nature genetics

Article

https://doi.org/10.1038/s41588-024-01894-5

Rare coding variant analysis for human diseases across biobanks and ancestries

Received: 26 February 2023

Accepted: 1 August 2024

Published online: 29 August 2024

Check for updates

Sean J. Jurgens[1,2,3,19], Xin Wang[1,3,19], Seung Hoan Choi[1,4], Lu-Chen Weng[1,3], Satoshi Koyama[1,3], James P. Pirruccello[1,5], Trang Nguyen[6], Patrick Smadbeck[6], Dongkeun Jang[6,7], Mark Chaffin[1], Roddy Walsh[2], Carolina Roselli[1], Amanda L. Elliott[7,8,9,10], Leonoor F. J. M. Wijdeveld[1,11], Kiran J. Biddinger[1], Shinwan Kany[1,12,13], Joel T. Rämö[1,3,10], Pradeep Natarajan[1,3,14], Krishna G. Aragam[1,3], Jason Flannick[6,15,16], Noël P. Burtt[6,7], Connie R. Bezzina[2], Steven A. Lubitz[1,3,17], Kathryn L. Lunetta[4,18] & Patrick T. Ellinor[1,3,17] ✉

Large-scale sequencing has enabled unparalleled opportunities to investigate the role of rare coding variation in human phenotypic variability. Here, we present a pan-ancestry analysis of sequencing data from three large biobanks, including the All of Us research program. Using mixed-effects models, we performed gene-based rare variant testing for 601 diseases across 748,879 individuals, including 155,236 with ancestry dissimilar to European. We identified 363 significant associations, which highlighted core genes for the human disease phenome and identified potential novel associations, including *UBR3* for cardiometabolic disease and *YLPM1* for psychiatric disease. Pan-ancestry burden testing represented an inclusive and useful approach for discovery in diverse datasets, although we also highlight the importance of ancestry-specific sensitivity analyses in this setting. Finally, we found that effect sizes for rare protein-disrupting variants were concordant between samples similar to European ancestry and other genetic ancestries ($\beta_{\mathrm{Deming}} = 0.7–1.0$). Our results have implications for multi-ancestry and cross-biobank approaches in sequencing association studies for human disease.

In recent years, the advent of large-scale sequencing has propelled studies into the role of rare coding variation in human phenotypic variability, including the human disease phenome[1–6]. However, for binary disease endpoints, previous work has had limitations in terms of power and/or statistical methodology. These limitations have included the use of simple tests that do not account for ancestry and other covariates, or models that produce miscalibrated test statistics for highly imbalanced phenotypes. Furthermore, discovery analyses typically focused on individuals of European (EUR) genetic ancestry[1,4,7,8], limiting interpretability, transferability and equity of genomic findings[9–11].

Several contemporary biobank initiatives have prioritized the inclusion of samples from understudied groups, with a goal to increase equitable understanding of health and disease[2,12–14]. Notably, the All of Us (AoU) research program has completed whole-genome sequencing (WGS) on almost 250,000 participants across the USA, of which most are enrolled from underrepresented communities and around half are of non-EUR genetic ancestry[12,15]. In this work, we set out to create a dataset of gene-based rare coding variant associations for human disease across large biobanks with sequencing data, and assess the role of diverse ancestral composition in rare variant association analyses.

Results

Ancestry distributions across three sequenced biobanks

We combined large-scale whole-exome sequence (WES) data from the UK Biobank (UKB) and the Mass General Brigham Biobank (MGB), with WGS data from AoU (Fig. 1). While several phenome-wide association

A full list of affiliations appears at the end of the paper. ✉e-mail: ellinor@mgh.harvard.edu

https://doi.org/10.1038/s41588-024-01894-5

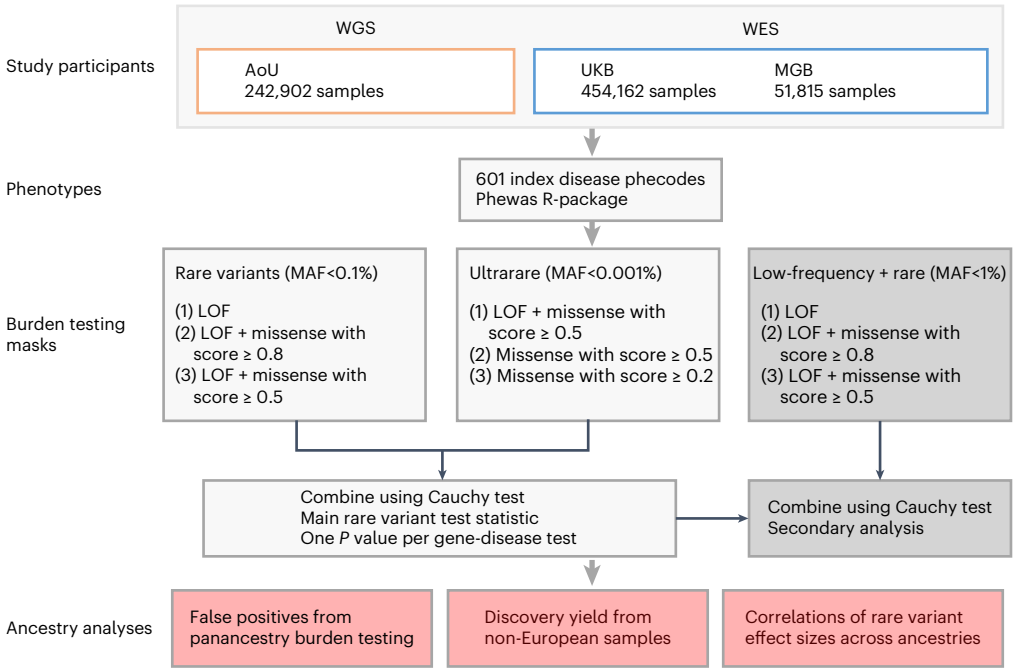


Fig. 1 | Study overview for rare variant discovery across human disease. Three studies were included in the analysis: AoU with WGS data, UKB with WES data and MGB with WES data. Over 600 disease Phecodes were identified using a hierarchal clustering algorithm. Disease Phecodes were analyzed using exome-wide gene-based testing of rare genetic variants using three masks (LOF, LOF+missense and ultrarare missense) after which *P* values were combined into a single *P* value using the Cauchy distribution for each gene–disease pair.

studies (PheWAS) for protein-coding variants have been published from the EUR ancestry subset of UKB[1,3,7], AoU and MGB represent less well-characterized cohorts. MGB is a health system biobank from eastern Massachusetts with a relatively high disease prevalence[16]. AoU is a diverse biobank that is actively enrolling participants from both health systems and via population-based ascertainment across the USA[12,17], with an emphasis on underrepresented groups. After quality-control procedures (Supplementary Note and Supplementary Fig. 1), we had a total of 748,879 individuals, including 454,162 from UKB, 242,902 from AoU and 51,815 from MGB (Supplementary Table 1).

Although ancestry is not truly categorical[18], we grouped individuals into principal continental ancestry groups based on their genetic similarity to samples from the 1000 Genomes project[19] (Supplementary Note), namely African (AFR), Admixed-American (AMR), East-Asian (EAS), EUR and South Asian (SAS) ancestries. Furthermore, individuals not falling clearly within predefined categories may not truly have 'mixtures' of such categorical ancestries; nevertheless, we refer to such samples as having 'admixed' ancestry. As expected, in our data, the ancestral diversity was greatest in AoU, with 49.9% of participants having a genetically determined ancestry other than EUR (most notably 21.0% AFR and 16.6% AMR ancestry). In contrast, 94.4% and 83.5% of samples from UKB and MGB were genetically determined to be of EUR ancestry (Fig. 2a and Supplementary Table 1). Across the three datasets, 119,660 individuals (16.0%) were similar to a defined continental ancestry other than EUR, and another 35,576 samples (4.7%) were of 'admixed' ancestry, totaling 155,236 samples with an ancestry dissimilar to EUR ancestry (20.7%) (Supplementary Table 1).

Phenotype and disease distributions

When comparing the three datasets, we found that several quantitative measures were similar across cohorts, including standing height, HDL cholesterol and creatinine levels (Supplementary Table 1). Nevertheless, several measures substantially differed between UKB and the US-based cohorts. Body mass index (BMI) and HbA1c levels were lower in UKB than in MGB and AoU, which potentially reflects higher obesity rates in the USA as compared with the UK[20]. LDL cholesterol, triglycerides and blood pressure levels were lower in AoU and MGB as compared with UKB (Supplementary Table 1), which might reflect different practices regarding hypertension control between the countries[21,22] and increased utilization of lipid-lowering therapy over the past decade, especially in the USA[23,24]. Overall, quantitative measures were comparable between AoU and MGB.

To define disease endpoints for our main genetic analyses, we created up to 1,866 phecodes (disease phenotypes) from International Classification of Disease (ICD) code mappings, after which we pruned down to a list of 601 index codes using a hierarchal clustering algorithm (Methods), as applied previously[7]. This approach was chosen to limit the number of highly correlated phenotypes—and thereby many potentially redundant rare variant associations—that have been found in many previous PheWAS approaches. In our primary analyses, we set a minimum of 50 cases, which left 546 phecodes in UKB, 601 in AoU and 601 in MGB (Supplementary Table 2). We note that not all AoU participants had complete electronic health record (EHR) linkage, although inclusion of such samples did not meaningfully affect any genetic association analyses (Supplementary Note and Supplementary Fig. 2).

In keeping with the health-system-based ascertainment of MGB, we found markedly higher phecode prevalence estimates in MGB as compared with UKB and AoU (Fig. 2b and Supplementary Note), as well as higher rates of likely pathogenic variants for cardiomyopathy (Extended Data Fig. 1). Furthermore, we found that disease prevalence estimates were generally lower in UKB than in AoU, probably due to sampling procedures and due to slightly different ICD coding systems (ICD in UKB and ICD-CM in AoU and MGB). Despite the differences between cohorts, disease prevalence estimates were highly correlated across datasets (Spearman's *r* in range of 0.7 and 0.9; Supplementary Fig. 3). Furthermore, gene-based effect sizes for three masks correlated reasonably between datasets (Supplementary Fig. 4).

Rare variant meta-analysis across biobanks

For the 601 phecode endpoints, we then performed exome-wide, gene-based burden testing in each dataset followed by a meta-analysis.

https://doi.org/10.1038/s41588-024-01894-5

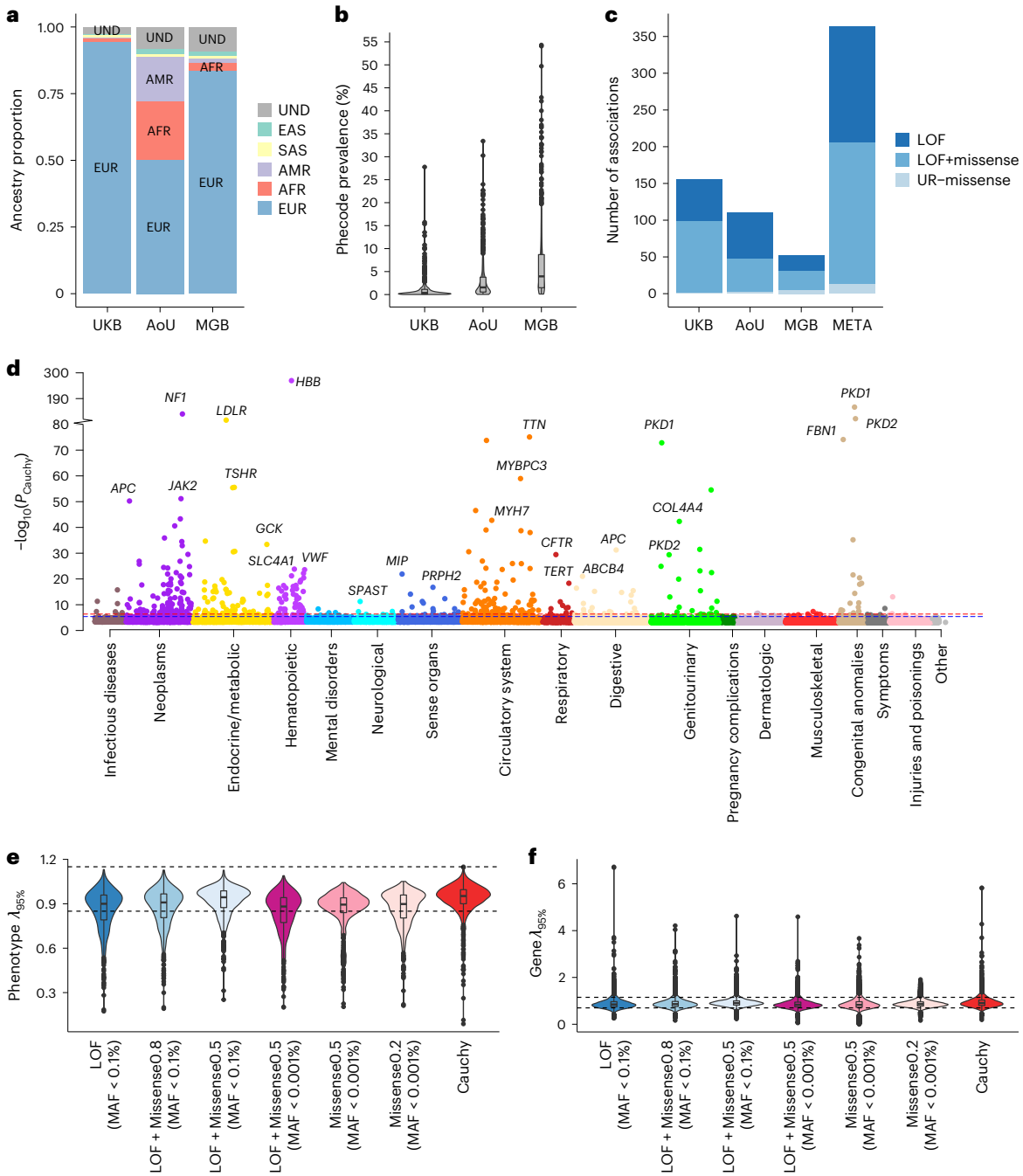


Fig. 2 | Multi-ancestry meta-analysis of rare genetic variation across three sequenced biobanks in over 750,000 individuals identifies 363 rare variant associations. a, Stacked bar chart with the proportion of each continental ancestry on the *y* axis and dataset on the *x* axis. Ancestral diversity was largest in AoU. **b**, Violin plot with overlaid boxplot showing the prevalence of Phecodes on the *y* axis and each dataset on the *x* axis. Plotted Phecodes were those included in the analysis with at least 50 cases in each dataset ($n = 546$). **c**, Stacked bar chart showing the number of identified disease associations (Cauchy $Q < 0.01$) on the *y* axis and each dataset and *x* axis, as well as the meta-analysis results. Bars are stacked by the class of mask that yielded the lowest *P* value (from LOF masks, LOF+missense masks and ultrarare missense masks). **d**, Multitrait gene-based Manhattan plot highlighting results from the overall meta-analysis, each dot representing one gene-trait test, with the $-\log_{10}$ of the Cauchy *P* value on the *y* axis and different disease categories on the *x* axis. For disease categories with strong associations, the top three nonredundant genes are annotated with the gene names. **e**, Violin plot with overlaid boxplot showing the distribution of inflation factors by phenotype (λ estimated at 95 percentile) on the *y* axis, and different rare variant masks on the *x* axis, as well as the distribution for the Cauchy combination results (on the far right). Dotted lines show the 0.75 and 1.25 cutoffs for inflation factor on the *y* axis. The number of phenotypes is 601 in all violins. **f**, Distribution of inflation factors by gene across the different masks and for the Cauchy combination results (on the far right), where the number of genes equals 14,388; 15,529; 17,809; 16,742; 15,462; 18,238 and 18,456. Cauchy *P* values represent the omnibus *P* value of all masks for a gene–phecode pair (unadjusted for multiple testing) after combining them using the Cauchy distribution. The Cauchy *Q* values represent the Benjamini–Hochberg FDR adjustments of these Cauchy *P* values. *P* values for mask–phecode pairs (before the Cauchy combination) were derived from *Z*-score-based meta-analyses of score tests from logistic mixed-effects models with SPA. All statistical tests and *P* values are two-sided. All boxplots show median (center), 25th percentile (bottom of box), 75th percentile (top of box), smallest/largest value within 1.5 × interquartile range from hinge (bottom/top whiskers, respectively), and datapoints outside of this range (dots). UND, undefined ancestry.

