 95% CI, the fitted model values (excluding covariates) and 95% lower and upper confidence bounds, were extracted using the visreg package[169] v.2.7.0 and visualized using the BPG package[125].

For testing the effect of non-genetic factors, fixed effects were defined for the medication dose, route, sex, age, initial BMI, 10 first genetic PCs (excluding PC6 due to a strong collinearity with PC5 that disrupted the model convergence) and weeks on semaglutide. The model included both

random intercepts and slopes for weeks on semaglutide. Bonferroni correction was applied to control for multiple testing.

For testing differences across ancestries, fixed effects were defined for the medication dose, route, sex, age, initial weight a polynomial function of weeks on semaglutide, and an interaction between a polynomial function of weeks and genetic ancestry categorical class (EUR, AFR, EAS, SAS and AMR). The model included both random intercepts and slopes for the polynomial function of weeks on semaglutide. An ANOVA was performed on the model with a Bonferroni correction to assess the global effects of ancestry and ancestry×weeks interaction. When a significant result was found, post hoc comparisons between groups were conducted using the summary lmerTest function (v.3.1.3[168]), applying a Bonferroni correction to all 12 class or class × time interactions.

To test the relationship between PGS and weight loss relying on longitudinal data, scaled BMI (PGS000027) and DM2 PGS (PGS000729) were divided into three equal bins each: low, intermediate and high scores. Related individuals based on their genetic similarity were excluded (defined using PLINK v2.0a[137] with the --king-cutoff 0.05), Then, for each trait, fixed effects were defined for the categorical PGS bins, medication dose, route, sex, age, initial weight, 10 first genetic PCs (excluding PC6 due to a strong collinearity with PC5 that disrupted the model convergence) and a polynomial function of weeks. Random intercepts were defined to account for repeated measures within individuals. Bonferroni correction was applied to control for multiple testing.

To test the relationship between PGS and weight loss relying on a simplified model where the maximum weight loss was considered for each patient, linear regression was applied. The model was adjusted for the medication dose and route, 10 genetic PCs, age, sex, and initial weight. Visualization was made with BPG[125], with smoothed data and 95% CI using loess.as R function (fANOVA package v.0.6.1[170]).

***GWAS***. The analysis was performed to test a relationship between the maximum weight loss on semaglutide and common genetic variants, for each broad-scale ancestry using SAIGE[171]. For step one, array observed variants were used following PLINK v2.0a[127] filtering, with the flags: --maf 0.01 --mind 0.1 --geno 0.1 --hwe 1e-6 . For the second step, array observed and imputed variants were used following plink filtering with --maf 0.01 --geno 0.05 --hwe 1e-6. A quantitative analysis was conducted with the traitType flag. The medication dose and route, the first five genetic PCs, age, sex, and initial weight were used as covariates. METAL[172] was used for meta-analysis.

***Gene-level tests***. To identify genes genetically associated with weight loss, we used Regenie[70] with an additive model for gene-level tests. Only EUR semaglutide users were used, and per each the maximum weight loss record with the corresponding number of weeks was considered. The list of candidate genes was limited to bonferroni-significant proteins whose plasma abundance was altered by semaglutide treatment[110]. We considered variants within these genes with a predicted moderate or high impact on the protein function, according to VEP v.1.2[90]. For step one, the variant list was limited to observed array SNPs following a PLINK v2.0a[137] filtering with: --maf 0.01 --mac 100 --indep-pairwise 1000 100 0.9 --chr 1-22 --snps-only --geno 0.1 --hwe 1e-15. In step 2, we applied PLINK v2.0a filtering to variants across all EUR individuals using the parameters --geno 0.05 and --hwe 1e-6. Subsequently, the file was filtered to include only

semaglutide users and was used in Step 2. The medication dose and route, the first 10 genetic PCs, age, sex, and initial weight were used as covariates. Bonferroni correction was applied to control for multiple testing.

## Competing Interests

P.S is a consultant for 10X Genomics, Illumina, Foresight Diagnostics, Natera, and Twinstrand. P.C.B. sits on the scientific advisory boards of Sage Bionetworks, Intersect Diagnostics and BioSymetric.

All other authors declare no conflict of interest.

## Author Contributions

D.H.G obtained funding for the biobank, and P.C.B, P.S, N.Zaitlen, B.P, A.B, V.A.A and T.S.C provided funding for analysis. D.H.G, P.C.B , P.S, and N.Zaitlen designed, and B.P, A.B, V.A.A, and T.S.C oversaw aspects of the study. R.H, J.F and V.T performed clinical data curation. T.N.Y, M.F.E.M, Y.P, N.Zeltser and M.T performed QC of the whole exome sequencing data. VT, S.A, N.Zeltser, and R.H.-W performed imputation of the array data, processing and QC. R.H, M.M, T.T, and V.P performed curation and annotation of the whole-exome sequencing data. R.H defined broad-scale ancestries, and A.W fine-scale ancestries with support from C.C and R.S. R.H, M.P.M, and J.F performed the demographic investigation. R.H performed analysis of PGS, clinical phenotypes and ancestries, ACMG variant differences, and semaglutide efficacy. M.P.M performed PheWAS and ExWAS analyses, curated computationally predicted deleterious variants, and constructed the web portal. K.Q performed locus pruning. L.O.C assisted in setting up the web server. A.W curated ClinGen P/LP variants and tested differences in frequencies of alleles across ancestries. J.A provided statistical support. R.H, M.P.M, A.W, P.S, and D.H.G wrote and edited the first drafts of the paper, and all others edited subsequent versions. L.E, M.A, C.L, M.E.B, P.T and D.B.M were involved in building and maintaining the biobank, the data and its management. M.E.B, P.Thapaliya, P.Tung and D.B.M provided data curation, security, database and computer infrastructure.

## Data and Materials Availability

Due to privacy regulations, individual-level data cannot be published. Comprehensive processed data and summary statistics are available in the supplementary tables or through our browser (atlas-phewas.mednet.ucla.edu).

## Ethical Statement

The Institutional Review Board of the University of California Los Angeles gave ethical approval for this work. The IRB Number for the ATLAS Initiative protocol is IRB#17-001013.

## Acknowledgments

We are deeply grateful to the participants of the UCLA Health Biobank and their families, without whom this research would not have been possible. We are grateful to the DGSOM and UC Health for their support and development of the ATLAS biobank, and to Allen and Charlotte Ginsburg for

their support of the Institute for Precision Health, which oversees ATLAS. WES data for this study were generated as part of the partnership of UCLA Health with Regeneron Genetics Center (RGC). R.H was supported by EMBO Postdoctoral Fellowship ALTF 1131-2021 and the Prostate Cancer Foundation Young Investigator Award 22YOUN32. M.P.M was supported by the UCLA-Caltech Medical Scientist Training Program (T32-GM008042) and by the National Institute of Mental Health (F30-MH135712). A.W was supported by the National Human Genome Research Institute (F31-HG013462). J.F was supported by the National Institute of Biomedical Imaging and Bioengineering Medical Imaging Informatics Training Grant (T32-EB016640). UCLA CTSI supported D.H.G, A.B, and C.L. T.S.C was supported by NIH grants K08AG065519-01A1, R01AG085518-01A1, U54NS123746 and California Department of Public Health, Chronic Disease Control Branch, Alzheimer's Disease Program, under Contract #22-10079 and #24-10127. N.Zeltser was supported by the National Human Genome Research Institute (T32-HG002536) and by the National Cancer Institute (F31-CA281168). V.A.A. and N.Zaitlen are supported by National Human Genome Research Insitute-R01HG011345.

## References


1. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022).

2. Chambers, D. A., Feero, W. G. & Khoury, M. J. Convergence of Implementation Science, Precision Medicine, and the Learning Health Care System: A New Model for Biomedical Research. *JAMA* **315**, 1941 (2016).

3. The All of Us Research Program Genomics Investigators *et al.* Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).

4. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).

5. Feng, Y.-C. A. *et al.* Taiwan Biobank: A rich biomedical research database of the Taiwanese population. *Cell Genomics* **2**, 100197 (2022).

6. Johnson, J. S. *et al.* Mapping anorexia nervosa genes to clinical phenotypes. *Psychol. Med.* **53**, 2619–2633 (2023).

7. Roden, D. *et al.* Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).

8. Carey, D. J. *et al.* The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. *Genet. Med.* **18**, 906–913 (2016).

9. Zawistowski, M. *et al.* The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. *Cell Genomics* **3**, 100257 (2023).

10. the eMERGE Team *et al.* The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* **4**, 13 (2011).

11. COVID-19 Host Genetics Initiative. A second update on mapping the human genetic architecture of COVID-19. *Nature* **621**, E7–E26 (2023).

12. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022).

13. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).

14. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

15. Kullo, I. J. *et al.* The PRIMED Consortium: Reducing disparities in polygenic risk assessment. *Am. J. Hum. Genet.* **111**, 2594–2606 (2024).

16. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).