https://doi.org/10.1038/s41588-024-01894-5

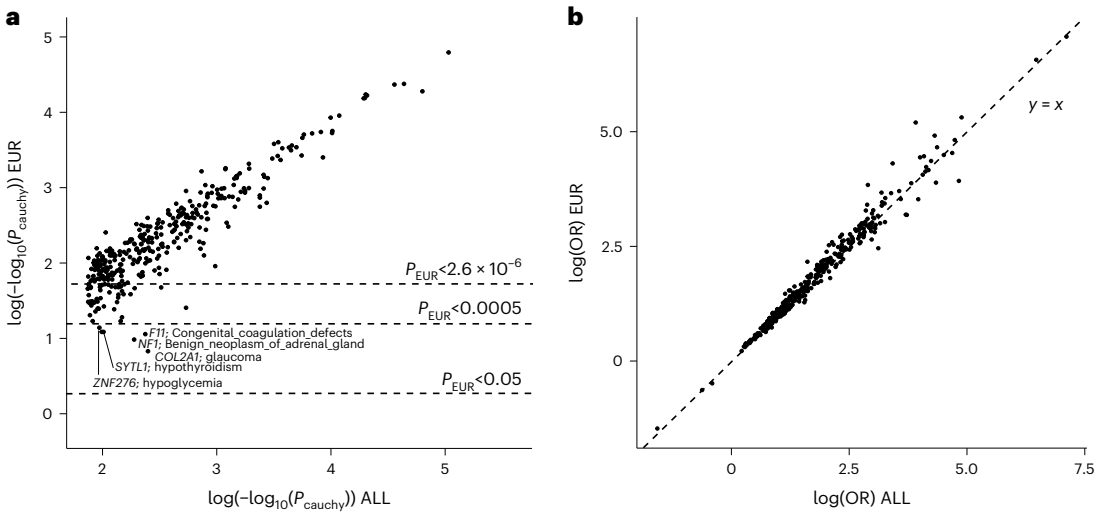


Fig. 3 | Assessment of bias from inclusion of non-EUR samples among the significant associations. a, Scatterplot with each dot representing a gene–phecode pair that reached test-wide significance in our primary analysis ($Q < 0.01$), with $-\log_{10}(P_{Cauchy})$ from the primary analysis on the *x* axis, and the $-\log_{10}(P_{Cauchy})$ derived from a EUR ancestry sensitivity analysis on the *y* axis (both log-transformed for clarity). Specific cutoffs on the *y* axis are highlighted using dotted lines. Any strong deviation of *P* values could indicate bias in our multi-ancestry approach, or alternatively indicate markedly lower power among EUR samples. No associations were abolished when restricting to EUR samples. There were six additional strongly attenuated genes ($0.05 > P_{EUR} > 0.0005$). Among these, several represent known gene–disease links (Supplementary Note). **b**, Scatterplot with the effect sizes for significant associations from the primary analysis on the *x* axis, with the effect sizes from EUR-only sensitivity analyses on the *y* axis. The effect size for the most significant mask is plotted for each gene–phecode pair, restricting to masks that had adequate allele counts in both the primary analysis and in the sensitivity analysis (cMAC ≥ 20). Any large deviations from the dotted line ($x = y$) indicate bias from our multi-ancestry approach. Strikingly, no strong deviations of effect sizes were observed in this sensitivity analysis. For eight associations, there were insufficient alleles among EUR ancestry samples to compute an effect size, although represented well-known gene–disease links (Supplementary Note). Together, these results show that the bias from inclusion of non-EUR samples was not substantial. Bias is defined here as the spurious change in effect sizes/test statistics that is caused by inclusion of several ancestries but is not caused by true biological differences. Cauchy *P* values represent the omnibus *P* value of all masks for a gene–phecode pair (unadjusted for multiple testing) after combining them using the Cauchy distribution. The Cauchy *Q* values represent the Benjamini–Hochberg FDR adjustments of these *P* values. *P*- values for mask–phecode pairs (before the Cauchy combination) were derived from *Z*-score-based meta-analyses of score tests from logistic mixed-effects models with SPA. All statistical tests and *P* values are two-sided. ORs were estimated using inverse-variance-weighted meta-analysis of two-sided Firth's logistic regression results. ALL, all-ancestry individuals.

We assessed six rare variant masks, including various combinations of loss-of-function (LOF) variants and missense variants, and using various frequency filters (maximum continental population minor allele frequency ($MAF_{population\text{-}max}$) < 0.1% and < 0.001%). For each gene–phecode pair, we combined *P* values from each mask into one *P* value using the Cauchy distribution (Methods; Fig. 1). For various sensitivity analyses, we also performed burden testing inclusive of both rare and low-frequency variants ($MAF_{population\text{-}max}$ < 1%; Fig. 1). Analyses in AoU and MGB yielded mostly positive-control associations, including many that were identified in the large UKB dataset; conversely, there were associations where AoU/MGB afforded better yield (that is, associations did not reach significance in UKB) (Supplementary Note).

Per-cohort test statistics were very well calibrated for all masks, as well as for the Cauchy combination, highlighting the robustness of our mixed-effects regression framework (Supplementary Figs. 5–10). In an initial meta-analysis, however, we found an earlier-than-expected deviation of test statistics (Supplementary Figs. 11 and 12). This inflation was due largely to the meta-analysis of AoU and MGB. Given partial recruitment from same sites, we investigated test-wide correlations of test statistics between AoU and MGB in more detail; phecodes showed a median of ~0.05 exome-wide correlation in test statistics (Supplementary Fig. 13 and Supplementary Note). We therefore applied a *Z*-score-based meta-analysis with correction for sample overlap, which markedly improved the calibration (especially for high allele count masks that were most affected; Supplementary Figs. 14 and 15). As large sequencing biobanks continue to grow, issues relating to sample overlap will also increase; in future, central biobank policies might need reconsideration to allow identification of overlapping participants. Considering acceptable calibration of test statistics using our approach, we proceeded with the overlap-corrected meta-analysis.

Genetic association data quality

When assessing individual datasets, the largest number of significant associations was observed within the UKB ($n = 185$ at Benjamini–Hochberg false-discovery rate (FDR) $Q < 0.01$; FDR across all genes by all phecodes), while MGB yielded the fewest associations ($n = 52$ at $Q < 0.01$; Fig. 2c, Supplementary Figs. 5–8 and Supplementary Table 3). Across the 11,060,516 unique gene–phecode pairs in our multi-ancestry meta-analysis, 363 gene-based associations reached significance at an overall FDR $Q < 0.01$ (Fig. 2d and Supplementary Table 3), comprising 165 unique phenotypes and 123 unique genes. Of note, 464 signals would have been identified in a naïve meta-analysis without correction for sample overlap; in a meta-analysis omitting MGB, we would have identified 319 significant associations. After correction for sample overlap, meta-analysis test statistics were reasonably calibrated within different bins for disease case counts and rare variant carrier counts (Supplementary Figs. 14–17 and Supplementary Table 4). Consistently, no individual phecode showed evidence of marked test statistic inflation in our final meta-analysis (all $\lambda_{95\%} < 1.16$; Fig. 2e).

In contrast, we found several genes with strong inflation ($n = 198$ genes with $\lambda_{95\%} > 1.5$, 11 genes with $\lambda_{95\%} > 2.5$; Fig. 2f and Supplementary Table 5). Per gene inflation may be caused by uncorrected confounders (that is, overlap, population stratification), stochastics (given small number of tests per gene; ≤601) or, alternatively, by widespread deleteriousness or pleiotropy of rare variants in the gene. In support of the latter, we found that many inflated genes represent known causes

of Mendelian disease (for example, *PKD1*, *APC*, *TTN* and *FBN1*), for which inflation was most prominent in relevant disease categories (Supplementary Fig. 18). Furthermore, inflated genes were enriched for LOF intolerance[25,26] (LOEUF < 0.5: odds ratio (OR) 2.8, 95% confidence interval (CI) [2.1, 3.7], $P = 5.5 \times 10^{-12}$; pLI > 0.9: OR 2.6, 95% CI [1.9, 3.6], $P = 4.3 \times 10^{-9}$; two-sided Fisher exact tests). In a sensitivity analysis restricting to samples of EUR ancestry, we observed a markedly better test statistic calibration for a minority of genes, but for most it was not substantial (Supplementary Fig. 18). Finally, we observed that a matched analysis of two synonymous masks yielded no signals at FDR $Q < 0.1$ (Supplementary Note), with the most significant signals including *IGLL5* for white blood cell-related traits[27,28]. These results suggest that a substantial proportion of gene-based inflation in our primary analysis was due to deleterious effects and/or stochastics, although a degree of finer (subcontinental) population stratification cannot be excluded.

Assessment of bias from pan-ancestry analyses in diverse populations

Given the increasing numbers and size of ancestrally diverse biobanks (for example, AoU), it is important to understand whether pan-ancestry burden testing yields reasonable results. We therefore assessed the potential bias introduced by performing pan-ancestry analyses. To this end, we performed sensitivity analyses restricting to individuals with genetic ancestry similar to EUR ancestry (Fig. 3, Supplementary Note and Supplementary Table 6). For the 363 significant signals, we compared the *P* values from the all-ancestry and EUR-ancestry analysis, which flagged six potentially problematic associations that were markedly weaker in EUR ancestry individuals (Fig. 3a). However, several of the weakened signals represented well-known gene–disease links, and comparison of log(OR) estimates showed a very high consistency between the pan-ancestry and EUR-ancestry analysis (Fig. 3b).

To assess the effect of ancestry bias in a highly diverse dataset, we then repeated these analyses restricting to AoU only (Supplementary Note). In AoU, we found a limited number of likely false-positive signals driven by potential ancestry bias (2 of 111–121 signals; one gene; Supplementary Fig. 19). Finally, we assessed whether genes were associated with our categorical ancestry outcomes. While several gene burdens were associated with ancestry, the significant genes from our primary analysis did not overlap ancestry-associated genes (Supplementary Note).

Together, our findings indicate that pan-ancestry burden testing—using mixed regression-based methods—may be a reasonable and inclusive approach to identify rare variant association signals in diverse datasets where cases and controls are well-represented across continental ancestries. At the same time, our results outline important ancestry-specific sensitivity analyses that should be considered to scrutinize such signals.

Somatic variation impacting sequencing association studies

We noticed several genes associated with clonal hematopoiesis of indeterminate potential (CHIP) among the inflated genes[29–31]. We explored somatic variation further, through prediction of age by rare variant carrier status, and by evaluation of the phenotypic associations found for known CHIP and known somatic leukemia genes (Supplementary Note). As expected, known CHIP genes were associated most strongly with age (*DNMT3A*, *TET2*, *SRSF2*, *SF3B1* and *ASXL1*; Extended Data Fig. 2a). These CHIP genes—and several known somatic leukemia genes (*TP53*, *NOTCH1*, *IDH2*, *KLHL6*, *RUNX1*, *CHD2* and *DDX41*)—were also associated with hematological traits and leukemic outcomes (Supplementary Tables 6–8 and Extended Data Fig. 2b). We conclude that somatic variation affecting hematological outcomes and CHIP genes is likely, although most of the significant associations are probably causal—albeit by somatic variation rather than germline variation. The effect of somatic variation in driving associations for nonhematological traits seems small in our dataset (Supplementary Note). Nevertheless, we advise careful interpretation of results from sequencing of blood-derived DNA for hematological outcomes and known hematological genes.

Genetic effects of core genes for the human disease phenome

Among the 363 significant associations in our meta-analysis, 301 were reported directly in the Online Mendelian Inheritance in Man (OMIM) database or were plausibly related to entries in this database (82.9%; Supplementary Table 8). Indeed, the significant signals from our analyses highlight pleiotropic disease genes (that is, those associated with several disease outcomes and sequelae) and genes associated with large effect sizes (Fig. 4), pointing towards core genes for the human disease phenome.

Notable examples include associations of *FBN1*—a causative gene for Marfan syndrome (MIM 154700)—with 13 diseases across cardiovascular and genetic disease codes (Fig. 4a and Supplementary Tables 6–8). *FBN1* showed the largest effect size for 'chromosomal anomalies' and 'genetic disorders' (OR_{LOF} 569.08, $P_{Cauchy} = 9.3 \times 10^{-75}$; Fig. 4b, c and Supplementary Table 6). Similarly, the known adenomatosis poli gene *APC* was associated with colorectal cancer (MIM 175100; $P_{Cauchy} = 2.8 \times 10^{-18}$, OR_{LOF} 12.7) and 22 other codes related largely to gastrointestinal disease (Fig. 4 and Supplementary Tables 6–8). The largest number of gene-based associations was identified for *PKD1*—a gene causative in autosomal polycystic kidney disease (MIM 173900). *PKD1* associated with 29 codes (Fig. 4a and Supplementary Tables 6–8), most notably genitourinary congenital anomalies ($P_{Cauchy} = 1.1 \times 10^{-153}$, OR_{LOF} 78.71) and chronic renal failure ($P_{Cauchy} = 1.8 \times 10^{-73}$; OR_{LOF} 17.36). Notably, the present analysis identifies various disease sequelae associated with known Mendelian diseases genes (such as *PKD1* and *APC*), many of which were not identified in previous PheWAS approaches (Supplementary Table 8).

Similarly, our large sample size allowed identification of many Mendelian gene–disease links that were not observed in previous PheWAS (Supplementary Table 8). For example, *PTEN* was associated with several phenotypes that recapitulate Cowden syndrome (MIM 158350), including congenital anomalies and thyroid disease; *LMNA* and *TNNT2* were associated with cardiomyopathy and various sequelae (MIM 601494; 135150); *CFTR* was associated with cystic fibrosis (MIM 219700); *FLCN* was associated with congenital anomalies and renal cancers (MIM 135150); *SMAD3*, *COL3A1* and *LDLR* were associated with vascular aneurysms (MIM 613795, 130050, 143890); *PAX6* was associated with congenital diseases of the eye (MIM 120430); (potentially somatic) variants in *TP53* were associated with various cancers; *NODAL* was associated with congenital heart disease (MIM 270100) and *SOD1* was associated with anterior horn cell disease (MIM 105400). These results highlight how the continued growth in sequencing is enabling an increased detection of bona fide Mendelian contributors to the disease phenome.

Most of the significant associations were driven by masks that combined LOF variants with missense variants (for example, a LOF+missense mask had the lowest *P* value; $n = 193$ associations), while LOF-only masks drove results for 157 gene–phecode pairs and missense-only masks drove results in only 13 cases (Fig. 2c). For instance, among the highly pleiotropic genes, associations for *LMNA*, *TP53*, *BRCA2* and *BRCA1* were strongest for LOF+missense masks, while associations for *MYH7* were driven largely by ultrarare missense variation (Fig. 4a and Supplementary Tables 6 and 7), consistent with genetic mechanisms in cardiomyopathy[32,33].

Understanding the effect sizes conferred by rare variants from a genome-first view may enable unbiased interpretation of risk and allow comparison with common variant effects. For disease categories with multiple associations, we tabulated the distributions of effect sizes (Fig. 4b and Supplementary Table 9). For 'circulatory system', the median OR for LOF variants was 4.5 (first quantile–third quantile, Q1–Q3 (2.7–16.6), 53 pairs), with the largest effect identified for *FBN1*

https://doi.org/10.1038/s41588-024-01894-5

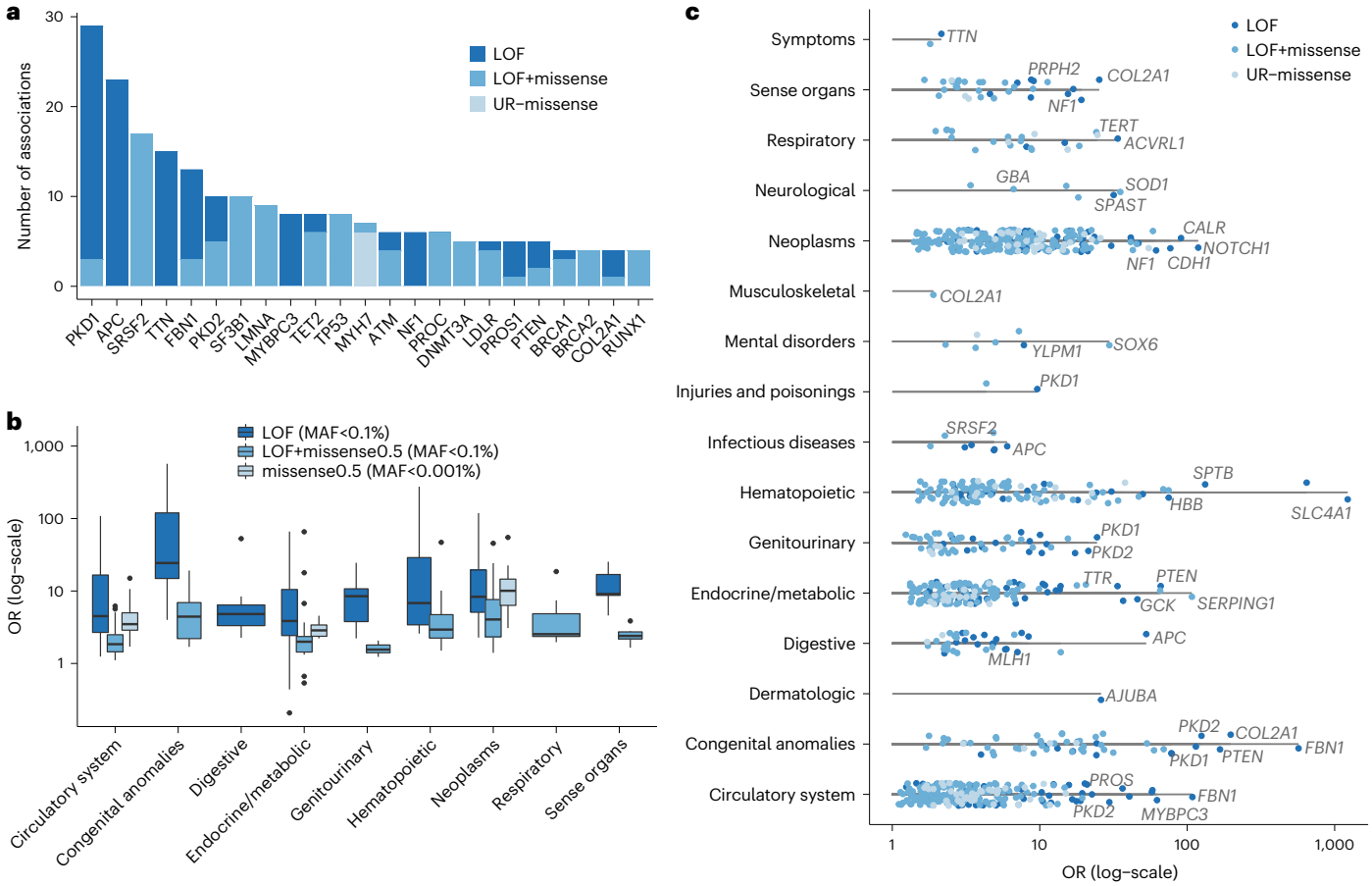


Fig. 4 | Large genetic effect sizes and pleiotropic associations identify core genes for the human disease phenome. a, Stacked barcharts for all genes from the meta-analysis that showed at least three associations, with the number of associations on the *y* axis and gene on the *x* axis. Bars are stacked by the class of the best mask (LOF, LOF+missense masks or ultrarare (UR) missense masks) for each gene–trait association. **b**, Grouped boxplots showing rare variant effect size distributions per Phecode category, with log-scaled ORs on the *y* axis and categories on the *x* axis. The figure is restricted to gene–trait associations reaching Cauchy $Q < 0.01$ and restricting to rare variant masks with $P < 2.6 \times 10^{-6}$. Per category, only masks with at least seven associations are shown (and therefore some categories do not show all masks and not all categories are plotted). In boxplots, the number of contributing associations from left to right equals 53, 44, 23, 15, 9, 16, 37, 30, 9, 15, 8, 22, 32, 54, 58, 18, 9, 9 and 8. All boxplots show median (center), 25th percentile (bottom of box), 75th percentile (top of box), smallest/largest value within 1.5 × interquartile range from hinge (bottom/top whiskers, respectively) and datapoints outside of this range (dots). **c**, Multiple jittered lollipop chart showing rare variant effect sizes for each Phecode category. The *x* axis shows the log-scaled OR with each dot representing an association (restricting to gene–trait associations with Cauchy $Q < 0.01$ and rare variant masks with $P < 2.6 \times 10^{-6}$), and Phecode categories on the *y* axis. Horizontal lines start at 1 and end at the largest estimated effect size within the category. Dots are colored by class of rare variant mask. Select genes are annotated within each category to highlight large effect size genes for the respective category. In all panels, ORs were estimated using inverse-variance-weighted meta-analysis of two-sided Firth's logistic regression results, while mask–phecode *P* values were estimated from *Z*-score-based meta-analysis of score tests from logistic mixed-effects models with SPA. Cauchy *P* values represent the omnibus *P* value of all masks for a gene–phecode pair after combining them using the Cauchy distribution (unadjusted for multiple testing), while the Cauchy *Q* values represent the Benjamini–Hochberg FDR adjustments of these *P* values. All statistical tests and *P* values are two-sided.

and aortic aneurysm (OR_{LOF} 108.5; Fig. 4c and Supplementary Tables 6 and 9). The category 'neoplasms' showed a median OR_{LOF} of 8.3 (Q1–Q3 (5.1–19.5), 54 pairs) with the largest effect conferred by (potentially somatic) variants in the leukemia gene *NOTCH1* (OR_{LOF} 118.9; Fig. 4b,c). As expected, the largest median effect size was identified for associations in the category 'congenital anomalies' (OR_{LOF} of 24.3 (Q1–Q3 (15.0–119.8), 15 pairs; Fig. 4b and Supplementary Table 9). Although we acknowledge that association yield is determined by statistical power, and therefore larger sample sizes may identify additional smaller-effect associations, our current work provides a useful reference of human Mendelian variation for common disease categories within the adult population.

Power from pan-ancestry approaches

In our pan-ancestry approach, we identified markedly more significant associations than a restrictive approach including only individuals with ancestries similar to the largest continental ancestry in our dataset (EUR ancestry; 18.2% fewer associations, $n = 297$). The improved yield may reflect the larger total sample size or additional power afforded by the inclusion of diverse ancestries. To assess this more formally, we downsampled AoU to an ancestrally diverse dataset of equal sample size to the EUR ancestry subset (Methods)—$n = 106{,}057$ samples with complete EHR linkage—and reran our main rare variant analyses. We observed comparable or slightly fewer numbers of significant signals in the ancestrally diverse subsets of AoU than in the EUR subset (Supplementary Table 10). When including low-frequency variant masks ($MAF_{population\text{-}max} < 1\%$), we still did not observe a yield benefit of the ancestrally diverse subset as compared with the EUR-only subset (Extended Data Fig. 3).

Case frequencies across different ancestries may have contributed to this finding. Overall, disease prevalences were higher among samples with genetically determined EUR ancestry, as compared with

other individuals (Extended Data Fig. 3c and Supplementary Tables 11). It currently remains unclear whether this may represent an artifact of AoU sampling[34] or potentially reflects broader bias in medical care in underrepresented populations[35,36]. Despite this, removal of phecodes that were enriched strongly among EUR ancestry samples did not alter the results markedly (Supplementary Table 12).

For common variant genome-wide association studies, it has been shown that ancestry-specific variants may contribute strongly to genetic signal and discovery[37–39]. By contrast, our analyses did not establish a marked increase in discovery yield from ancestral diversity for rare variant burden testing of disease phenotypes, at a phenome-wide scale. This result may partly represent higher disease frequency among individuals genetically similar to EUR ancestry. In addition, we note that (1) certain distinct rare variant signals may exist in populations dissimilar to EUR ancestry; (2) specific phenotypes might have increased yield in populations dissimilar to EUR ancestry, for instance if the phenotype is enriched in that population (Extended Data Fig. 3c); and (3) our results may not translate to founder populations and populations with high degrees of consanguinity[8,40]. Furthermore, it is possible that sample sizes for underrepresented groups currently remain too small to confer meaningful boosts in power for burden testing.

Rare variant signals informing disease biology

We identified several biologically plausible gene–disease links, which were described recently in biobank studies[4,5,41,42] (Supplementary Tables 6–8). These included *PIEZO1*, which encodes a mechano-sensing protein, with varicose veins (OR_{LOF} 1.92; $P_{Cauchy} = 5.3 \times 10^{-8}$); *AJUBA*, which encodes a protein involved in cell–cell adhesion, with erythematosquamous dermatosis (OR_{LOF} 26.1; $P_{Cauchy} = 2.1 \times 10^{-7}$); and *GIGYF1*, which encodes a regulator of insulin-like-growth-factor signaling, with type 2 diabetes (OR_{LOF} 3.3; $P_{Cauchy} = 4.8 \times 10^{-14}$).

Overall, 42.4% of significant associations (154 out of 363) did not reach significance in two previous biobank-scale PheWAS (Supplementary Table 8); 8.8% of associations (32 out of 363) were also not reported in the OMIM database. Among these signals, several were consistent with recent literature. For instance, the association between SRCAP-complex-encoding genes *DMAP1* (OR_{LOF} 3.7; $P_{Cauchy} = 6.3 \times 10^{-8}$) and *YEATS4* ($OR_{LOF}$ 3.8; $P_{Cauchy} = 5.6 \times 10^{-8}$) with benign neoplasms of uterus[43]; *APOB*, which encodes a lipid particle apolipoprotein, with chronic liver disease and cirrhosis[44] (OR_{LOF} 2.3; $P_{Cauchy} = 1.0 \times 10^{-8}$); and *NOS3*, which encodes a nitric oxide synthase, with ischemic heart disease[45] (OR_{LOF} 1.7; $P_{Cauchy} = 9.1 \times 10^{-9}$).

We additionally focus on select novel findings, restricting to those that survived sensitivity analyses (Supplementary Table 6). For instance, we found that rare variants in *YLPM1* were associated significantly with bipolar disorder ($P_{Cauchy} = 8.1 \times 10^{-9}$; OR_{LOF} 3.9) and personality disorders ($P_{Cauchy} = 2.0 \times 10^{-7}$; OR_{LOF} 7.8; Extended Data Fig. 4). Common variants near *YLPM1* are associated with mood instability, depressed affect and neuroticism[46–48], and OpenTargets[49] reports strong colocalization for a *YLPM1* eQTL with 'feeling worry' and 'feeling nervous' (posterior probability of colocalization >0.8). Furthermore, a rare *YLPM1* missense variant was among several variants cosegregating with apparent autosomal-dominant bipolar disorder in one pedigree[50]. In a recent exome sequencing study of bipolar disorder, ultrarare *YLPM1* LOF and missense variants reached nominal significance (OR 3.4; $P = 0.01$, one-sided Fisher exact test), although some case overlap with our discovery samples is possible[51]. *YLPM1* is expressed in many tissues, including brain, although it has not been widely studied functionally.

We further found that *UBR3* variants were associated with an adverse metabolic profile, including hypertension ($P_{Cauchy} = 6.7 \times 10^{-9}$; OR_{LOF} 2.8), type 2 diabetes ($P_{Cauchy} = 3.8 \times 10^{-8}$; OR_{LOF} 3.6) and a suggestive signal for obesity ($P_{Cauchy} = 1.8 \times 10^{-6}$; OR_{LOF} 2.6; Extended Data Fig. 5). In OpenTargets, there was moderate evidence for colocalization between a common *UBR3* sQTL and BMI-adjusted waist-to-hip ratio (posterior probability for colocalization of 0.66). A previous mutational screen in mice identified *Ubr3* loss as a strong inducer of increased weight and fat-to-lean mass in both male and female mice[52], and a paralog of *UBR3*, *UBR2*, was found in a recent sequencing study for BMI[53]. *UBR3* and *UBR2* are highly constrained (pLI = 1) and encode ubiquitin protein ligase components[54].

Other novel associations include *MIB1*, which encodes a Notch signaling protein[55] found to regulate pancreatic β-cell formation in mice[56], with type 2 diabetes ($P_{Cauchy} = 5.3 \times 10^{-8}$; OR_{LOF} 1.3), and *SYTL1*, which encodes a synaptotagmin, a protein class involved in neuronal and endocrine exocytosis[57,58], with hypothyroidism ($P_{Cauchy} = 6.5 \times 10^{-8}$; OR_{LOF} 1.7). Although we identified initial replication evidence for these genes (Supplementary Note), new associations will require further external replication in other large datasets.

Consistency of rare variant effects across ancestries

Finally, we asked whether our multi-ancestry dataset could answer whether the effects of rare coding variation for human disease are consistent across ancestries. To this end, we used a three-sample approach to assess whether effects are consistent between EUR ancestry samples and individuals of other genetic ancestries. We first identified suggestively significant signals ($P < 2.6 \times 10^{-6}$) from a meta-analysis of EUR individuals from UKB and MGB, and then assessed the effect sizes of these signals in the diverse AoU dataset (Methods).

Phenome-wide significant burden effect sizes from EUR ancestry samples correlated well with the estimated effects from other ancestries (Fig. 5 and Supplementary Table 13), similar to previous *trans*-ancestry findings for common variants[59–61] and quantitative traits[1]. To better incorporate measurement error and assess calibration, we then used Deming regression (Methods). We found highly significant slopes (all $P < 2.4 \times 10^{-8}$) for LOF and missense masks, which, in most cases, were consistent with a calibration of 1 (Fig. 5 and Supplementary Table 13). For instance, for LOF variant masks, the regression slope was 0.9 for EUR versus AFR ancestry ($P = 3.9 \times 10^{-23}$, 95% CI (0.72, 1.08)), and 0.9 for EUR versus AMR ancestry ($P = 6.4 \times 10^{-47}$, 95% CI (0.78, 1.02)).

We then performed several sensitivity analyses. These included the removal of genes associated with age and/or leukemic outcomes, and analyses accounting for bins of effective sample size. These analyses produced largely consistent results, although estimated coefficients for ultrarare missense variants tended to be somewhat attenuated (Fig. 5 and Supplementary Tables 14 and 15). Finally, for LOF variants, we used a random-effects inverse-variance-weighted (IVW) approach to combine phenotype-specific results for phecodes with at least three qualifying genes. Although we caution against overinterpretation of the individually noisy estimates, the meta-analysis yielded consistent results when comparing EUR ancestry with non-EUR ancestry ($\beta_{Deming\text{-}IVW}$ 0.82–0.87, $P < 2 \times 10^{-17}$) and when comparing EUR ancestry with AFR ancestry ($\beta_{Deming\text{-}IVW}$ 0.84–0.87, $P < 0.004$; Supplementary Table 15).

Broadly, our results provide evidence that effect sizes for rare LOF variants have reasonable consistency between EUR and other genetic ancestries, justifying further *trans*-ancestry approaches to improve discovery power in disease sequencing association studies. In addition, these analyses support the notion that causal variants share high consistency in their effects across different ancestries[62]. Nevertheless, our analyses assume homogeneous effects across phenotypes and genes; subgroup analyses with respect to specific diseases and genes were not adequately powered at the current sample size and remain directions for future work. Furthermore, our analyses were not powered to assess other principal continental ancestries (for example, Asian ancestry) at this time.