17. Tsuo, K. *et al.* All of Us diversity and scale improve polygenic prediction contextually with greatest improvements for under-represented populations. *BioRxiv Prepr. Serv. Biol.* 2024.08.06.606846 (2024) doi:10.1101/2024.08.06.606846.

18. Jurgens, S. J. *et al.* Rare coding variant analysis for human diseases across biobanks and ancestries. *Nat. Genet.* **56**, 1811–1820 (2024).

19. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).

20. Venner, E. *et al.* The frequency of pathogenic variation in the All of Us cohort reveals ancestry-driven disparities. *Commun. Biol.* **7**, 174 (2024).

21. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).

22. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

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

24. Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).

25. Lajonchere, C. *et al.* An Integrated, Scalable, Electronic Video Consent Process to Power Precision Health Research: Large, Population-Based, Cohort Implementation and Scalability Study. *J. Med. Internet Res.* **23**, e31121 (2021).

26. Johnson, R. *et al.* Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. *Genome Med.* **14**, 104 (2022).

27. Johnson, R. *et al.* The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank. *Cell Genomics* **3**, 100243 (2023).

28. Feigin, V. L. *et al.* The global burden of neurological disorders: translating evidence into policy. *Lancet Neurol.* **19**, 255–265 (2020).

29. Murray, C. J. L. Findings from the Global Burden of Disease Study 2021. *The Lancet* **403**, 2259–2262 (2024).

30. Ferrari, A. J. *et al.* Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet* **403**, 2133–2161 (2024).

31. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

32. Sirugo, G., Tishkoff, S. A. & Williams, S. M. The quagmire of race, genetic ancestry, and health disparities. *J. Clin. Invest.* **131**, e150255 (2021).

33. Adigbli, G. Race, science and (im)precision medicine. *Nat. Med.* **26**, 1675–1676 (2020).

34. Krainc, T. & Fuentes, A. Genetic ancestry in precision medicine is reshaping the race debate. *Proc. Natl. Acad. Sci.* **119**, e2203033119 (2022).

35. Borrell, L. N. *et al.* Race and Genetic Ancestry in Medicine — A Time for Reckoning with Racism. *N. Engl. J. Med.* **384**, 474–480 (2021).

36. Quan, H. *et al.* Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am. J. Epidemiol.* **173**, 676–682 (2011).

37. Quan, H. *et al.* Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* **43**, 1130–1139 (2005).

38. American Cancer Society. Cancer Facts & Figures. (2024).

39. Gillott, R. G., Willan, K., Kain, K., Sivananthan, U. M. & Tayebjee, M. H. South Asian ethnicity is associated with a lower prevalence of atrial fibrillation despite greater prevalence of established risk factors: a population-based study in Bradford Metropolitan District. *Europace* euw010 (2016) doi:10.1093/europace/euw010.

40. Agbonlahor, O. *et al.* Racial/Ethnic Discrimination and Cardiometabolic Diseases: A Systematic Review. *J. Racial Ethn. Health Disparities* **11**, 783–807 (2024).

41. Akinyemi, R. O. *et al.* Dementia in Africa: Current evidence, knowledge gaps, and future directions. *Alzheimers Dement.* **18**, 790–809 (2022).

42. Cohen, F. *et al.* Prevalence and burden of migraine in the United States: A systematic review. *Headache J. Head Face Pain* **64**, 516–532 (2024).

43. Siddiqi, S. *et al.* Race and ethnicity matter! Moving Parkinson's risk research towards diversity and inclusiveness. *Npj Park. Dis.* **11**, 45 (2025).

44. Gutiérrez-Rojas, L., Porras-Segovia, A., Dunne, H., Andrade-González, N. & Cervilla, J. A. Prevalence and correlates of major depressive disorder: a systematic review. *Rev. Bras. Psiquiatr. Sao Paulo Braz. 1999* **42**, 657–672 (2020).

45. Baxter, A. J., Scott, K. M., Vos, T. & Whiteford, H. A. Global prevalence of anxiety disorders: a systematic review and meta-regression. *Psychol. Med.* **43**, 897–910 (2013).

46. Banerjee, P. N., Filippi, D. & Allen Hauser, W. The descriptive epidemiology of epilepsy—A review. *Epilepsy Res.* **85**, 31–45 (2009).

47. Adamu, A., Chen, R., Li, A. & Xue, G. Epilepsy in Asian countries. *Acta Epileptol.* **5**, 25 (2023).

48. Minsky, S., Vega, W., Miskimen, T., Gara, M. & Escobar, J. Diagnostic Patterns in Latino, African American, and European American Psychiatric Patients. *Arch. Gen. Psychiatry* **60**, 637 (2003).

49. Hwang, S. H. J. *et al.* Higher prevalence of bipolar I disorder among Asian and Latino compared to Caucasian patients receiving treatment. *Asia-Pac. Psychiatry* **2**, 156–165 (2010).

50. Mirrakhimov, A. E., Sooronbaev, T. & Mirrakhimov, E. M. Prevalence of obstructive sleep apnea in Asian adults: a systematic review of the literature. *BMC Pulm. Med.* **13**, 10 (2013).

51. Villaneuva, A. T. C., Buchanan, P. R., Yee, B. J. & Grunstein, R. R. Ethnicity and obstructive sleep apnoea. *Sleep Med. Rev.* **9**, 419–436 (2005).

52. Redline, S. *et al.* Racial differences in sleep-disordered breathing in African-Americans and Caucasians. *Am. J. Respir. Crit. Care Med.* **155**, 186–192 (1997).

53. Ong, K. C. & Clerk, A. A. Comparison of the severity of sleep-disordered breathing in Asian and Caucasian patients seen at a sleep disorders center. *Respir. Med.* **92**, 843–848 (1998).

54. Abdelmoumen, I. *et al.* Boricua Founder Variant in FRRS1L Causes Epileptic Encephalopathy With Hyperkinetic Movements. *J. Child Neurol.* **36**, 93–98 (2021).

55. Belbin, G. M. *et al.* Leveraging health systems data to characterize a large effect variant conferring risk for liver disease in Puerto Ricans. *Am. J. Hum. Genet.* **108**, 2099–2111 (2021).

56. Caggiano, C. *et al.* Disease risk and healthcare utilization among ancestrally diverse groups in the Los Angeles region. *Nat. Med.* **29**, 1845–1856 (2023).

57. Belbin, G. M. *et al.* Toward a fine-scale population health monitoring system. *Cell* **184**, 2068-2083.e11 (2021).

58. Dai, C. L. *et al.* Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis. *Am. J. Hum. Genet.* **106**, 371–388 (2020).

59. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PloS One* **12**, e0175508 (2017).

60. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).

61. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).

62. Prasad, P. & Krishnan, E. Filipino Gout: A Review. *Arthritis Care Res.* **66**, 337–343 (2014).

63. Feliciano-Astacio, B. E. *et al.* The Puerto Rico Alzheimer Disease Initiative (PRADI): A Multisource Ascertainment Approach. *Front. Genet.* **10**, 538 (2019).

64. Umemura, S. *et al.* The Japanese Society of Hypertension Guidelines for the Management of Hypertension (JSH 2019). *Hypertens. Res.* **42**, 1235–1481 (2019).

65. Zhang, N., Wang, J., Li, Y. & Jiang, B. Prevalence of primary open angle glaucoma in the last 20 years: a meta-analysis and systematic review. *Sci. Rep.* **11**, 13762 (2021).

66. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).

67. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

68. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

69. Curtis, D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* **28**, 85–89 (2018).

70. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

71. Kavanagh, P. L., Fasipe, T. A. & Wun, T. Sickle Cell Disease: A Review. *JAMA* **328**, 57–68 (2022).

72. Hsu, L.-A., Wu, S., Teng, M.-S. & Ko, Y.-L. Causal links of α-thalassemia indices and cardiometabolic traits and diabetes: MR study. *Life Sci. Alliance* **6**, e202302204 (2023).

73. Romeo, S. *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.* **40**, 1461–1465 (2008).

74. Yang, L. *et al.* Haploinsufficiency for Stard7 is associated with enhanced allergic responses in lung and skin. *J. Immunol. Baltim. Md 1950* **194**, 5635–5643 (2015).

75. Peng, D.-F., Hu, T.-L., Soutto, M., Belkhiri, A. & El-Rifai, W. Loss of glutathione peroxidase 7 promotes TNF-α-induced NF-κB activation in Barrett's carcinogenesis. *Carcinogenesis* **35**, 1620–1628 (2014).

76. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

77. Viswanathan, L. & Rao, S. S. Intestinal Disaccharidase Deficiency in Adults: Evaluation and Treatment. *Curr. Gastroenterol. Rep.* **25**, 134–139 (2023).

78. Garg, A. *et al.* The molecular basis of Human FN3K mediated phosphorylation of glycated substrates. *Nat. Commun.* **16**, 941 (2025).

79. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

80. Gomez, F., Hirbo, J. & Tishkoff, S. A. Genetic Variation and Adaptation in Africa: Implications for Human Evolution and Disease. *Cold Spring Harb. Perspect. Biol.* **6**, a008524–a008524 (2014).

81. Apostolou, P. & Fostira, F. Hereditary Breast Cancer: The Era of New Susceptibility Genes. *BioMed Res. Int.* **2013**, 1–11 (2013).

82. Naccashian, Z., Hattar-Pollara, M., Ho, C. A. & Ayvazian, S. P. Prevalence and Predictors of Diabetes Mellitus and Hypertension in Armenian Americans in Los Angeles. *Diabetes Educ.* **44**, 130–143 (2018).

83. Amaryan, G., Sarkisian, T., Tadevosyan, A. & Braegger, C. Familial Mediterranean fever in Armenian children with inflammatory bowel disease. *Front. Pediatr.* **11**, 1288523 (2023).

84. Piel, F. B., Steinberg, M. H. & Rees, D. C. Sickle Cell Disease. *N. Engl. J. Med.* **376**, 1561–1573 (2017).

85. Goodrich, J. K. *et al.* Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nat. Commun.* **12**, 3505 (2021).

86. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).

87. Tanchoco, C. C., Cruz, A. J., Duante, C. A. & Litonjua, A. D. Prevalence of metabolic syndrome among Filipino adults aged 20 years and over. *Asia Pac. J. Clin. Nutr.* **12**, 271–276 (2003).

88. Miller, D. T. *et al.* ACMG SF v3.2 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. Off. J. Am. Coll. Med. Genet.* **25**, 100866 (2023).

89. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

90. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

91. Rastogi, R. *et al.* Critical assessment of missense variant effect predictors on disease-relevant variant data. *Hum. Genet.* **144**, 281–293 (2025).

92. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).

93. Erdem, M. *et al.* Identification of an mRNA isoform switch for HNRNPA1 in breast cancers. *Sci. Rep.* **11**, 24444 (2021).

94. Cullup, T. *et al.* Recessive mutations in EPG5 cause Vici syndrome, a multisystem disorder with defective autophagy. *Nat. Genet.* **45**, 83–87 (2013).

95. Zhang, S. *et al.* The regulation, function, and role of lipophagy, a form of selective autophagy, in metabolic disorders. *Cell Death Dis.* **13**, 132 (2022).

96. Stein, C. S. *et al.* Osmoregulation of ceroid neuronal lipofuscinosis type 3 in the renal medulla. *Am. J. Physiol. Cell Physiol.* **298**, C1388-1400 (2010).

97. Huang, S. *et al.* PPAR-γ in Macrophages Limits Pulmonary Inflammation and Promotes Host Recovery following Respiratory Viral Infection. *J. Virol.* **93**, e00030-19 (2019).

98. Szot, J. O. *et al.* A metabolic signature for NADSYN1-dependent congenital NAD deficiency disorder. *J. Clin. Invest.* **134**, e174824 (2024).

99. Schuurs-Hoeijmakers, J. H. M. *et al.* Mutations in DDHD2, encoding an intracellular phospholipase A(1), cause a recessive form of complex hereditary spastic paraplegia. *Am. J. Hum. Genet.* **91**, 1073–1081 (2012).

100. Liao, X. *et al.* Effect of mechanical unloading on genome-wide DNA methylation profile of the failing human heart. *JCI Insight* **8**, e161788 (2023).

101. Fujimoto, B. A. *et al.* The exocyst complex regulates insulin-stimulated glucose uptake of skeletal muscle cells. *Am. J. Physiol. Endocrinol. Metab.* **317**, E957–E972 (2019).

102. Deng, H., Gong, X., Ji, G., Li, C. & Cheng, S. KIF2C promotes clear cell renal cell carcinoma progression via activating JAK2/STAT3 signaling pathway. *Mol. Cell. Probes* **72**, 101938 (2023).

103. Lin, F. *et al.* Kidney-specific inactivation of the KIF3A subunit of kinesin-II inhibits renal ciliogenesis and produces polycystic kidney disease. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5286–5291 (2003).

104. den Heijer, J. M. *et al.* A Large-Scale Full GBA1 Gene Screening in Parkinson's Disease in the Netherlands. *Mov. Disord. Off. J. Mov. Disord. Soc.* **35**, 1667–1674 (2020).

105. Ibrahim, S. *et al.* LDLR variant classification for improved cardiovascular risk prediction in familial hypercholesterolemia. *Atherosclerosis* **397**, 117610 (2024).

106. Wilding, J. P. H. *et al.* Once-Weekly Semaglutide in Adults with Overweight or Obesity. *N. Engl. J. Med.* **384**, 989–1002 (2021).

107. Ryan, D. H. *et al.* Long-term weight loss effects of semaglutide in obesity without diabetes in the SELECT trial. *Nat. Med.* **30**, 2049–2057 (2024).

108. Garvey, W. T. *et al.* Two-year effects of semaglutide in adults with overweight or obesity: the STEP 5 trial. *Nat. Med.* **28**, 2083–2091 (2022).

109. German, J. *et al.* Association between plausible genetic factors and weight loss from GLP1-RA and bariatric surgery. *Nat. Med.* (2025) doi:10.1038/s41591-025-03645-3.

110. Maretty, L. *et al.* Proteomic changes upon treatment with semaglutide in individuals with obesity. *Nat. Med.* (2025) doi:10.1038/s41591-024-03355-2.

111. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

112. Kannel, W. B. Diabetes and Cardiovascular Disease: The Framingham Study. *JAMA* **241**, 2035 (1979).

113. Tsao, C. W. & Vasan, R. S. The Framingham Heart Study: past, present and future. *Int. J. Epidemiol.* **44**, 1763–1766 (2015).

114. Wiley, L. K. *et al.* Building a vertically integrated genomic learning health system: The biobank at the Colorado Center for Personalized Medicine. *Am. J. Hum. Genet.* **111**, 11–23 (2024).

115. German, J. *et al.* Association between plausible genetic factors and weight loss from GLP1-RA and bariatric surgery: a multi-ancestry study in 10 960 individuals from 9 biobanks. Preprint at https://doi.org/10.1101/2024.09.11.24313458 (2024).

116. Palma-Martínez, M. J. *et al.* The multi-scale complexity of human genetic variation beyond continental groups. Preprint at https://doi.org/10.1101/2024.12.11.627824 (2024).

117. Smith, N. *et al.* Health plan administrative records versus birth certificate records: quality of race and ethnicity information in children. *BMC Health Serv. Res.* **10**, 316 (2010).

118. Larena, M. *et al.* Multiple migrations to the Philippines during the last 50,000 years. *Proc. Natl. Acad. Sci.* **118**, e2026132118 (2021).

119. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679 (1999).

120. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinforma. Oxf. Engl.* **27**, 2156–2158 (2011).

121. O'Malley, K. J. *et al.* Measuring diagnoses: ICD code accuracy. *Health Serv. Res.* **40**, 1620–1639 (2005).

122. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

123. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).

124. Gasparini, A. comorbidity: An R package for computing comorbidity scores. *J. Open Source Softw.* **3**, 648 (2018).

125. P'ng, C. *et al.* BPG: Seamless, automated and interactive visualization of scientific data. *BMC Bioinformatics* **20**, 42 (2019).

126. Lambert, S. A. *et al.* Enhancing the Polygenic Score Catalog with tools for score calculation and ancestry normalization. *Nat. Genet.* **56**, 1989–1994 (2024).

127. Lambert, S. A. *et al.* The Polygenic Score Catalog: new functionality and tools to enable FAIR research. Preprint at https://doi.org/10.1101/2024.05.29.24307783 (2024).

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

129. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

130. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).

131. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

132. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

133. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat. Genet.* **55**, 1243–1249 (2023).

134. Shemirani, R. *et al.* Rapid detection of identity-by-descent tracts for mega-scale datasets. *Nat. Commun.* **12**, 3546 (2021).

135. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

136. Staudt, C. L., Sazonovs, A. & Meyerhenke, H. NetworKit: A tool suite for large-scale complex network analysis. *Netw. Sci.* **4**, 508–530 (2016).

137. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

138. Caggiano, C. *et al.* Disease risk and healthcare utilization among ancestrally diverse groups in the Los Angeles region. *Nat. Med.* **29**, 1845–1856 (2023).

139. Puhr, R., Heinze, G., Nold, M., Lusa, L. & Geroldinger, A. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat. Med.* **36**, 2302–2317 (2017).

140. Krasheninina, O. *et al.* Open-source mapping and variant calling for large-scale NGS data from original base-quality scores. Preprint at https://doi.org/10.1101/2020.12.15.356360 (2020).

141. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

142. Picard toolkit. Broad Institute (2019).

143. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

144. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinforma. Oxf. Engl.* **36**, 5582–5589 (2021).

145. Sun, K. Y. *et al.* A deep catalogue of protein-coding variation in 983,578 individuals. *Nature* **631**, 583–592 (2024).

146. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).

147. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinforma. Oxf. Engl.* **32**, 3047–3048 (2016).

148. Cleary, J. G. *et al.* Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. Preprint at https://doi.org/10.1101/023754 (2015).

149. Patel, Y. *et al.* Metapipeline-DNA: A Comprehensive Germline & Somatic Genomics Nextflow Pipeline. Preprint at https://doi.org/10.1101/2024.09.04.611267 (2024).

150. Pinto, B. J., O'Connor, B., Schatz, M. C., Zarate, S. & Wilson, M. A. Concerning the eXclusion in human genomics: the choice of sex chromosome representation in the human genome drastically affects the number of identified variants. *G3 Bethesda Md* **13**, jkad169 (2023).

151. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).

152. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).

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

154. Jurgens, S. J. *et al.* Rare coding variant analysis for human diseases across biobanks and ancestries. *Nat. Genet.* **56**, 1811–1820 (2024).

155. Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J. & Hocking, T. D. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *Am. J. Hum. Genet.* **103**, 474–483 (2018).

156. Li, C., Zhi, D., Wang, K. & Liu, X. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. *Genome Med.* **14**, 115 (2022).

157. Feng, B.-J. PERCH: A Unified Framework for Disease Gene Prioritization. *Hum. Mutat.* **38**, 243–251 (2017).

158. Wu, Y., Li, R., Sun, S., Weile, J. & Roth, F. P. Improved pathogenicity prediction for rare human missense variants. *Am. J. Hum. Genet.* **108**, 1891–1906 (2021).

159. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

160. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

161. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* **11**, 5918 (2020).

162. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).

163. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2016239118 (2021).

164. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).

165. Olson, N. D. *et al.* PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genomics* **2**, 100129, S2666-979X(22)00058–1 (2022).

166. Yao, F., Müller, H.-G. & Wang, J.-L. Functional Data Analysis for Sparse Longitudinal Data. *J. Am. Stat. Assoc.* **100**, 577–590 (2005).

167. Liu, B. & Müller, H.-G. Estimating Derivatives for Samples of Sparsely Observed Functions, With Application to Online Auction Dynamics. *J. Am. Stat. Assoc.* **104**, 704–717 (2009).

168. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. **lmerTest** Package: Tests in Linear Mixed Effects Models. *J. Stat. Softw.* **82**, (2017).

169. Breheny, P. & Burchett, W. Visualization of Regression Models Using visreg. *R J.* **9**, 56 (2017).

170. Fruth, J., Roustant, O. & Kuhnt, S. Total interaction index: A variance-based sensitivity index for second-order interaction screening. *J. Stat. Plan. Inference* **147**, 212–223 (2014).

171. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

172. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

**Table 1. Baseline demographic information on the UCLA ATLAS population**

| | | UCLA ATLAS | | |
|---|---|---|---|---|
| **Variable** | **Category** | **Overall** | **Alive** | **Deceased** |
| **n** | | 92512 | 88555 | 3957 |
| **Age, mean (SD)** | | 54.0 (17.3) | 53.4 (17.2) | 67.4 (15.2) |
| **Self-reported sex, n (%)** | Female | 51545 (55.7) | 49834 (56.3) | 1711 (43.2) |
| | Male | 40930 (44.2) | 38684 (43.7) | 2246 (56.8) |
| | Unspecified, X | 37 (0.0) | 37 (0.0) | |
| **Self-reported race, n (%)** | American Indian, Alaska Native | 823 (0.9) | 800 (0.9) | 23 (0.6) |
| | Asian | 11105 (12.0) | 10733 (12.1) | 372 (9.4) |
| | Black, African American | 4246 (4.6) | 4019 (4.5) | 227 (5.7) |
| | Caribbean/West Indian | 165 (0.2) | 165 (0.2) | |
| | Middle Eastern or North African | 2364 (2.6) | 2315 (2.6) | 49 (1.2) |
| | Native Hawaiian, Guamian or Chamorro, Samoan, Other Pacific Islander | 269 (0.3) | 261 (0.3) | 8 (0.2) |
| | Other Race | 3437 (3.7) | 2978 (3.4) | 459 (11.6) |
| | Unknown, declined to specify | 12111 (13.1) | 11824 (13.4) | 287 (7.3) |
| | White | 57992 (62.7) | 55460 (62.6) | 2532 (64.0) |
| **Self-reported ethnicity, n (%)** | Hispanic/Latino, Cuban, Hispanic/Spanish origin, Mexican, Mexican American, Chicano/a, Puerto Rican | 13706 (14.8) | 12956 (14.6) | 750 (19.0) |

| | | | | |
|---|---|---|---|---|
| | Non-Hispanic/Latino | 71538 (77.3) | 68493 (77.3) | 3045 (77.0) |
| | Unknown, declined to specify | 7268 (7.9) | 7106 (8.0) | 162 (4.1) |
| **Self-reported smoking, n (%)** | Former | 24621 (26.6) | 23025 (26.0) | 1596 (40.3) |
| | Never | 63516 (68.7) | 61339 (69.3) | 2177 (55.0) |
| | Passive Smoke Exposure - Never Smoker | 119 (0.1) | 112 (0.1) | 7 (0.2) |
| | Smoker | 3589 (3.9) | 3461 (3.9) | 128 (3.2) |
| | Unknown, declined to specify | 667 (0.7) | 618 (0.7) | 49 (1.2) |
| **Types of Encounters, n (%)** | Inpatient and Outpatient Encounters | 29637 (32.1) | 26394 (29.8) | 3243 (82.0) |
| | Only Outpatient Encounters | 62828 (67.9) | 62115 (70.2) | 713 (18.0) |
| **Total Encounters, mean (SD)** | | 115.9 (145.8) | 109.1 (136.9) | 266.8 (231.2) |
| Total Inpatient Encounters, mean (SD) | | 0.8 (2.2) | 0.7 (1.9) | 3.8 (5.1) |
| **Total Inpatient Days, mean (SD)** | | 16.4 (33.7) | 13.6 (29.0) | 38.6 (54.4) |

**Table 2. Summary of WES variants**

| | Autosomal variants count | | Median autosomal variants per participant (IQR) | |
|---|---|---|---|---|
| | Variants | Variants with MAF < 1% | Variants | Variants with MAF < 1% |
| Total | 11,571,436 | 11,373,307 | 74,476 (1,631) | 1,925 (654) |
| Targeted regions | 5,445,278 | 5,374,412 | 26,297 (491) | 844 (259) |
| **Variant type** | | | | |
| SNVs | 5,221,310 | 5,175,173 | 24,867 (471) | 799 (248) |
| Indels | 223,956 | 199,364 | 1,401 (45) | 40 (14) |
| Multi-allelic | 447,191 | 439,265 | 4,061 (107) | 136 (42) |
| **Predicted function** | | | | |
| Synonymous | 1,531,497 | 1,510,732 | 10,693 (209) | 266 (100) |
| Missense | 2,893,307 | 2,873,731 | 9,821 (220) | 403 (110) |
| LOF | 173,235 | 172,462 | 159 (15) | 15 (6) |