Publicly available data via the Human Disease Knowledge Portal

We have released a web portal to browse our gene-based results through the Human Disease Knowledge Portal (https://hugeamp.org:8000/research.html?pageid=600_traits_app_home). Users may

https://doi.org/10.1038/s41588-024-01894-5

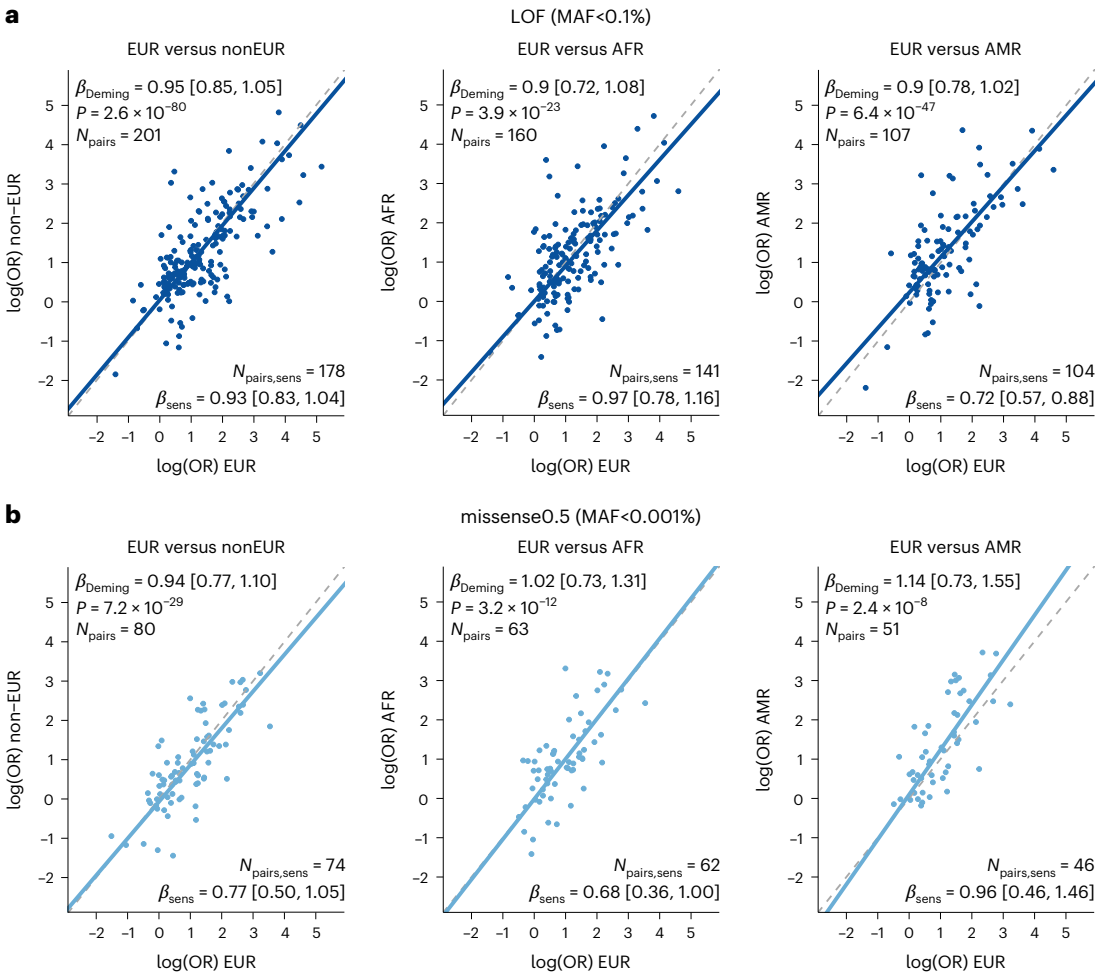


Fig. 5 | Effect sizes of rare coding variants for disease correlate between genetic EUR and other genetic ancestries. Scatterplots with the effect sizes from EUR-ancestry analyses on the *x* axis with the respective effect sizes estimated among individuals dissimilar to EUR ancestry on the *y* axis. In each panel, a three-sample design was applied: significant mask–disease pairs were identified from a EUR meta-analysis of UKB and MGB (significance determined at $P < 2.6 \times 10^{-6}$), after which those mask–disease pairs were assessed within different ancestry groups from the AoU dataset. Each panel shows effect sizes (that is, log(OR)) for EUR analysis on the *x* axis and effect sizes from other ancestries on the *y* axis; the left panels show EUR versus all non-EUR samples, the middle panels show EUR versus AFR samples and the right panels show EUR versus AMR samples. **a**, Results for rare LOF variant masks with at least 20 carriers in both ancestry assignments. **b**, Results for ultrarare missense0.5 variant masks with at least 20 carriers in both ancestry assignments. Linear trend lines from error-in-variable total-least-squares Deming regression are added to the plots. Statistics from Deming regression, including estimated β (95% CI) and *P* values, are added in text in the top left corners. A regression coefficient (β_{sens}) and 95% CI is also provided in the bottom right corners, showing results from a combined sensitivity analysis where genes associated with age or leukemic outcomes are removed, and where analyses are adjusted for quantiles of effective sample size (Supplementary Table 14). All ORs were estimated using Firth's logistic regression models among unrelated participants. Deming regression was run using beta coefficients and their standard errors, making the analysis comparable with York regression with the assumption of uncorrelated errors. Standard errors were computed using Jackknife estimators. All statistical tests and *P* values are two-sided. sens, sensitivity analysis.

browser results from individual datasets (UKB, AoU and MGB), various meta-analyses (including uncorrected and sample overlap-corrected meta-analyses), different ancestries (all-comers or EUR ancestry only) and various mask filters ($MAF_{population\text{-}max} < 0.1\%$; $MAF_{population\text{-}max} < 1\%$). For instance, our portal highlights eight and nine exome-wide significant ($P < 1 \times 10^{-6}$) genes for cardiomyopathy and diabetes mellitus, respectively (Extended Data Figs. 6 and 7). For de novo discovery or replication, researchers might want to restrict to specific subsets of our data; to this end, the individual cohort results and results for meta-analyses of UKB + AoU and UKB + MGB are available.

Discussion

The present work is imperfect and subject to several limitations. First, phecode definitions may imperfectly capture disease status, and therefore a degree of phenotype misclassification is likely. Second, the overlapping samples between AoU and MGB may have introduced bias despite applying an overlap-aware meta-analysis, as rare variant burdens may be affected differently by sample overlap compared with single common variants. Reassuringly, however, the uncorrected inflation was most notable for higher allele count burdens—expected to behave more similarly to common variants—whereas few significant results were observed for low allele count burdens (Supplementary Figs. 16 and 17). Third, while our main signals were not driven by continental population stratification, it is possible that finer population stratification introduced some bias. Fourth, we applied a more liberal cutoff based on an FDR of 1% in our PheWAS analyses. For all the above reasons, any specific new gene–disease links will require replication in independent datasets. Fifth, our statistical analyses were focused specifically on rare variant burden testing and might not translate to rare single variant and/or variance component tests. Relatedly, our analyses of discovery yield in diverse datasets may still have been limited by sample size. Future studies with even larger diverse datasets

might be needed to identify benefits for rare variant burden testing, especially considering our focus on binary outcomes. Finally, our analyses were restricted to protein-coding genes, while rare noncoding regions remain largely unexplored on a population scale. The AoU research program aims to eventually release WGS data on over 1million participants, and the UKB recently made WGS data available on almost 500 thousand samples; these data will be instrumental to extend our findings to additional populations and noncoding regions.

In conclusion, through pan-ancestry meta-analysis of over 750,000 sequences, we present a dataset of gene-based rare variant associations across a wide range of human disease phenotypes. Our results provide insights into the consistent effects of ultrarare coding genetic variation for human disease across ancestries, while providing analytical implications for future sequencing approaches. These findings are of relevance given the important and continued efforts to sequence underrepresented populations[10–13,63,64]. To propel use of our data, we have made our results available for download and browsing in the Human Disease Knowledge Portal.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01894-5.

References

1. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
2. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
3. Wang, Q. et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
4. Karczewski, K. J. et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168 (2022).
5. Jurgens, S. J. et al. Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nat. Genet.* **54**, 240–250 (2022).
6. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
7. Sun, B. B. et al. Genetic associations of protein-coding variants in human disease. *Nature* **603**, 95–102 (2022).
8. Heyne, H. O. et al. Mono- and biallelic variant effects on disease at biobank scale. *Nature* **613**, 519–525 (2023).
9. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489–494 (2009).
10. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
11. Hindorff, L. A. et al. Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
12. Ramirez, H. A. et al. The All of Us Research Program: data quality, utility, and diversity. *Patterns* **3**, 100570 (2022).
13. Gurdasani, D. et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).
14. Gaziano, J. M. et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
15. All of Us Research Program Genomics Investigators. Genomic data in the All of Us research program. *Nature* **627**, 340–346 (2024).
16. Koyama, S. et al. Decoding genetics, ancestry, and geospatial context for precision health. Preprint at *medRxiv* https://doi.org/10.1101/2023.10.24.23297096 (2023).
17. Denny, J. C. et al. The 'All of Us' research program. *N. Engl. J. Med.* **381**, 668–676 (2019).
18. Ding, Y. et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).
19. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
20. Janssen, F., Bardoutsos, A. & Vidra, N. Obesity prevalence in the long-term future in 18 European countries and in the USA. *Obes. Facts* **13**, 514–527 (2020).
21. Marshall, A. et al. Comparison of hypertension healthcare outcomes among older people in the USA and England. *J. Epidemiol. Community Health* **70**, 264–270 (2016).
22. Joffres, M. et al. Hypertension prevalence, awareness, treatment and control in national surveys from England, the USA and Canada, and correlation with stroke and ischaemic heart disease mortality: a cross-sectional study. *BMJ Open* **3**, e003423 (2013).
23. Matyori, A., Brown, C. P., Ali, A. & Sherbeny, F. Statins utilization trends and expenditures in the U.S. before and after the implementation of the 2013 ACC/AHA guidelines. *Saudi Pharm. J.* **31**, 795–800 (2023).
24. Gao, Y., Shah, L. M., Ding, J. & Martin, S. S. US trends in cholesterol screening, lipid levels, and lipid-lowering medication use in US adults, 1999 to 2018. *J. Am. Heart Assoc.* **12**, e028205 (2023).
25. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
26. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
27. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
28. Jurgens, S. J. et al. Adjusting for common variant polygenic scores improves yield in rare variant association analyses. *Nat. Genet.* **55**, 544–548 (2023).
29. Jaiswal, S. Clonal hematopoiesis and nonhematologic disorders. *Blood* **136**, 1606–1614 (2020).
30. Asada, S. & Kitamura, T. Clonal hematopoiesis and associated diseases: a review of recent findings. *Cancer Sci.* **112**, 3962–3971 (2021).
31. Mitchell, E. et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
32. Ingles, J. et al. Evaluating the clinical validity of hypertrophic cardiomyopathy genes. *Circ. Genom. Precis Med* **12**, e002460 (2019).
33. Walsh, R. et al. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.* **19**, 192–203 (2017).
34. National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Women in Science, Engineering, and Medicine; Committee on Improving the Representation of Women and Underrepresented Minorities in Clinical Trials Research. *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups* (National Academies Press, 2022).
35. Ward, E. et al. Cancer disparities by race/ethnicity and socioeconomic status. *CA Cancer J. Clin.* **54**, 78–93 (2004).
36. Suther, S. & Kiros, G. E. Barriers to the use of genetic testing: a study of racial and ethnic disparities. *Genet. Med.* **11**, 655–662 (2009).
37. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
38. Vujkovic, M. et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691 (2020).

39. Graham, S. E. et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
40. Wall, J. D. et al. South Asian medical cohorts reveal strong founder effects and high rates of homozygosity. *Nat. Commun.* **14**, 3377 (2023).
41. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
42. Deaton, A. M. et al. Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *Sci. Rep.* **11**, 21565 (2021).
43. Välimäki, N. et al. Inherited mutations affecting the SRCAP complex are central in moderate-penetrance predisposition to uterine leiomyomas. *Am. J. Hum. Genet.* **110**, 460–474 (2023).
44. Haas, M. E. et al. Machine learning enables new insights into genetic contributions to liver fat accumulation. *Cell Genom.* **1**, 100066 (2021).
45. Khera, A. V. et al. Gene sequencing identifies perturbation in nitric oxide signaling as a nonlipid molecular subtype of coronary artery disease. *Circ. Genom. Precis. Med.* **15**, e003598 (2022).
46. Ward, J. et al. Genome-wide analysis in UK Biobank identifies four loci associated with mood instability and genetic correlation with major depressive disorder, anxiety disorder and schizophrenia. *Transl. Psychiatry* **7**, 1264 (2017).
47. Luciano, M. et al. Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* **50**, 6–11 (2018).
48. Nagel, M. et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **50**, 920–927 (2018).
49. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
50. Liu, F. R. et al. Pedigree-based study to identify GOLGB1 as a risk gene for bipolar disorder. *Transl. Psychiatry* **12**, 390 (2022).
51. Palmer, D. S. et al. Exome sequencing in bipolar disorder identifies *AKAP11* as a risk gene shared with schizophrenia. *Nat. Genet.* **54**, 541–547 (2022).
52. Cui, J. et al. Disruption of Gpr45 causes reduced hypothalamic POMC expression and obesity. *J. Clin. Invest.* **126**, 3192–3206 (2016).
53. Akbari, P. et al. Sequencing of 640,000 exomes identifies *GPR75* variants associated with protection from obesity. *Science* **373**, eabf8683 (2021).
54. Yamazaki, O., Hirohama, D., Ishizawa, K. & Shibata, S. Role of the ubiquitin proteasome system in the regulation of blood pressure: a review. *Int. J. Mol. Sci.* **21**, 5358 (2020).
55. Li, X. Y., Zhai, W. J. & Teng, C. B. Notch signaling in pancreatic development. *Int. J. Mol. Sci.* **17**, 48 (2015).
56. Horn, S. et al. Mind bomb 1 is required for pancreatic β-cell formation. *Proc. Natl Acad. Sci. USA* **109**, 7356–7361 (2012).
57. Potter, G. B., Facchinetti, F., Beaudoin, G. M. & Thompson, C. C. Neuronal expression of synaptotagmin-related gene 1 is regulated by thyroid hormone during cerebellar development. *J. Neurosci.* **21**, 4373–4380 (2001).
58. Moghadam, P. K. & Jackson, M. B. The functional significance of synaptotagmin diversity in neuroendocrine secretion. *Front Endocrinol. (Lausanne)* **4**, 124 (2013).
59. Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet* **99**, 76–88 (2016).
60. Galinsky, K. J. et al. Estimating cross-population genetic correlations of causal effect sizes. *Genet. Epidemiol.* **43**, 180–188 (2019).
61. Yengo, L. et al. A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
62. Hou, K. et al. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).
63. Ziyatdinov, A. et al. Genotyping, sequencing and analysis of 140,000 adults from the Mexico City Prospective Study. *Nature* **622**, 784–793 (2023).
64. Fatumo, S. & Inouye, M. African genomes hold the key to accurate genetic risk prediction. *Nat. Hum. Behav.* **7**, 295–296 (2023).

[1]Cardiovascular Disease Initiative, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Experimental Cardiology, Heart Center, Amsterdam Cardiovascular Sciences, Heart Failure and Arrhythmias, Amsterdam UMC location University of Amsterdam, Amsterdam, The Netherlands. [3]Cardiovascular Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [4]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [5]Division of Cardiology, University of California, San Francisco, CA, USA. [6]Metabolism Program, The Broad Institute of Harvard and MIT, Cambridge, MA, USA. [7]Program in Medical and Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, MA, USA. [8]Department of Psychiatry and Center for Genomic Medicine, Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital,Harvard Medical School, Boston, MA, USA. [9]Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [10]Institute for Molecular Medicine Finland (FIMM), Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland. [11]Department of Physiology, Amsterdam UMC location VU, Amsterdam, The Netherlands. [12]Department of Cardiology, University Heart and Vascular Center Hamburg-Eppendorf, Hamburg, Germany. [13]German Center for Cardiovascular Research (DZHK), Partner Site Hamburg/Kiel/Lübeck, Hamburg, Germany. [14]Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [15]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [16]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [17]Demoulas Center for Cardiac Arrhythmias, Massachusetts General Hospital, Boston, MA, USA. [18]NHLBI and Boston University's Framingham Heart Study, Framingham, MA, USA. [19]These authors contributed equally: Sean J. Jurgens, Xin Wang. ✉e-mail: ellinor@mgh.harvard.edu

Methods

Study datasets

In the present study we utilized three large biobanks with available sequencing data and linkage to EHRs.

UK Biobank. The UKB is a large population-based prospective cohort study from the UK that included over 500,000 individuals with deep phenotypic data, including medical interviews, EHR linkage and death registry linkage[65,66]. Participants were recruited between 2006 and 2010 at the ages of 40–69 years[66]. Relevant genomic data currently includes exome sequencing on over 450,000 samples funded through industry partnerships[1,67]. Exomes were captured using the revised version of the IDT xGen Exome Research Panel v.1.0 on Illumina NovaSeq 6000 machines (https://www.ukbiobank.ac.uk/media/najcnoaz/access_064-uk-biobank-exome-release-faq_v11-1_final-002.pdf). Alignment using BWA-MEM, calling using DeepVariant, and joint genotyping using GLNexus have been described in detail elsewhere (https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/UKB_WES_Protocol.pdf). In the present study, we utilized the OQFE exome call set and closely followed a previously published pipeline to perform stringent quality control (QC) of the exome sequencing data, including genotype QC, variant QC and sample QC[5]. Details on custom QC, principal component analysis, ancestry inference and relatedness inference are described in the Supplementary Note. After QC, we were left with 18,752,405 high-quality autosomal variants and 454,210 high-quality samples, of which 454,162 could be linked to their phenotypic data. The UKB resource was approved by the UKB Research Ethics Committee, and all participants provided written informed consent to participate. Use of UKB data was performed under application number 17488 and was approved by the Mass General Brigham Institutional Review Board.

All of Us. The AoU research program of the National Institutes of Health (NIH) is a longitudinal cohort study that aims to include 1 million racially, ancestrally and demographically diverse participants from across the USA, combining phenotypic data from various sources including patient-derived information and EHR linkage[68]. One of the goals set by AoU was to recruit individuals that have been, and continue to be, underrepresented in biomedical research because of limited access to healthcare[12,68]. Consistently, AoU prioritized underrepresented participants for genome sequencing and data collection and included them in the first few releases of the dataset, resulting in a diverse research population with rich phenotypic data. As part of the release in April 2023, WGS was performed on approximately 250,000 participants using Illumina NovaSeq 6000 machines following manufacturer's best practices. The same protocol for library preparation (PCR Free Kapa HyperPrep) and software for variant calling (DRAGEN v.3.4.12) were used to keep consistent WGS data generated from different AoU Genome Centers. A stringent central QC procedure was applied, as described in the program's genomic quality report (https://support.researchallofus.org/hc/en-us/articles/4617899955092-All-of-Us-Genomic-Quality-Report-), leaving 245,394 samples (47.7% described as racial/ethnic minorities). We performed further genotype, variant and sample QC procedures on the exome-region call set (contains variants that are within the exon regions of the Gencode v.42 basic transcripts, with padding of 15 bases on either side of each exon) released by the program, resulting in 242,902 eligible samples and 31,247,262 high-quality genetic variants. Details on the QC procedure, ancestry inference, principal component analysis and relatedness inference are described in the Supplementary Note. All enrolled participants provided informed consent to AoU. Use of AoU data was approved under a data use agreement between the Massachusetts General Hospital and the AoU research program.

Mass General Brigham Biobank. The MGB (formerly known as Partners Biobank) is an ongoing observational research project enrolling participants from a multicenter health system in eastern Massachusetts[69]. Participants are enrolled with broad-based consent collected by local research coordinators, either as part of a collaborative research study or electronically through a patient portal[70]. Demographic data, blood samples and surveys are collected at baseline and linked to EHR data. All adult patients provided informed consent to participate. A small number of children were enrolled with Institutional Review Board-approved assent forms; upon reaching 18 years of age all enrolled children had to provide consent or were removed from the study. The Human Research Committee of MGB approved the Biobank protocol (2009P002312). Exome sequencing has currently been completed for over 53,000 MGB participants, partly within the Centers for Common Disease Genomics initiative of the National Human Genome Research Institute and partly through industry partnership with IBM health. Samples were sequenced on Illumina NovaSeq machines with a custom exome panel (TWIST Human Core Exome), with a target of at least 20× coverage at >85% of target sites. Alignment, processing and joint calling of variants were performed using the Genome Analysis ToolKit (GATK v.4.1) following GATK best practices, after which we applied a stringent QC pipeline on the sequencing data (comparable with the pipeline applied in the UKB). Details on QC, ancestry inference, principal component analysis and relatedness inference are described in the Supplementary Note. After stringent QC, we were left with 12,421,458 autosomal genetic variants across 52,059 high-quality samples, of which 51,815 could be matched to their EHRs.

Ancestry definitions

In all analyses across all datasets, ancestry labels were based on inference from the genetic data. In all datasets, we defined labels for continental ancestries, namely EUR, EAS, SAS, AFR and AMR ancestries. Methods for genetic inference of ancestry differed between UKB and MGB, as compared with AoU. Methodology for ancestry inference is described in the Supplementary Note.

Phenotype construction

We defined a harmonized set of disease endpoints across the included datasets. To this end, we used the R package PheWAS (v.1.0, https://github.com/PheWAS/PheWAS) to create disease phecodes mapped from various ICD-10 billing codes[71]. We required at least one instance of an ICD code to define a sample as a case, while all other samples were considered controls for the given phecode. Prevalent and incident cases were pooled. In MGB and AoU, 1,866 and 1,835 phecodes could be mapped from ICD-10-CM code data, respectively, while in UKB available ICD-10 code data allowed mapping to 1,591 phecodes. In UKB, we manually curated a select number of traits, which had low case numbers in UKB due to absence of available ICD-10 codes (but had high case numbers in AoU/MGB; Supplementary Note). Given the high degree of correlation between various phecodes, we then utilized a clustering algorithm to identify important index phecodes[7]; we performed the clustering algorithm within the most phenotypically rich dataset, MGB. We first excluded any phecode with <50 cases in MGB (leaving 1,770 phecodes), which we then used to create a cosine similarity matrix and a cosine distance matrix (1 − similarity matrix). We used Ward's method to hierarchically cluster the cosine distance matrix, using a clustering tree height cutoff of 1.0 to define meaningful phecode clusters. We defined the index phecode as the phenotype with the highest case count within a cluster, utilizing the sum of case counts across UKB, a previous release of AoU (n = 98k) and MGB. Therefore, it is possible that a given index phecode is not present in each dataset; however, we keep the phecode yielding a high overall case number to increase statistical power for downstream genetic analysis. The clustering process left 519 index phecodes; we manually inspected the codes that were removed and pulled back 82 phecodes, leaving a final set of 601 largely independent phecodes for analysis. Of the 601 phecodes, 555, 601 and 601 codes were found in UKB, AoU and MGB, respectively, of which 546, 601 and 601 had at least 50 cases.

Variant annotation

In each dataset, variants were annotated using dbNSFP (v.4.2a for MGB and v.4.3a for UKB and AoU[72]) and the loss-of-function transcript effect estimator (LOFTEE[25]) plug-in implemented in the variant effect predictor (VEP; v.105)[73] (https://github.com/konradjk/loftee). VEP was used to ascertain the most severe consequence of a given variant for each gene. LOFTEE was implemented to identify high-confidence LOF variants, which include frameshift indels, stop-gain variants and splice site-disrupting variants. LOFs flagged by LOFTEE as dubious were removed. Missense variants were assigned a missense score representing the proportion of bioinformatics tools predicting a damaging effect, following previously published methods[5]. In short, we used information from 30 tools included in the dbNSFP database to score each missense variant by the number of tools predicting a damaging/deleterious effect, and divided this value by the number of tools that gave a prediction. Missense variants with fewer than seven predictions were removed. For instance, if 14 tools predicted a damaging effect and 28 total tools gave a prediction, then the missense score would equal 0.5 (14/28). Details on the contributing tools are provided in the Supplementary Note. Finally, variants were annotated with the highest continental allele frequency from gnomAD v.2 exomes (extracting frequencies for EUR, EAS, SAS, AFR and AMR superpopulations) denoted as 'gnomAD popmax'[25]. Within a dataset, the highest MAF between gnomAD popmax and the within-dataset MAF was designated the $MAF_{population\text{-}max}$.

Rare variant analyses

In each dataset, we performed exome-wide rare variant collapsing tests across the included disease phecodes with ≥50 cases. We assessed six rare variant masks in our main discovery analysis, namely:

(1) 'rare LOF' mask restricting to LOF variants with $MAF_{population\text{-}max} < 0.1\%$ (that is, MAF < 0.1% in the dataset and gnomAD popmax < 0.1%),
(2) 'rare LOF+missense0.8' mask including both LOF variants and predicted-deleterious missense variants with missense score > 0.8 and $MAF_{population\text{-}max} < 0.1\%$,
(3) 'rare LOF+missense0.5' mask including both LOF variants and predicted-deleterious missense variants with missense score > 0.5 and $MAF_{population\text{-}max} < 0.1\%$,
(4) 'ultrarare LOF+missense0.5' mask including both LOF variants and predicted-deleterious missense variants with missense score > 0.5 and $MAF_{population\text{-}max} < 0.001\%$ (for within-dataset filtering, we used MAC < 5 if more inclusive)
(5) 'ultrarare missense0.5' mask restricting to missense variants with missense score > 0.5 and $MAF_{population\text{-}max} < 0.001\%$ (for within-dataset filtering, we used MAC < 5 if more inclusive)
(6) 'ultrarare missense0.2' mask restricting to missense variants with missense score > 0.2 and $MAF_{population\text{-}max} < 0.001\%$ (for within-dataset filtering, we used MAC < 5 if more inclusive)

The stringent frequency cutoffs were chosen to limit results to very rare genetic variation in an attempt to enforce orthogonality to conventional common variant genome-wide association study results[1,5].

In secondary analyses, we also performed burden testing inclusive of low-frequency variants ($MAF_{population\text{-}max} < 1\%$; Fig. 1), for

(1) LOF variant mask ($MAF_{population\text{-}max} < 1\%$),
(2) LOF+missense0.8 mask ($MAF_{population\text{-}max} < 1\%$), and
(3) LOF+missense0.5 mask ($MAF_{population\text{-}max} < 1\%$)

For a given phenotype, rare variant masks were analyzed in a two-sided logistic mixed-effects score test using custom software (https://github.com/seanjosephjurgens/UKBB_200KWES_CVD/tree/v1.2), which is a previously described adaptation[5] of the R package GENESIS (v.2.18)[74]. Fixed effects included age, age^2, sex, sequencing batch (if applicable; Supplementary Note), ancestral principal components 1 to 4, and any other component among the first 5 to 20 components if associated with the phecode (nominal $P < 0.05$ among unrelated samples). In AoU, only the first 16 components were available. We accounted for relatedness by including a sparse kinship matrix as a random effect (Supplementary Note), and P values were computed using the saddle-point approximation (SPA) to account for case–control imbalance[75]. In cases where the mixed-effects model failed to converge, analyses were conducted using regular logistic regression among unrelated individuals. Missing genetic data were imputed to zero. For tests reaching nominal significance ($P < 0.05$), ORs, and s.e. were estimated using an approximate Firth's bias-reduced logistic regression[76,77] in the unrelated subset of each dataset.

Meta-analyses

To compute meta-statistics, we used a score-based meta-analysis approach. For each phenotype-mask test, we computed the $score_{meta}$ as the sum of study-specific score statistics, and the $score\ variance_{meta}$ as the sum of study-specific score variances[78]. To account for case–control imbalance in our meta-analysis, we recomputed the score variances in each dataset using the SPA P values before meta-analysis[79] (Supplementary Note). To prevent false positives driven by low minor allele count, we removed any tests with cumulative minor allele count (cMAC) < 10 in the study-specific results before meta-analysis. Because AoU does not allow extraction of summary statistics describing results from <20 individuals, the minimum number of alternative allele carriers for AoU was set to 20 before extraction of data from the AoU web portal. After meta-analysis, we removed any results with cMAC < 20. Therefore, our meta-analysis results include only tests with cMAC ≥ 20, where each contributing study has cMAC ≥ 10. Effect sizes for significant associations were estimating using an inverse-variance weighted meta-analysis of ORs and SEs.

Because we found that there was evidence of sample overlap and/or cryptic relatedness between AoU and MGB (median 0.05 test statistic correlation), we then applied an approach to correct the meta-analytical P values for this issue (Supplementary Note). In short, we used a weighted Z-score meta-analysis that (1) first estimates the spurious test statistic correlation across datasets, estimated separately for each phenotype (we found that correlations were approximately consistent across masks; Supplementary Fig. 13); and (2) then corrects the meta-analytical weights, accounting for the spurious correlation. While not perfect, we found that this approach yielded a substantially better calibration of meta-analytical test statistics. We note that this correction does not directly correct the effect size estimation, and therefore the variance of the effect sizes might be underestimated; nevertheless, we found that the corrected P values were reasonable for hypothesis testing.

To compute a single P value per gene–phecode pair, we used the Cauchy distribution to combine the mask-specific P values (from all six different masks) into a single omnibus P value. The Cauchy distribution allows for valid aggregation of several, potentially correlated, test statistics into a single test statistic[80]. A Benjamini–Hochberg FDR correction was then applied to these Cauchy P values to compute multiple testing-corrected Q-values, taking into account all gene–phecode pairs in one FDR correction. Q values < 0.01 were considered significant. All discovery analyses and meta-analyses considered all samples irrespective of ancestry. In sensitivity analyses, all discovery and meta-analyses were repeated restricting to samples determined genetically to be similar to EUR ancestry. We also analyzed matched synonymous variants, as described in the Supplementary Note.

Assessment of power benefits from diverse ancestries

To investigate the effect of ancestral diversity on discovery yield, we compared the number of identified associations at various significance cutoffs, when using all samples, and when using only samples of genetically determined EUR ancestry. We assessed the number of signals at Bonferroni-corrected significance ($P < 1 \times 10^{-7}$) and at standard

exome-wide significance ($P < 2.6 \times 10^{-6}$). To disentangle whether differences in number of discovered associations were due to the diminished sample size for the EUR-only analysis compared with the entire dataset, we then applied a downsampling approach. We downsampled the AoU dataset in such a way (removed samples) so the remaining sample size matched the sample size of the EUR subset of AoU. While doing so, we ensured the most ancestrally diverse composition of the downsampled dataset (Supplementary Note). As such, we created two equally sized subsets of AoU: one of exclusively EUR ancestry, and one highly diverse dataset. We performed exome-wide discovery analyses as described for our primary analysis. We then assessed the discovery yield—measured by number of significant associations at $P < 1 \times 10^{-7}$ and exome-wide significance—in both datasets. We further performed meta-analyses where we combined either dataset with UKB, to assess whether an ancestrally diverse dataset may improve yield when combined in meta-analysis with a large homogenous set (Supplementary Table 10).

Assessment of rare variant effect sizes across ancestry

We then aimed to assess whether rare variant effects estimated from genetically determined EUR samples were consistent in other ancestries. For this analysis, we considered rare LOF variant (MAF < 0.1%) masks and ultrarare missense0.5 (MAF < 0.001%) variant masks. We employed a three-sample design to avoid bias from Winner's curse. First, we identified suggestively significant mask–phecode associations from a EUR ancestry meta-analysis of UKB and MGB, defined as $P < 2.6 \times 10^{-6}$ among samples of genetically determined EUR ancestry. We then estimated effect sizes for those masks in various subsets of AoU. For instance, our main analysis focused on the effect sizes of those signals among EUR ancestry individuals in AoU, and the respective effect sizes among samples with a genetically defined ancestry dissimilar to EUR (AFR, AMR, EAS, SAS, admixed). Effect sizes and standard errors for both groups were estimated using Firth's regression among unrelated samples, requiring ≥20 rare variant carriers in both groups. In secondary analyses, we performed similar comparisons, this time comparing effect sizes from EUR ancestry samples with the respective effect sizes from two defined ancestry groups with sufficient sample size in AoU, namely AFR and AMR ancestries.

To compare phenome-wide effect sizes between different groups, we computed Pearson correlation estimates, quantifying the correlation between rare variant effect sizes from EUR samples against the respective effect sizes in non-EUR samples. Because effect sizes from our analysis are estimated with error (large s.e. given low numbers of carriers) this can downward bias correlation and regression estimates, a phenomenon known as attenuation bias[81]. Given known s.e. of our estimates, we also computed disattenuated correlation coefficients providing upper bound estimates of the possible true correlations between EUR effect sizes and non-EUR effect sizes (Supplementary Note). We then aimed to build regression models quantifying the relationship between EUR and non-EUR effect sizes. To incorporate the error in effect estimates, we used Deming regression[62,82]—a form of error-in-variables total-least-squares regression—to regress non-EUR effect sizes on EUR effect sizes (using function deming() in R package deming v.1.4). Since the s.e. values for each beta coefficient were known, these were fed directly into the regression model. As such, the assumption of equal error ratios was relaxed, making the regressions comparable with York regression with the assumption of uncorrelated errors. Regression weights were applied to account for potential heteroscedasticity, and s.e. was computed using Jackknife estimators for all regressions including more than eight datapoints.