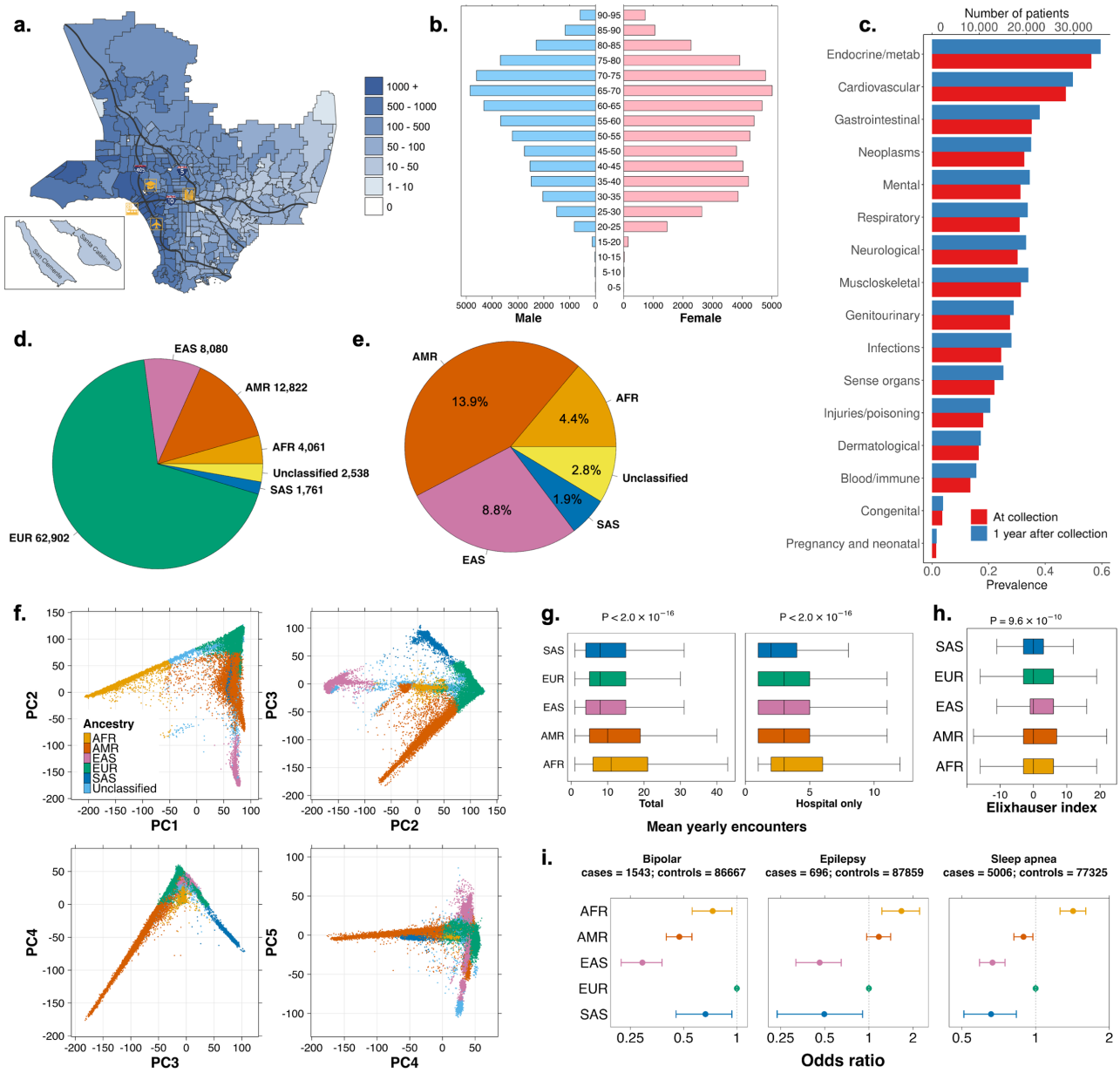


**Figure 1. Overview of the UCLA ATLAS biobank. a.** Choropleth map of ATLAS participants within Los Angeles County, with major highways and landmarks labelled **b.** Distribution of ATLAS participants by age and sex **c.** Prevalence of phecode groups in the UCLA ATLAS population at the time of collection and within one year of the ATLAS launch date. **d.** Genetic ancestry sample sizes. **e.** Genetic ancestry fractions of non-EUR populations. Although EUR are not presented, percentages were calculated using all ATLAS populations, including EUR. **f.** Genetic PCs of ATLAS individuals. Colors are according to the ancestry predictions as inferred from 1000 Genomes. **g.** Mean yearly encounters vary across genetic ancestries. **h.** Comorbidity index across genetic ancestries. **i.** New associations between clinical phenotypes and broad-scale ancestries.

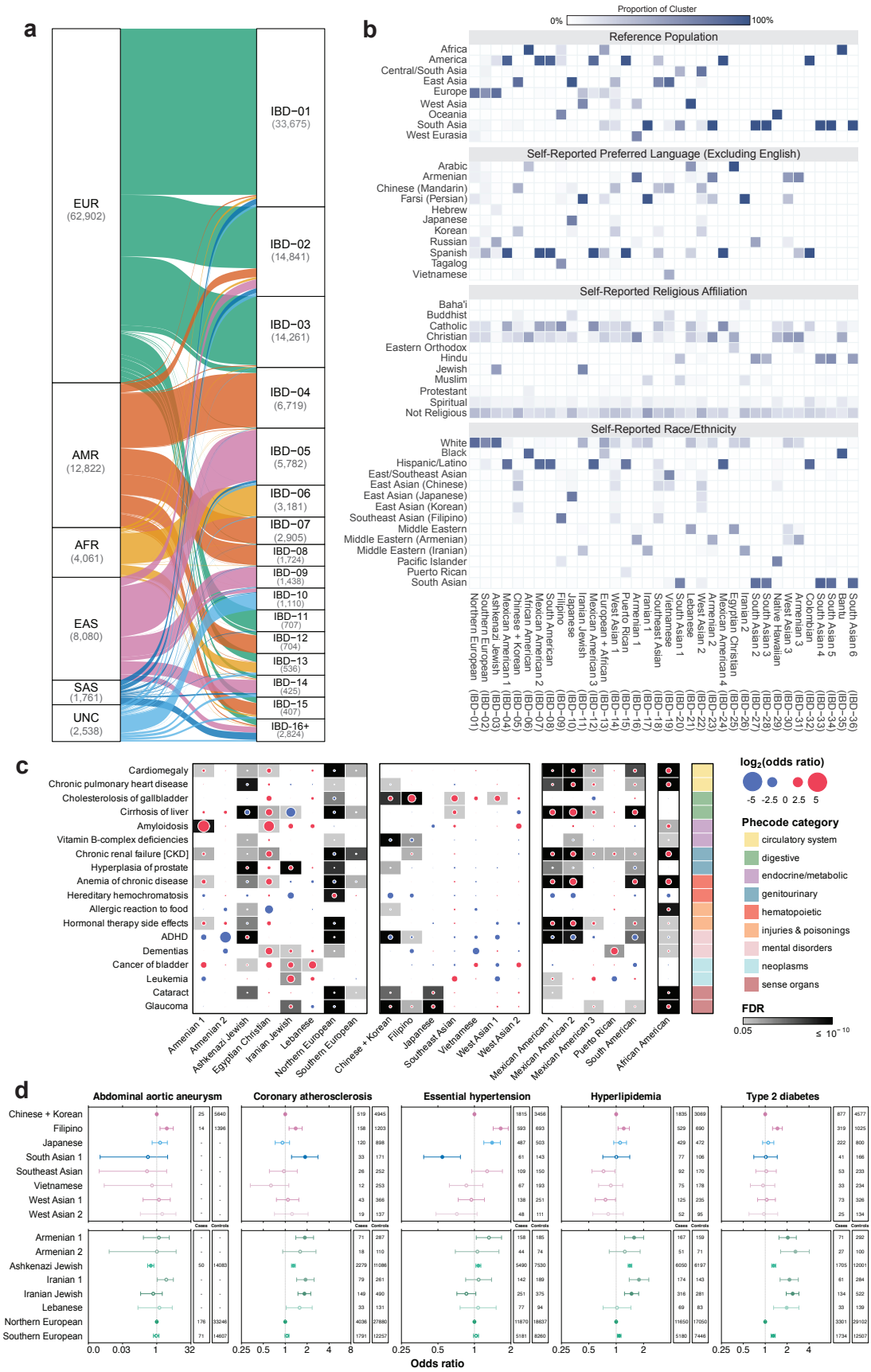
a
EUR (62,902)
AMR (12,822)
AFR (4,061)
EAS (8,080)
SAS (1,761)
UNC (2,538)
IBD−01 (33,675)
IBD−02 (14,841)
IBD−03 (14,261)
IBD−04 (6,719)
IBD−05 (5,782)
IBD−06 (3,181)
IBD−07 (2,905)
IBD−08 (1,724)
IBD−09 (1,438)
IBD−10 (1,110)
IBD−11 (707)
IBD−12 (704)
IBD−13 (536)
IBD−14 (425)
IBD−15 (407)
IBD−16+ (2,824)
b
Proportion of Cluster
0%
100%
Reference Population
Africa
America
Central/South Asia
East Asia
Europe
West Asia
Oceania
South Asia
West Eurasia
Self-Reported Preferred Language (Excluding English)
Arabic
Armenian
Chinese (Mandarin)
Farsi (Persian)
Hebrew
Japanese
Korean
Russian
Spanish
Tagalog
Vietnamese
Self-Reported Religious Affiliation
Baha'i
Buddhist
Catholic
Christian
Eastern Orthodox
Hindu
Jewish
Muslim
Protestant
Spiritual
Not Religious
Self-Reported Race/Ethnicity
White
Black
Hispanic/Latino
East/Southeast Asian
East Asian (Chinese)
East Asian (Japanese)
East Asian (Korean)
Southeast Asian (Filipino)
Middle Eastern
Middle Eastern (Armenian)
Middle Eastern (Iranian)
Pacific Islander
Puerto Rican
South Asian
Northern European (IBD−01)
Southern European (IBD−02)
Ashkenazi Jewish (IBD−03)
Mexican American 1 (IBD−04)
Chinese + Korean (IBD−05)
African American (IBD−06)
Mexican American 2 (IBD−07)
South American (IBD−08)
Filipino (IBD−09)
Japanese (IBD−10)
Iranian Jewish (IBD−11)
Mexican American 3 (IBD−12)
European + African (IBD−13)
West Asian 1 (IBD−14)
Puerto Rican (IBD−15)
Armenian 1 (IBD−16)
Iranian 1 (IBD−17)
Southeast Asian (IBD−18)
Vietnamese (IBD−19)
South Asian 1 (IBD−20)
Lebanese (IBD−21)
West Asian 2 (IBD−22)
Armenian 2 (IBD−23)
Mexican American 4 (IBD−24)
Egyptian Christian (IBD−25)
Iranian 2 (IBD−26)
South Asian 2 (IBD−27)
South Asian 3 (IBD−28)
Native Hawaiian (IBD−29)
Vietnamese 3 (IBD−30)
Armenian 3 (IBD−31)
Colombian (IBD−32)
South Asian 4 (IBD−33)
Bantu (IBD−34)
South Asian 5 (IBD−35)
South Asian 6 (IBD−36)
c
Cardiomegaly
Chronic pulmonary heart disease
Cholesterolosis of gallbladder
Cirrhosis of liver
Amyloidosis
Vitamin B-complex deficiencies
Chronic renal failure [CKD]
Hyperplasia of prostate
Anemia of chronic disease
Hereditary hemochromatosis
Allergic reaction to food
Hormonal therapy side effects
ADHD
Dementias
Cancer of bladder
Leukemia
Cataract
Glaucoma
Armenian 1
Armenian 2
Ashkenazi Jewish
Egyptian Christian
Iranian Jewish
Lebanese
Northern European
Southern European
Chinese + Korean
Filipino
Japanese
Southeast Asian
Vietnamese
West Asian 1
West Asian 2
Mexican American 1
Mexican American 2
Mexican American 3
Puerto Rican
South American
African American
log2(odds ratio)
-5 -2.5 0 2.5 5
Phecode category
circulatory system
digestive
endocrine/metabolic
genitourinary
hematopoietic
injuries & poisonings
mental disorders
neoplasms
sense organs
FDR
0.05
≤ 10−10
d
Abdominal aortic aneurysm
Coronary atherosclerosis
Essential hypertension
Hyperlipidemia
Type 2 diabetes
Chinese + Korean
Filipino
Japanese
South Asian 1
Southeast Asian
Vietnamese
West Asian 1
West Asian 2
Armenian 1
Armenian 2
Ashkenazi Jewish
Iranian 1
Iranian Jewish
Lebanese
Northern European
Southern European
Cases
Controls
0.0 0.03 1 32
0.25 0.5 1 2 4 8
0.25 0.5 1 2
0.25 0.5 1 2
0.25 0.5 1 2 4
Odds ratio