In sensitivity analyses, we removed genes associated with leukemic outcomes and/or age, to assess potential effects from somatic variation on phenome-wide effect size correlations. We also performed analyses accounting for bins of effective sample size, to better account for differential discovery power across different phenotypes (Supplementary Table 14); in these analyses, we performed Deming regression within quantiles determined by the effective sample size (computed within EUR ancestry samples), and then performed a random-effects IVW meta-analysis to combine the results from the quantiles. For LOF variants, we finally performed analyses using phenotype-specific effect size correlations. To this end, we used phenotypes with at least three qualifying genes and performed Deming regression for each phenotype separately. We then used the IVW approach to combine the phenotype-specific Deming coefficients.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Results from our gene-based association analyses are available for browsing and download through our online portal (https://hugeamp.org:8000/research.html?pageid=600_traits_app_home). Bulk download of summary statistics is possible via the Cardiovascular Disease Knowledge Portal (https://cvd.hugeamp.org/downloads.html). Access to individual-level UKB data, both phenotypic and genetic, is available to bona fide researchers through application on the UKB website (https://www.ukbiobank.ac.uk). The final release of the exome sequencing dataset of UKB is available only through the DNAnexus Research Analysis Platform (https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform). Additional information about registration for access to the data is available at http://www.ukbiobank.ac.uk/register-apply/. Use of UKB data was performed under application number 17488. Access to individual phenotypic and genetic data from AoU is currently available to bona fide researchers within the USA through the AoU Researcher Workbench, a cloud-based computing platform (https://www.researchallofus.org/register/). A publicly available data browser is provided by the research program (https://databrowser.researchallofus.org/). Access to individual-level data for participants from the MGB is currently not publicly available. Other datasets used in this manuscript include: the dbNSFP database v.4.2a and v.4.3a (https://sites.google.com/site/jpopgen/dbNSFP); gnomAD exomes v.2.1 (https://gnomad.broadinstitute.org/downloads); the OMIM database (omim.org) accessed on 25 August 2022; Ensembl release 105 (https://www.ensembl.org/info/data/index.html); and the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) accessed in December 2022.

Code availability

QC of individual-level data was performed using Hail v.0.2 (https://hail.is) as well as PLINK v.2.0.a (https://www.cog-genomics.org/plink/2.0/). Variant annotation was performed using VEP v.105 (https://github.com/Ensembl/ensembl-vep). Main rare variant association analyses were performed using an adaptation of the R package GENESIS v.2.18 (https://rdrr.io/bioc/GENESIS/man/GENESIS-package.html), which has previously been made available by us through the GitHub repository https://github.com/seanjosephjurgens/UKBB_200KWES_CVD / v.1.2 (https://doi.org/10.5281/zenodo.11638262). Meta-analyses were performed using custom code available in the same repository, and using METAL (2017-12-21 release). Analyses that were run in R, were run within R v.4 (https://www.r-project.org).

References

65. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
66. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
67. Szustakowski, J. D. et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).

68. Cronin, R. M. et al. Development of the initial surveys for the All of Us Research Program. *Epidemiology* **30**, 597–608 (2019).
69. Karlson, E. W., Boutin, N. T., Hoffnagle, A. G. & Allen, N. L. Building the Partners HealthCare Biobank at Partners Personalized Medicine: informed consent, return of research results, recruitment lessons and operational considerations. *J. Pers. Med.* **6**, 2 (2016).
70. Boutin, N. T. et al. Implementation of electronic consent at a biobank: an opportunity for precision medicine research. *J. Pers. Med.* **6**, 17 (2016).
71. Wu, P. et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inf.* **7**, e14325 (2019).
72. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* **37**, 235–241 (2016).
73. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
74. Gogarten, S. M. et al. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346–5348 (2019).
75. 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).
76. Heinze, G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat. Med.* **25**, 4216–4226 (2006).
77. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
78. Tang, Z. Z. & Lin, D. Y. MASS: meta-analysis of score statistics for sequencing studies. *Bioinformatics* **29**, 1803–1805 (2013).
79. Zhao, Z. et al. UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. *Am. J. Hum. Genet.* **106**, 3–12 (2020).
80. Liu, Y. et al. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* **104**, 410–421 (2019).
81. Muchinsky, P. M. The correction for attenuation.*Educ. Psychol. Meas.* **56**, 63–75 (1996).
82. Deming, W. E. *Statistical Adjustment of Data* (Wiley, 1943).

Acknowledgements

We gratefully thank all participants of UKB, AoU and MGB Biobank, as this study would not have been possible without their contributions. We also thank the NIH's AoU Research Program, the UKB resource (under application number 17488) and the MGB team, for making available the participant data examined in this study. P.T.E. was supported by funding from the NIH (1RO1HL092577, 1R01HL157635), by a grant from the American Heart Association (18SFRN34110082, 961045) and from the European Union (MAESTRIA 965286). This work was also supported by an American Heart Association Strategically Focused Research Networks (SFRN) postdoctoral fellowship (18SFRN34110082) to L.-C.W. This work was supported by the John S. LaDue Memorial Fellowship for Cardiovascular Research, a Sarnoff Scholar award from the Sarnoff Cardiovascular Research Foundation and by a NIH grant (K08HL159346) to J.P.P. This work was further supported by a grant from the NIH (1K08HL153937) and a grant from the American Heart Association (862032) to K.G.A. This work was supported by a Sigrid Jusélius Fellowship to J.T.R. This work was also supported by an Amsterdam UMC doctoral fellowship and the Junior Clinical Scientist Fellowship (03-007-2022-0035) from the Dutch Heart Foundation, to S.J.J. This work was supported by the BioData Ecosystem fellowship to S.H.C. This work was also supported by a grant from the NIH (R01DK125490) to J.F.

Author contributions

S.J.J. and P.T.E. conceived and designed the study. S.J.J., X.W., S.H.C., L.-C.W., S. Koyama and J.P.P. performed data curation and data processing. S.J.J. and X.W. performed the main statistical and bioinformatic analyses, with S.H.C. providing important bioinformatic support. M.C., R.W., C.R., K.J.B., S. Kany, A.L.E., L.F.J.M.W. and J.T.R. contributed critically to the analysis plan. P.N., K.G.A., C.R.B., S.A.L., K.L.L., and P.T.E. supervised the study. T.N., P.S. and D.J. created the online web portal on the Human Disease Knowledge Portal. J.F. and N.P.B. supervised the creation of the online web portal. S.J.J., X.W. and P.T.E. wrote the manuscript. All authors critically revised and approved the manuscript.

Competing interests

P.T.E. has received sponsored research support from Bayer AG, Bristol Myers Squibb and Pfizer and Novo Nordisk. S.A.L. is an employee of Novartis as of July 2022. S.A.L. previously received sponsored research support from Bristol Myers Squibb, Pfizer, Boehringer Ingelheim, Fitbit, Medtronic, Premier and IBM, and has consulted for Bristol Myers Squibb, Pfizer, Blackstone Life Sciences and Invitae. P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific, Genentech/Roche and Novartis, personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Creative Education Concepts, CRISPR Therapeutics, Eli Lilly & Co, Esperion Therapeutics, Foresite Labs, Genentech/Roche, GV, HeartFlow, Magnet Biomedicine, Merck, Novartis, TenSixteen Bio and Tourmaline Bio, equity in Bolt, Candela, Mercury, MyOme, Parameter Health, Preciseli and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work. B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. S. Kany is supported by the Walter Benjamin Fellowship from the Deutsche Forschungsgemeinschaft (521832260). The remaining authors declare no competing interests.

Additional information

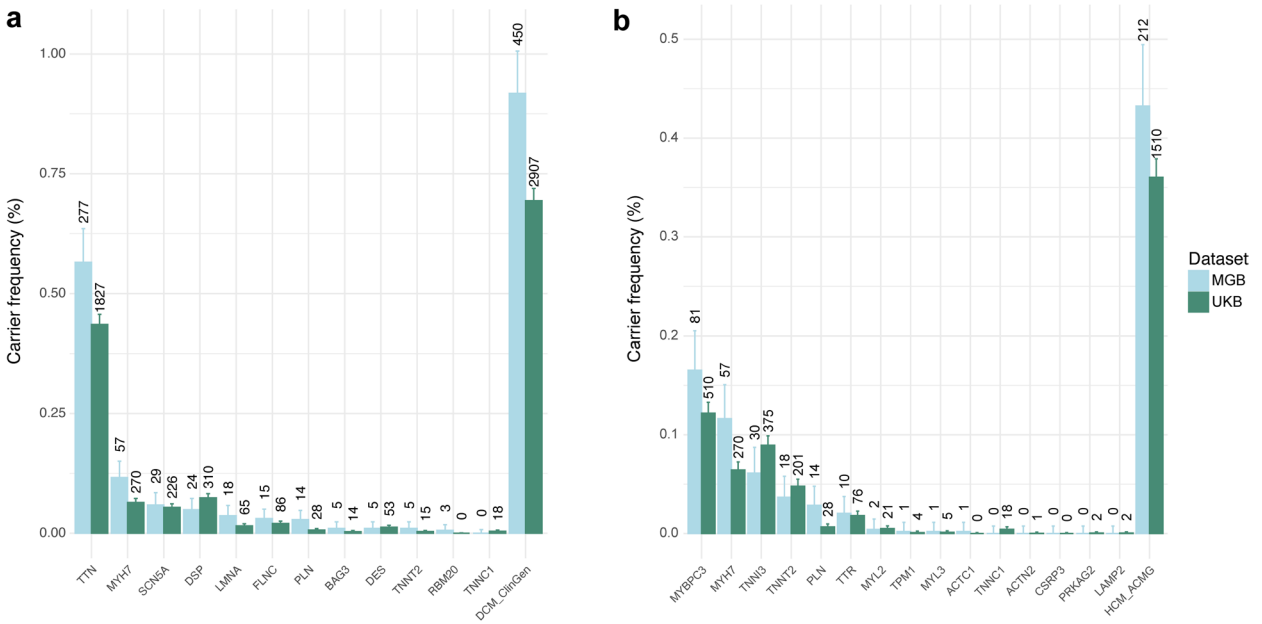
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01894-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01894-5.

Correspondence and requests for materials should be addressed to Patrick T. Ellinor.

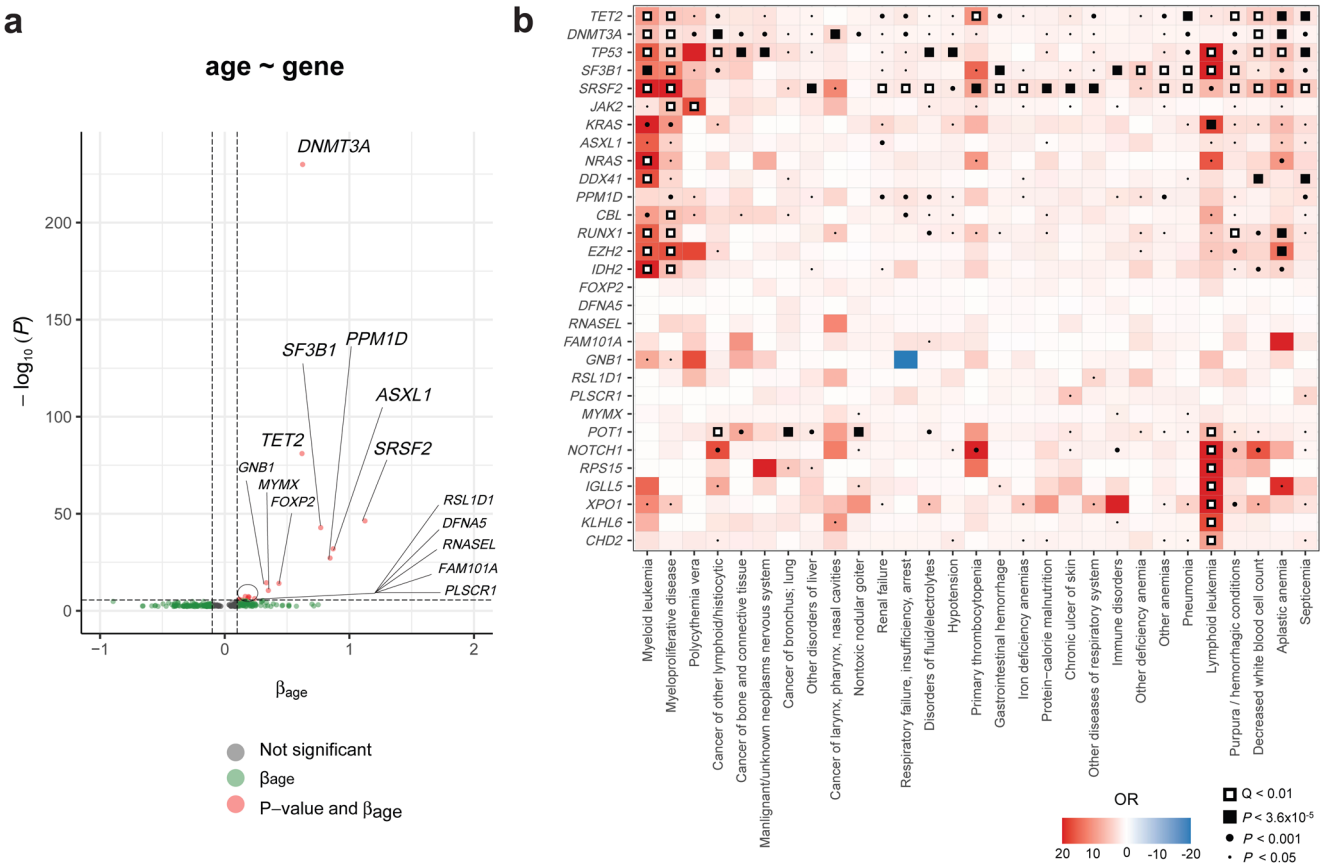
Peer review information *Nature Genetics* thanks Benjamin Sun, Seunggeun Lee and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-01894-5



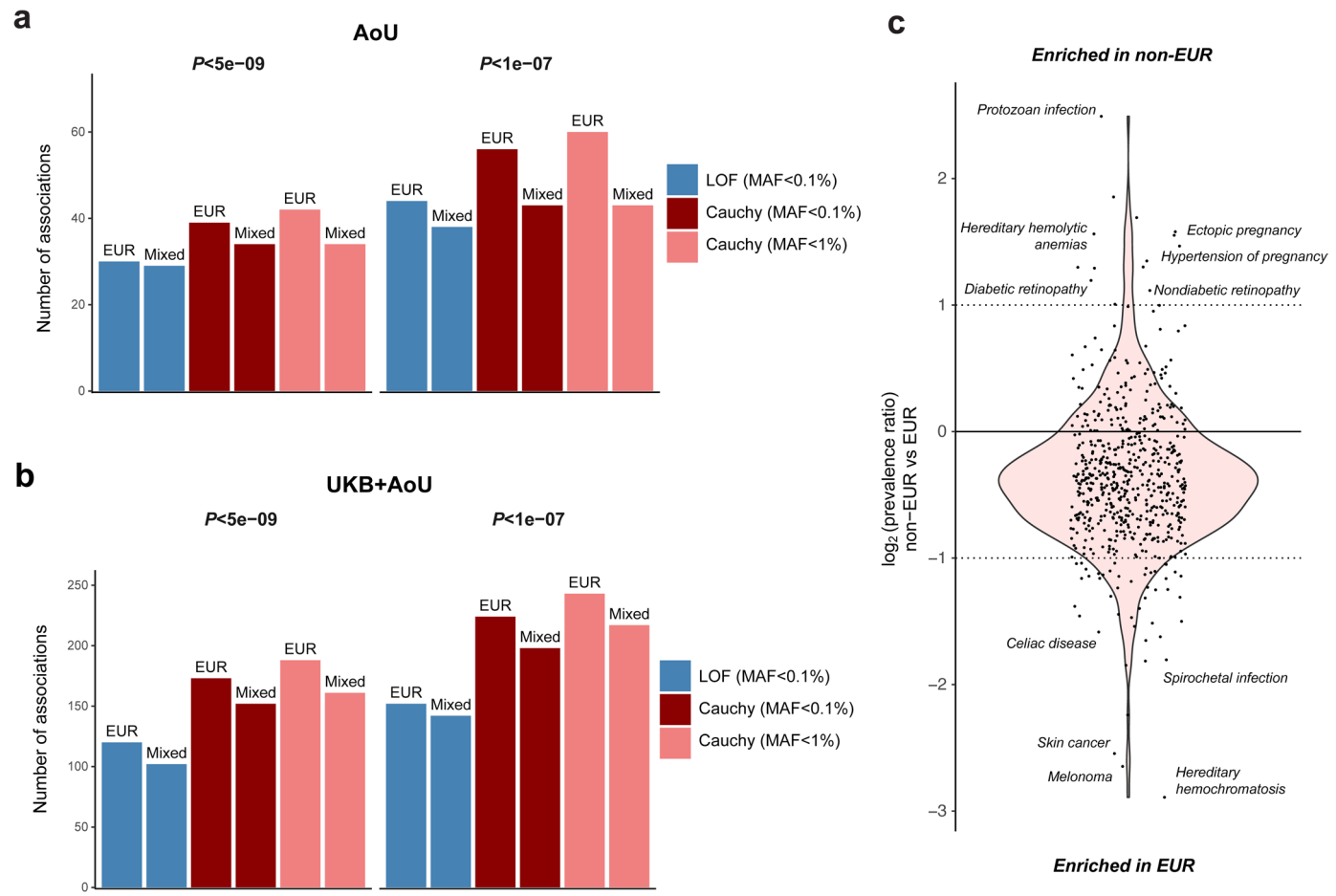
Extended Data Fig. 1 | Prevalence of rare likely pathogenic and pathogenic variants in cardiomyopathy genes across UKB and MGB datasets. a,**b**, Bar graphs reflect the percentage of biobank participants found to carry pathogenic or likely pathogenic variants for dilated cardiomyopathy (**a**) or hypertrophic cardiomyopathy (**b**) with 95% binomial confidence intervals, where light blue represent Massachusetts Biobank (MGB) and dark green represents UK Biobank (UKB). The absolute number of carriers identified in a given grouping is added above each bar. The total number of participants was $n = 51,815$ in MGB and $n = 454,162$ in UKB. Pathogenic or likely pathogenic variants reported in ClinVar and submitted by clinical testing labs from 2015 onwards were included, as well as high-confidence LOF variants affecting canonical transcripts for select genes (where truncation is considered pathogenic for the disease); for *TTN*, only variants affecting the cardiac exons were included. Variants were filtered to MAF < 0.1%. The combined prevalences (all genes combined) are shown on the far right of the panels. Overall, rare disease-causing variants for both disorders were more frequent in MGB vs. UKB (non-overlapping 95% binomial confidence intervals).