**Figure 2. Introduction to IBD mapping and fine-scale ancestry disease risk. a.** The distribution of genetic ancestry between fine- and broad-scale populations**. b.** Enrichment of reference populations, self-reported language, religion and race across fine-scale ancestries. **c.** The risk of phenotype-defined phecodes across fine-scale genetic ancestries. Each fine-scale population was tested against all ATLAS participants outside of this population. Only populations with at least 100 participants were tested, and phecodes with at least 100 cases across ATLAS, resulting in 23 tested populations and 1,253 phecodes. FDR was used to control for multiple testing. **d**. Cardiometabolic disease risk for each fine-scale group within the same broad-scale continental ancestry. Representative cardiometabolic phecodes were selected, and only populations with at least 100 participants were tested. In cases of small sample sizes, '–' was used instead of numeric values to protect patient privacy. Filled points represent significant results (FDR $\leq$ 0.05).

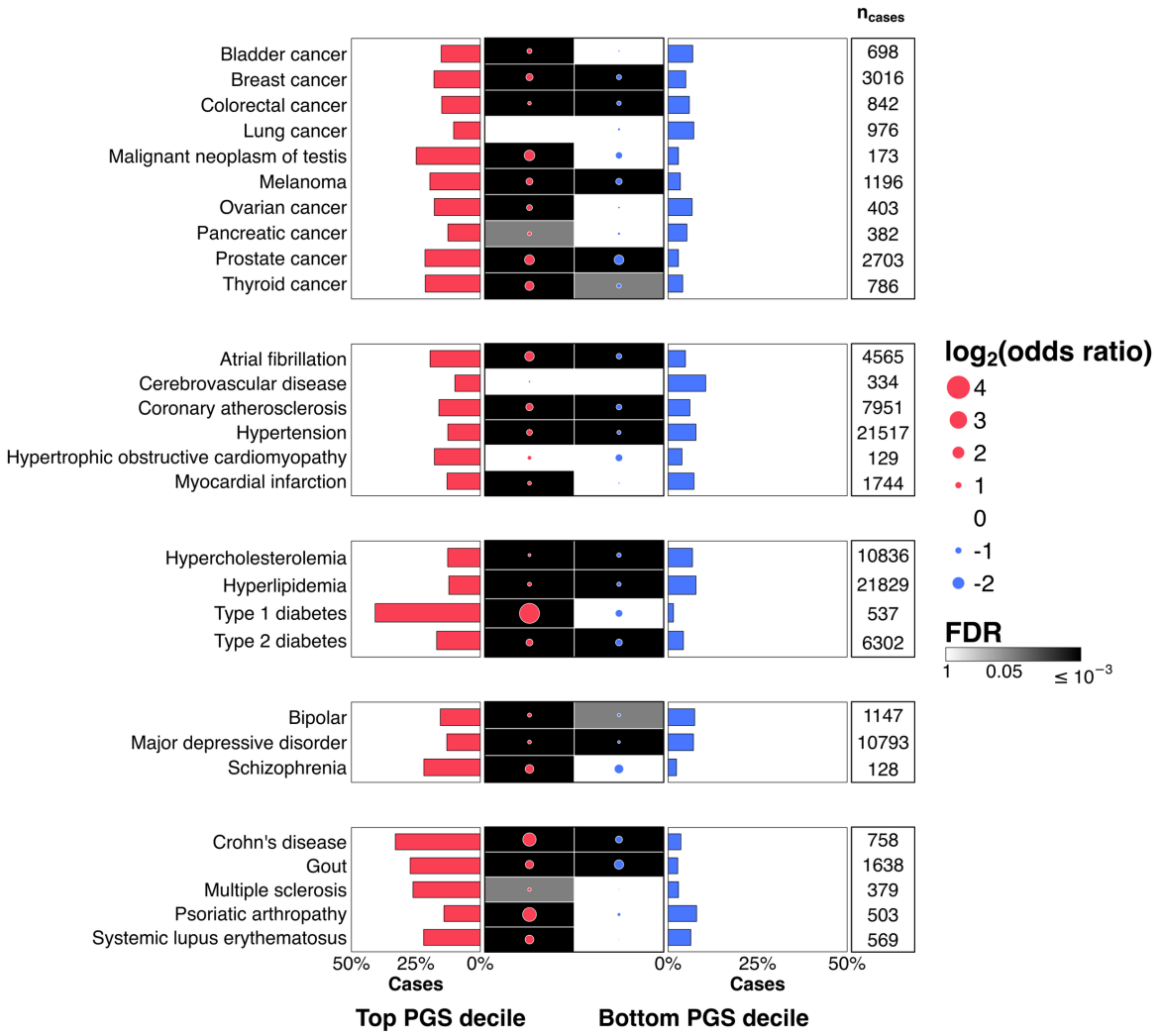


**Figure 3. Polygenic risk and disease diagnoses.** The impact of the top and bottom PGS on disease risk is evaluated through odds ratios (dotmps). The top and bottom PRS deciles were compared to the 5th declle to calculate the OR using a logistic regression model (Methods). The numbers of diagnosed patients assigned to the top or bottom PGS bins are shown in barplots**.** Red represents the top PGS decile, and blue the bottom PGS decile. Diseases are divided into categories, from top to bottom: cancer, cardiovascular, metabolic, neuro-psychiatric and immunne/autoimmune. Only non-related EUR individuals were included. FDR was used to control for multiple testing.