https://doi.org/10.1038/s41588-024-01894-5



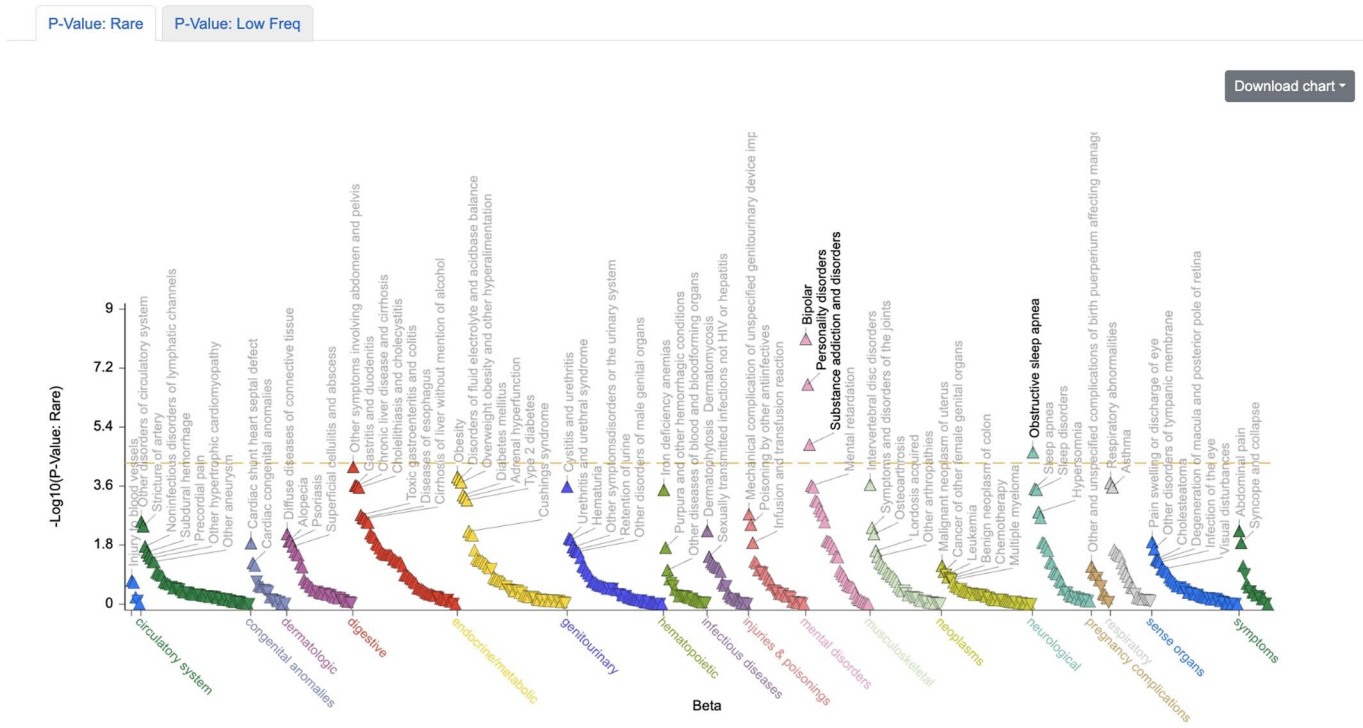
Extended Data Fig. 2 | Evidence of age-related somaticism and phenotype associations for potentially somatic gene variants. a, Volcano plot with results from linear regression models predicting age from rare variant carrier status (in a meta-analysis of UKB, AoU and MGB), with the -$\log_{10}(P_{Cauchy})$ on the *y*-axis and the estimated effect size per year for the most significant mask on the *x*-axis (β_{age}). The same pipeline was used as for our primary analysis. The horizontal dotted line shows a suggestive significance cutoff of $P < 1 \times 10^{-6}$, while the vertical lines highlight $\beta_{age} = -0.1$ and $\beta_{age} = 0.1$, respectively. Significant genes ($P < 1 \times 10^{-6}$) are annotated with their gene names; all genes that were significantly associated with any outcome in our primary analysis (Supplementary Table 16) and with age are also annotated with their gene names. Gene masks reaching $P < 1 \times 10^{-6}$ and $\beta_{age} > 0.1$ can be considered suggestively affected by age and therefore raise suspicion that they are affected by age-related somatic variants. Indeed, many of these genes are known clonal hematopoiesis of indeterminate potential (CHIP) genes (Supplementary Note). **b**, Heatmap with Phecodes on the *x*-axis and genes on the *y*-axis. The heatmap shows results for genes that reached significance in our PheWAS for leukemic/hematological outcomes and/or genes associated with age. These genes are plotted against a range of representative phenotypes suggestively associated at $P < 1 \times 10^{-5}$ with any of the genes. The color in each cell represents the odds ratio (OR) for the most significant mask (lowest nominal *P*-value) with red indicating increased disease risk (OR > 1) and blue indicating decreased disease risk (OR < 1). Significance levels are shown in each cell using circles and boxes, with a small dot representing nominal $P_{Cauchy} < 0.05$, a larger dot representing $P_{Cauchy} < 0.001$, a black box representing $P_{Cauchy} = 1 \times 10^{-5}$ and a black box with smaller white box representing $Q_{Cauchy} < 0.01$ in our primary analysis. ORs were estimated using inverse-variance weighted meta-analysis of two-sided Firth's logistic regression results. The reported *P*-values are Cauchy *P*-values that represent the omnibus *P*-value of all masks for a gene-phecode pair after combining them using the Cauchy distribution (unadjusted for multiple testing). *Q*-values represent the Benjamini-Hochberg FDR adjustments of these *P*-values. *P*-values for mask-phecode pairs (prior to the Cauchy combination) were derived from a *Z* score-based meta-analysis of score tests from logistic mixed-effects models with saddle-point-approximation. All statistical tests and *P*-values are two-sided. LOF, loss-of-function; OR, odds ratio.

https://doi.org/10.1038/s41588-024-01894-5



Extended Data Fig. 3 | No apparent discovery benefit in rare variant burden testing from ancestral diversity at current case numbers in AoU. a,**b**, Grouped barcharts with the number of significant signals identified from rare variant burden testing on the *y*-axis, comparing results from two different sub-samples of datasets. Blue bars represent results for LOF variants only ($MAF_{population\text{-}max}$ < 0.1%), while dark red bars show results for the Cauchy combination of 6 masks and light red bars represent a Cauchy combination of 9 masks (including 3 low-frequency variant masks at $MAF_{population\text{-}max}$ < 1%). Plot in **a** shows results for the subset of AoU consisting of individuals genetically similar to European ancestry (*n* = 106,057 samples with complete EHR linkage; EUR) on the left side of each comparison, while the right side shows results for an ancestrally diverse sub-sample of AoU of equal size (*n* = 106,057 samples with complete EHR linkage; Mixed). Results are restricted to 584 phecodes that were testable in both subsamples. Plot in **b** shows those same sub-samples of AoU in a meta-analysis with UKB, restricting to 530 phecodes that were testable across AoU subsamples and in the UKB dataset. **c**, Violin plot showing prevalence ratios for all 601 phecodes in AoU, where the prevalence ratios represent the ratio between prevalence among EUR samples and within individuals genetically dissimilar to European ancestry (non-EUR). Prevalence ratios are presented on the log_2-scale, where one unit difference represents a doubling/halving of the relative prevalence. The black line represents prevalence ratio of 0 (no difference), while the dotted lines represent prevalence ratios of 1 and −1. Select phenotypes enriched on either side are annotated. Many phecodes are relatively enriched in EUR as compared to non-EUR, which might contribute to the slightly diminished discovery yield within the ancestrally diverse subsample of AoU as compared to the EUR subsample. The Cauchy *P*-values represent the omnibus *P*-value of all relevant masks for a gene-phecode pair after combining them using the Cauchy distribution (unadjusted for multiple testing). *P*-values for mask-phecode pairs (prior to the Cauchy combination) in AoU were derived from a saddle-point-approximation score tests from logistic mixed-effects models, while meta-analysis *P*-values were derived from *Z* score-based meta-analysis of such score tests. All statistical tests and *P*-values are two-sided.

https://doi.org/10.1038/s41588-024-01894-5



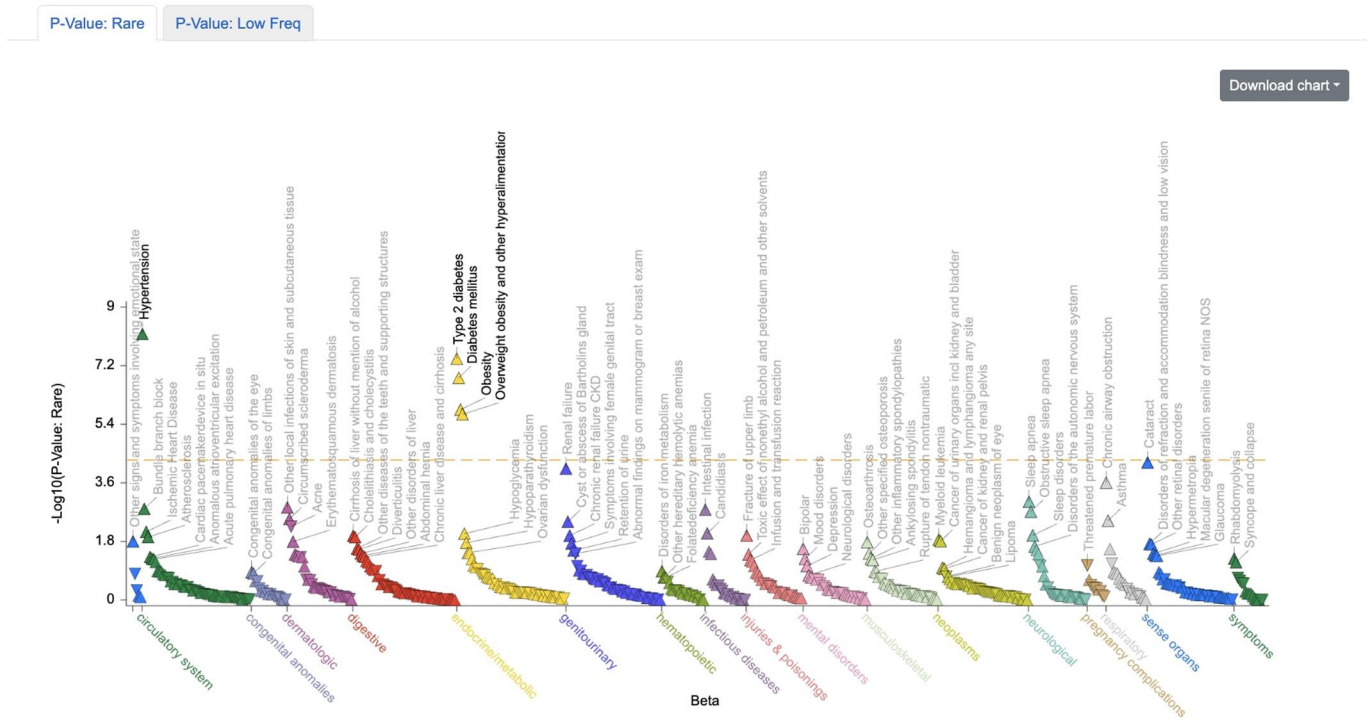
Total rows: 601

Save data / set table

Phenotype	Category	Meaning	P-Value: Rare	P-Value: Low Freq	Beta	Evidence
phecode_296.1	mental disorders	Bipolar	8.14e-9	1.21e-8	▲1.301	View
phecode_301.0	mental disorders	Personality disorders	2.00e-7	2.70e-7	▲1.612	View
phecode_316.0	mental disorders	Substance addiction and disorders	0.0000136	0.00002	▲0.931	View
phecode_327.32	neurological	Obstructive sleep apnea	0.0000233	0.0000176	▲1.017	View
phecode_579.0	digestive	Other symptoms involving abdomen and pelvis	0.0000652	0.0000893	▲0.941	View
phecode_278.1	endocrine/metabolic	Obesity	0.0001329	0.000165	▲0.591	View
phecode_276.0	endocrine/metabolic	Disorders of fluid electrolyte and acidbase balance	0.000158	0.0001199	▲0.815	View
phecode_278.0	endocrine/metabolic	Overweight obesity and other hyperalimentation	0.0001948	0.0002629	▲0.571	View
phecode_513.0	respiratory	Respiratory abnormalities	0.0002016	0.0002567	▲1.253	View
phecode_722.0	musculoskeletal	Intervertebral disc disorders	0.00023	0.0003211	▲0.612	View

Extended Data Fig. 4 | Broad Human Disease Knowledge Portal showing phenome-wide results for *YLPM1*. Output from a search for the gene *YLPM1* on the Broad Human Disease Knowledge Portal, which showcases the results from our primary meta-analysis of UKB, AoU and MGB. The top of the figure shows a dot plot with each dot representing a different phecode tested for association with *YLPM1*, where the *y*-axis shows the -$\log_{10}$(Cauchy *P*-value) and the *x*-axis represents different phenotypes grouped by broad phecode category. The arrows represent directionality, with an upwards arrow indicating that rare variants in *YLPM1* are associated with increased risk of the given phecode, and downwards arrows representing decreased risk; directionality is based in the 'Best Mask' which is the mask that yielded the lowest nominal *P*-value in burden testing. The dotted line represents the significance level used for phenome-wide testing of a single gene on the portal ($\alpha = 5 \times 10^{-5}$). Phenotypes reaching this level of significance are highlighted in black text. The bottom of the figure shows the associated results table as presented on the portal, including details on the most strongly associated phecodes, the Cauchy *P*-values for burden testing of rare variant masks, the Cauchy *P*-values for burden testing of rare and low-frequency masks, and the beta coefficient of the 'Best Mask' (ie the mask that reached the lowest nominal *P*-value in burden testing). Results can be queried through the following link: https://hugeamp.org:8000/research.html?ancestry=mixed&cohort=UKB_450k_AoU_250k_MGB_53k_META_overlapcorrected&file=600Traits.csv&gene=YLPM1&pageid=600_traits_app. Betas, which represent log(odds ratios), were estimated using inverse-variance weighted meta-analysis of two-sided Firth's logistic regression results. The reported *P*-values are Cauchy *P*-values that represent the omnibus *P*-value of all masks for a gene-phecode pair after combining them using the Cauchy distribution (unadjusted for multiple testing). *P*-values for mask-phecode pairs (prior to the Cauchy combination) were derived from *Z*-score-based meta-analysis of score tests from logistic mixed-effects models. All statistical tests and *P*-values are two-sided. LOF, loss-of-function; OR, odds ratio.