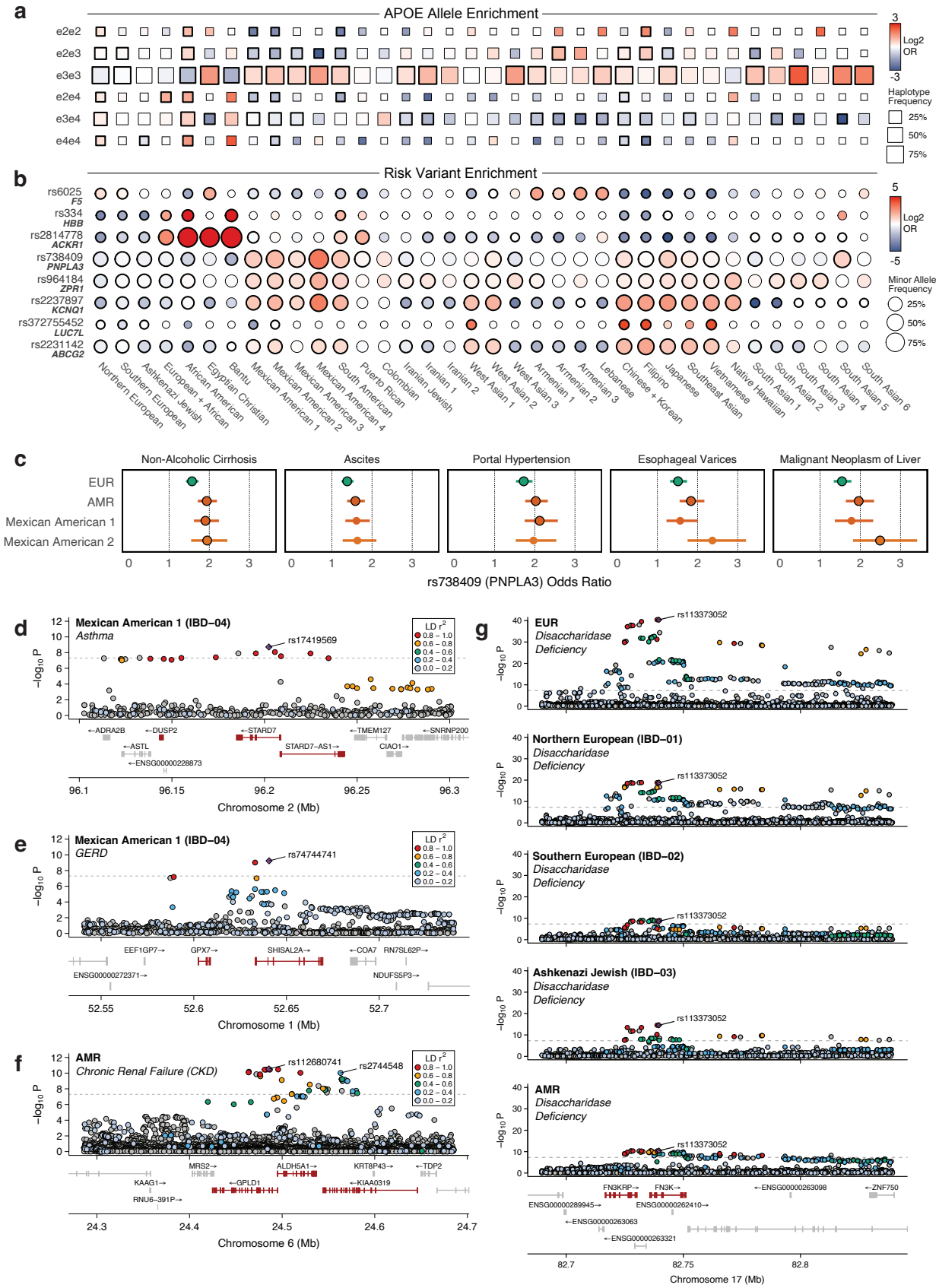
a
APOE Allele Enrichment
e2e2
e2e3
e3e3
e2e4
e3e4
e4e4
Log2 OR
Haplotype Frequency
25%
50%
75%
b
Risk Variant Enrichment
rs6025 F5
rs334 HBB
rs2814778 ACKR1
rs738409 PNPLA3
rs964184 ZPR1
rs2237897 KCNQ1
rs372755452 LUC7L
rs2231142 ABCG2
Minor Allele Frequency
Northern European
Southern European
Ashkenazi Jewish
European + African
African American
Egyptian Christian
Bantu
Mexican American 1
Mexican American 2
Mexican American 3
Mexican American 4
South American
Puerto Rican
Colombian
Iranian Jewish
Iranian 1
Iranian 2
West Asian 1
West Asian 2
West Asian 3
Armenian 1
Armenian 2
Armenian 3
Lebanese
Chinese + Korean
Filipino
Japanese
Southeast Asian
Vietnamese
Native Hawaiian
South Asian 1
South Asian 2
South Asian 3
South Asian 4
South Asian 5
South Asian 6
c
Non-Alcoholic Cirrhosis
Ascites
Portal Hypertension
Esophageal Varices
Malignant Neoplasm of Liver
EUR
AMR
Mexican American 1
Mexican American 2
rs738409 (PNPLA3) Odds Ratio
d
Mexican American 1 (IBD–04)
Asthma
rs17419569
LD r2
Chromosome 2 (Mb)
e
Mexican American 1 (IBD–04)
GERD
rs74744741
Chromosome 1 (Mb)
f
AMR
Chronic Renal Failure (CKD)
rs112680741
rs2744548
Chromosome 6 (Mb)
g
EUR
Disaccharidase Deficiency
rs113373052
Northern European (IBD–01)
Southern European (IBD–02)
Ashkenazi Jewish (IBD–03)
AMR
Chromosome 17 (Mb)

**Figure 4. Common Variation in the ATLAS Biobank. a.** APOE haplotype frequency across fine-scale cohorts. **b.** Allele frequency for known risk variants across fine-scale cohorts; shading indicates the level of over- (red) or under- (blue) enrichment of a haplotype/allele in a given cohort *via* Fisher's exact test. In panels **a-b**, a bold border indicates a significant result. **c-g.** selected PheWAS associations: **c.** non-alcoholic cirrhosis risk variant rs738409, indicating cohort-specific risk across a range of sequelae from liver cirrhosis. **d.** Locus plots for novel associations between rs738409-G and asthma, **e.** rs74744741-C and GERD, **f.** rs112680741-C and chronic renal failure, and **g.** rs7208565-T and intestinal disaccharidase deficiency. Point shading (blue to orange to red) indicates level of ancestry-specific LD with the lead SNP, while gene shading (red) indicates prioritized risk genes.

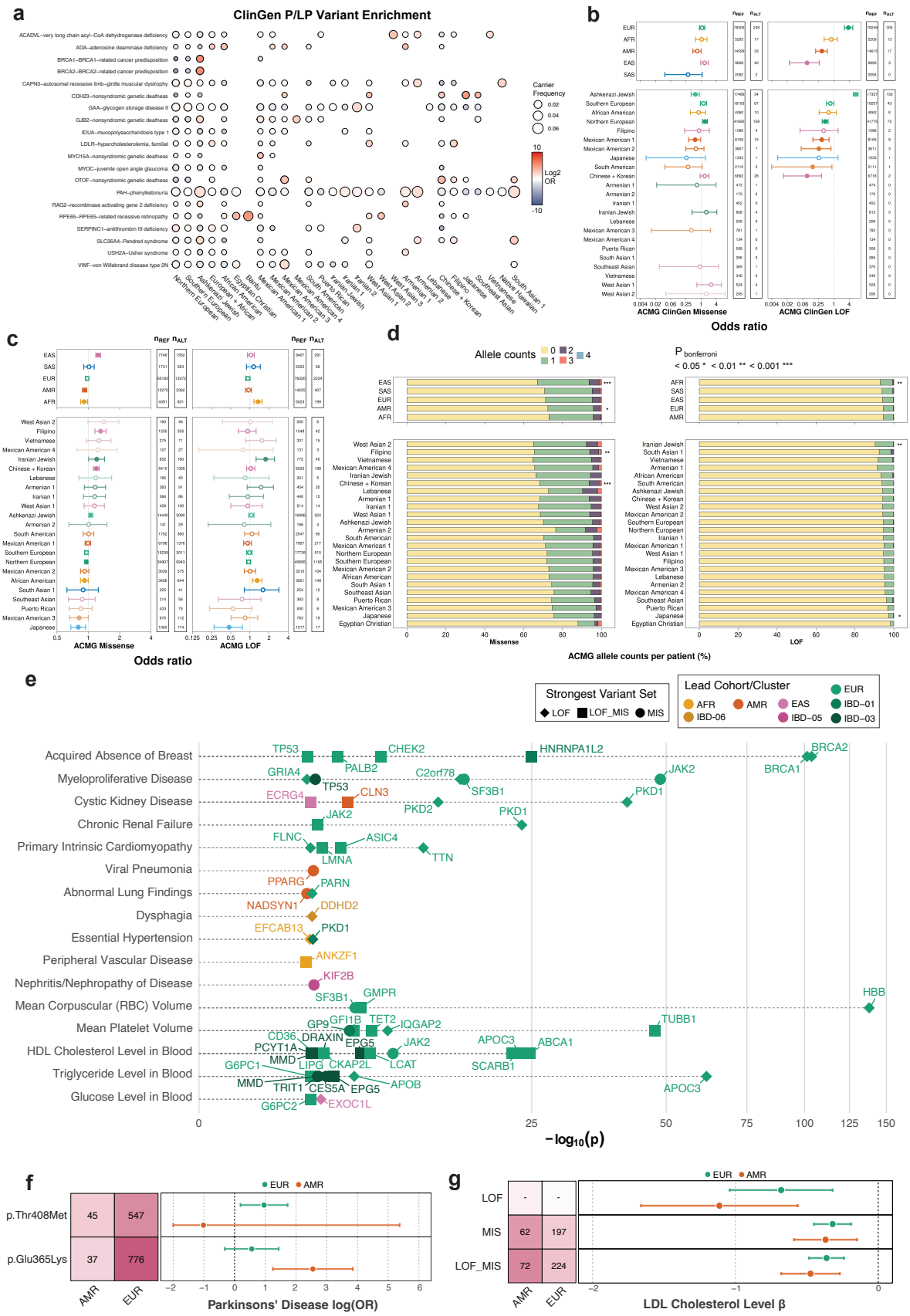
a
ClinGen P/LP Variant Enrichment
Carrier Frequency
Log2 OR
b
ACMG ClinGen Missense
ACMG ClinGen LOF
Odds ratio
c
ACMG Missense
ACMG LOF
Odds ratio
d
Allele counts
P bonferroni
< 0.05 * < 0.01 ** < 0.001 ***
Missense
LOF
ACMG allele counts per patient (%)
e
Strongest Variant Set
LOF
LOF_MIS
MIS
Lead Cohort/Cluster
AFR
AMR
EAS
EUR
IBD-06
IBD-05
IBD-03
Acquired Absence of Breast
Myeloproliferative Disease
Cystic Kidney Disease
Chronic Renal Failure
Primary Intrinsic Cardiomyopathy
Viral Pneumonia
Abnormal Lung Findings
Dysphagia
Essential Hypertension
Peripheral Vascular Disease
Nephritis/Nephropathy of Disease
Mean Corpuscular (RBC) Volume
Mean Platelet Volume
HDL Cholesterol Level in Blood
Triglyceride Level in Blood
Glucose Level in Blood
TP53
CHEK2
HNRNPA1L2
BRCA2
BRCA1
PALB2
GRIA4
C2orf78
JAK2
SF3B1
PKD1
ECRG4
CLN3
PKD2
FLNC
ASIC4
LMNA
TTN
PPARG
PARN
NADSYN1
DDHD2
EFCAB13
ANKZF1
KIF2B
GMPR
HBB
GP9
GFI1B
TET2
IQGAP2
TUBB1
CD36
DRAXIN
EPG5
APOC3
ABCA1
PCYT1A
MMD
G6PC1
LIPG
CKAP2L
LCAT
SCARB1
TRIT1
CES5A
APOB
G6PC2
EXOC1L
−log10(p)
f
EUR
AMR
p.Thr408Met
p.Glu365Lys
Parkinsons' Disease log(OR)
g
LOF
MIS
LOF_MIS
LDL Cholesterol Level β

**Figure 5. Rare variant findings. a.** ClinGen P/LP variant enrichment across fine-scale ancestries. **b.** Differences across populations in the total numbers of rare ClinGene P/LP variants in ACMG genes. **c.** Differences across populations in the total numbers of rare predicted damaging missense and LOF variants in ACMG genes. In **b-c,** $n_{REF}$ is the total number of reference alleles, and $n_{ALT}$ of alternative alleles, which are the rare P/LP ClinGen variants in b, and the rare, predicted LOF/damaging missense variants in **c**. **d.** The numbers of rare computationally predicted damaging missense and LOF alleles per individual across ancestries. In **b-d**, the top panels show broad-scale ancestries, and the bottom fine-scale ancestries. Mann-Whitney U test with a Bonferroni correction was applied to test the difference in the distribution of rare LOF and predicted damaging missense counts per individual between any broad- or fine-scale group compared to all others. Statistically significant differences indicated by an asterisk (*). **e**. ExWAS significant results for selected traits/ancestries. **f-g.** Ancestry-specific deleterious variant frequencies (heatmap) and effect sizes (forest plot) for **f.** *GBA1* and **g.** *PCSK9*. In cases of small sample sizes, '–' was used instead of numeric values to protect patient privacy.

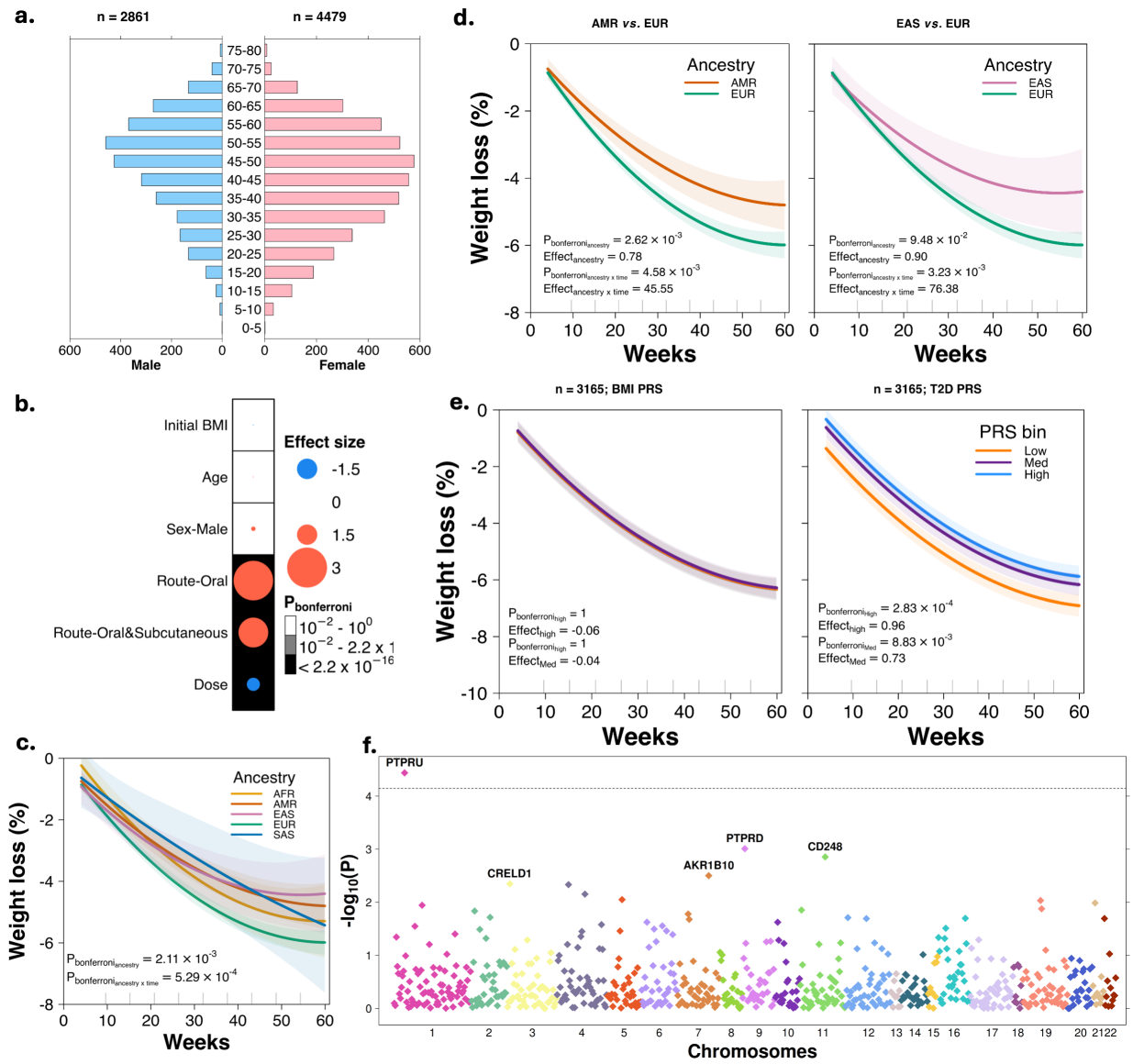


**Figure 6. Semaglutide efficacy. a.** Distribution of age by sex in semaglutide ATLAS users. **b.** The effect of non-genetic factors on weight loss in response to semaglutide. **c.** Weight loss patterns across genetic ancestry groups. Bonferroni-adjusted P-values are obtained from an ANOVA test on a linear mixed model with covariates. The overall number of weight measurements was 24,152, with a mean of 5 repeated measurements per patient. Ancestry sample sizes were: EUR , 3189; AFR, 373; AMR, 914 ; EAS, 291; SAS, 107. **d**. Differences in weight loss pattern between AMR and EUR populations (left) and EAS and EUR (right). Bonferroni-adjusted P-values were obtained from a linear mixed-effects model that included covariates. **e.** The relationship between PGS for BMI (left) and DM2 (right) and weight loss in response to semaglutide. Scaled PGS were divided into 3 groups: high, intermediate and low. The relationship between PGS bins and weight loss was tested using a linear mixed model. Only non-related EUR individuals were considered. In **c-e** linear mixed model fitted values are plotted with 95% CI, based on longitudinal data with repeated weight measurements. **f.** Semaglutide candidate protein gene-level test results. Presented are the -log10(P values) following a REGINIE additive model applied to test the genetic association between senaglutide-affected proteins and weight loss on semaglutide. The horizontal line shows the threshold for Bonferroni significant results. The top 5 hits are labeled by gene names. For plotting, gene positions across the genome were defined based on the first base pair.