https://doi.org/10.1038/s41588-024-01894-5



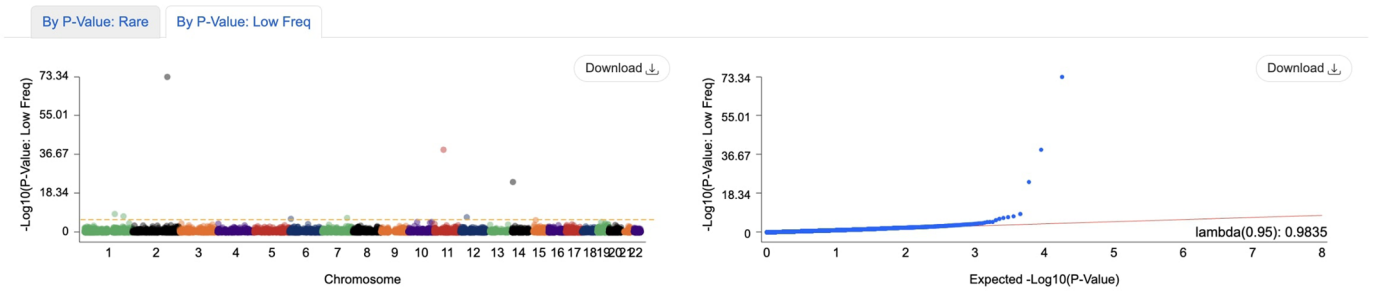
Total rows: 601

Save data / set table

Phenotype	Category	Meaning	P-Value: Rare	P-Value: Low Freq	Beta	Evidence
phecode_401.0	circulatory system	Hypertension	6.69e-9	5.42e-9	▲1.075	View
phecode_250.2	endocrine/metabolic	Type 2 diabetes	3.84e-8	2.88e-8	▲1.288	View
phecode_250.0	endocrine/metabolic	Diabetes mellitus	1.49e-7	1.12e-7	▲1.226	View
phecode_278.1	endocrine/metabolic	Obesity	1.41e-6	1.06e-6	▲0.985	View
phecode_278.0	endocrine/metabolic	Overweight obesity and other hyperalimentation	1.98e-6	1.48e-6	▲0.972	View
phecode_366.0	sense organs	Cataract	0.0000593	0.0000446	▲0.961	View
phecode_585.0	genitourinary	Renal failure	0.000092	0.0000691	▲1.293	View
phecode_496.0	respiratory	Chronic airway obstruction	0.0002478	0.0003527	▲0.613	View
phecode_327.3	neurological	Sleep apnea	0.0009586	0.0014338	▲0.327	View
phecode_686.0	dermatologic	Other local infections of skin and subcutaneous tissue	0.0014265	0.0010728	▲1.467	View

Extended Data Fig. 5 | Broad Human Disease Knowledge Portal showing phenome-wide results for *UBR3*. Output from a search for the gene *UBR3* on the Broad Human Disease Knowledge Portal, which showcases the results from our primary meta-analysis of UKB, AoU and MGB. The top of the figure shows a dot plot with each dot representing a different phecode, where the *y*-axis shows the $-\log_{10}$(Cauchy *P*-value) and the *x*-axis represents different phenotypes grouped by broad phecode category. The arrows represent directionality, with an upwards arrow indicating that rare variants in *UBR3* are associated with increased risk of the given phecode, and downwards arrows representing decreased risk; directionality is based in the 'Best Mask' which is the mask that yielded the lowest nominal *P*-value in burden testing. The dotted line represents the significance level used for phenome-wide testing of a single gene on the portal ($\alpha = 5 \times 10^{-5}$). Phenotypes reaching this level of significance are highlighted in black text. The bottom of the figure shows the associated results table as presented on the portal, including details on the most strongly associated phecodes, the Cauchy *P*-values for burden testing of rare variant masks, the Cauchy *P*-values for burden testing of rare and low-frequency masks, and the beta coefficient of the 'Best Mask' (ie the mask that reached the lowest nominal *P*-value in burden testing). Results can be queried through the following link: https://hugeamp.org:8000/research.html?ancestry=mixed&cohort=UKB_450k_AoU_250k_MGB_53k_META_overlapcorrected&file=600Traits.csv&gene=UBR3&pageid=600_traits_app. Betas, which represent log(odds ratios), were estimated using inverse-variance weighted meta-analysis of two-sided Firth's logistic regression results. The reported *P*-values are Cauchy *P*-values that represent the omnibus *P*-value of all masks for a gene-phecode pair after combining them using the Cauchy distribution (unadjusted for multiple testing). *P*-values for mask-phecode pairs (prior to the Cauchy combination) were derived from *Z*-score-based meta-analysis of score tests from logistic mixed-effects models. All statistical tests and *P*-values are two-sided. LOF, loss-of-function; OR, odds ratio.

https://doi.org/10.1038/s41588-024-01894-5



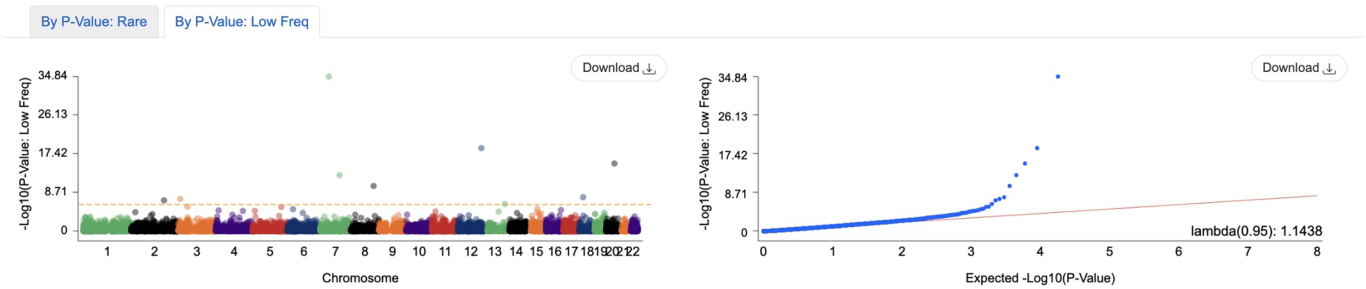
Total rows: 17971

Save data / set table

Gene	Region	P-Value: Rare	P-Value: Low Freq	P-Value: Best Mask	Beta	Best Mask	Type	Cases	Controls
TTN	2:179390716-179695529	3.04e-74	4.56e-74	5.07e-75	▲1.223	LOF (MAF<0.1%)	protein_coding	14691	734188
MYBPC3	11:47352957-47374253	1.72e-39	1.09e-39	2.11e-40	▲2.333	LOF (MAF<1%)	protein_coding	14691	734188
MYH7	14:23881949-23904869	1.21e-24	1.81e-24	2.62e-25	▲1.13	missense0.2 (MAF<0.001%)	protein_coding	14691	734188
LMNA	1:156052364-156109872	1.53e-9	2.30e-9	2.63e-10	▲1.29	LOF+missense0.5 (MAF<0.001%)	protein_coding	14691	734188
TNNT2	1:201328136-201346892	3.89e-8	2.90e-8	6.99e-9	▲0.877	LOF+missense0.5 (MAF<1%)	protein_coding	14691	734188
PKP2	12:32943679-33049711	3.66e-7	7.36e-8	1.63e-8	▲1.103	LOF+missense0.8 (MAF<1%)	protein_coding	14691	734188
FLNC	7:128470460-128499328	2.57e-7	1.66e-7	3.23e-8	▲1.622	LOF (MAF<1%)	protein_coding	14691	734188
DSP	6:7541850-7586947	6.10e-7	4.57e-7	1.03e-7	▲1.592	LOF (MAF<0.1%)	protein_coding	14691	734188
ACTC1	15:35082431-35087750	2.00e-6	1.93e-6	6.00e-7	▲1.552	LOF+missense0.5 (MAF<1%)	protein_coding	14691	734188

Extended Data Fig. 6 | Broad Human Disease Knowledge Portal showing exome-wide results for the phecode '*Cardiomyopathy*'. Output from a search for the phecode *Cardiomyopathy* on the Broad Human Disease Knowledge Portal, which showcases the results from our primary meta-analysis of UKB, AoU and MGB. A Manhattan plot is shown in the top left, with each dot representing a different gene tested for association with *Cardiomyopathy*, where the *y*-axis shows the -$\log_{10}$(Cauchy *P*-value) and the *x*-axis represents genomic coordinates. In this figure, results are restricted to 'rare variant' masks only (MAF < 0.1%). The dotted line represents the significance threshold used for a single phenotype on the portal ($\alpha = 1 \times 10^{-6}$). A quantile-quantile plot in the top right shows the observed genome-wide test statistics on the *y*-axis, against the expected test statistics under the null hypothesis on the *x*-axis; the red line represents the $x = y$ line. The bottom of the figure shows the associated results table as presented on the portal, including details on the most strongly associated genes, the Cauchy *P*-values for burden testing of rare variant masks, the Cauchy *P*-values for burden testing of rare and low-frequency masks, the *P*-value and beta coefficient for the 'Best Mask' (that is the mask with the lowest nominal *P*-value in burden testing), and information on case/control numbers. The table here is restricted to 9 genes with at least suggestive evidence ($P < 3 \times 10^{-6}$). Results can be queried through the following link: https://hugeamp.org:8000/research.html?ancestry=mixed&cohort=UKB_450k_AoU_250k_MGB_53k_META_overlapcorrected&file=600Traits.csv&pageid=600_traits_app&phenotype=phecode_425.0. Betas, which represent log(odds ratios), were estimated using inverse-variance weighted meta-analysis of two-sided Firth's logistic regression results. The reported *P*-values for 'rare' and 'low-freq' represent omnibus *P*-values of all relevant masks for a gene-phecode pair after combining them using the Cauchy distribution (unadjusted for multiple testing). *P*-values for mask-phecode pairs (prior to the Cauchy combination) were derived from *Z*-score-based meta-analysis of score tests from logistic mixed-effects models with saddle-point-approximation. All statistical tests and *P*-values are two-sided. LOF, loss-of-function; OR, odds ratio.

https://doi.org/10.1038/s41588-024-01894-5



Total rows: 17971

Save data / set table

Gene	Region	P-Value: Rare	P-Value: Low Freq	P-Value: Best Mask	Beta	Best Mask	Type	Cases	Controls
GCK	7:44182812-44237769	3.68e-35	1.46e-35	2.21e-36	▲1.64	LOF+missense0.8 (MAF<1%)	protein_coding	89991	658888
HNF1A	12:121416346-121440315	2.55e-19	1.91e-19	4.25e-20	▲2.628	LOF (MAF<0.1%)	protein_coding	89991	658888
HNF4A	20:42984339-43061485	3.33e-16	5.55e-16	5.99e-17	▲0.859	LOF+missense0.8 (MAF<0.1%)	protein_coding	89991	658888
GIGYF1	7:100277130-100291873	3.13e-13	2.34e-13	5.87e-14	▲1.143	LOF+missense0.8 (MAF<1%)	protein_coding	89991	658888
SLC30A8	8:117962512-118188953	5.44e-6	6.40e-11	7.11e-12	▼-0.402	LOF+missense0.5 (MAF<1%)	protein_coding	89991	658888
MIB1	18:19284918-19450914	2.82e-8	2.23e-8	4.72e-9	▲0.304	LOF+missense0.5 (MAF<0.1%)	protein_coding	89991	658888
PPARG	3:12328867-12475855	8.88e-8	5.35e-8	1.35e-8	▲1.152	LOF+missense0.8 (MAF<1%)	protein_coding	89991	658888
UBR3	2:170683964-170940641	1.49e-7	1.12e-7	3.21e-8	▲1.226	LOF (MAF<0.1%)	protein_coding	89991	658888
IRS2	13:110405042-110438930	9.90e-7	7.42e-7	2.54e-7	▲1.292	LOF (MAF<0.1%)	protein_coding	89991	658888
BSN	3:49591910-49708982	4.87e-6	2.81e-6	9.38e-7	▲1.029	LOF (MAF<1%)	protein_coding	89991	658888

Extended Data Fig. 7 | Broad Human Disease Knowledge Portal showing exome-wide results for the phecode '*Diabetes Mellitus*'. Output from a search for the phecode *Diabetes Mellitus* on the Broad Human Disease Knowledge Portal, which showcases the results from our primary meta-analysis of UKB, AoU and MGB. A Manhattan plot is shown in the top left, with each dot representing a different gene tested for association with *Diabetes Mellitus*, where the *y*-axis shows the -$\log_{10}$(Cauchy *P*-value) and the *x*-axis represents genomic coordinates. In this figure, results include both 'rare variant' and 'low-frequency' masks (MAF < 1%). The dotted line represents the significance threshold used for a single phenotype on the portal ($\alpha = 1 \times 10^{-6}$). A quantile-quantile plot in the top right shows the observed genome-wide test statistics on the *y*-axis, against the expected test statistics under the null hypothesis on the x-axis; the red line represents the $x = y$ line. The bottom of the figure shows the associated results table as presented on the portal, including details on the most strongly associated genes, the Cauchy *P*-values for burden testing of rare variant masks, the Cauchy *P*-values for burden testing of rare and low-frequency masks, the *P*-value and beta coefficient for the 'Best Mask' (that is the mask with the lowest nominal *P*-value in burden testing), and information on case/control numbers. The table here is restricted to 10 genes with at least suggestive evidence ($P < 3 \times 10^{-6}$). Results can be queried through the following link: https://hugeamp.org:8000/research.html?ancestry=mixed&cohort=UKB_450k_AoU_250k_MGB_53k_META_overlapcorrected&file=600Traits.csv&pageid=600_traits_app&phenotype=phecode_250.0. Betas, which represent log(odds ratios), were estimated using inverse-variance-weighted meta-analysis of two-sided Firth's logistic regression results. The reported *P*-values for 'rare' and 'low-freq' represent omnibus *P*-values of all relevant masks for a gene-phecode pair after combining them using the Cauchy distribution (unadjusted for multiple testing). *P*-values for mask-phecode pairs - prior to the Cauchy combination - were derived from *Z*-score-based meta-analysis of score tests from logistic mixed-effects models with saddle-point-approximation. All statistical tests and *P*-values are two-sided. LOF, loss-of-function; OR, odds ratio.

nature portfolio

Corresponding author(s): Patrick T. Ellinor

Last updated by author(s): Jun 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For the UK Biobank and All of Us sequencing datasets, data collection and data pre-processing were performed centrally, and therefore no commercial software was needed to collect data specific to the present study. Pre-processing of sequencing data for the Massachusetts General Brigham Biobank dataset was performed using the Genome Analysis Toolkit v4.1 (https://github.com/broadinstitute/gatk/releases), as described in the Supplementary Note.
Data analysis	Quality-control of individual level data was performed using Hail version 0.2 (https://hail.is) as well as PLINK version 2.0.a (https://www.cog-genomics.org/plink/2.0/). Variant annotation was performed using VEP version 105 (https://github.com/Ensembl/ensembl-vep). Main rare variant association analyses were performed using an adaptation of the R package GENESIS version 2.18 (https://rdrr.io/bioc/GENESIS/man/GENESIS-package.html), which has previously been made available by us through the GitHub repository https://github.com/seanjosephjurgens/UKBB_200KWES_CVD/ version 1.2 (DOI: 10.5281/zenodo.11638262). Meta-analyses were performed using custom code available in the same repository, and using METAL (2017-12-21 release). Analyses that were run in R, were run within R version 4 (https://www.r-project.org).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Results from our gene-based association analyses are available for browsing and download through our online portal (https://hugeamp.org:8000/research.html?pageid=600_traits_app_home). Bulk download of summary statistics is possible via the Cardiovascular Disease Knowledge Portal (https://cvd.hugeamp.org/downloads.html). Access to individual level UK Biobank data, both phenotypic and genetic, is available to bona fide researchers through application on the UK Biobank website (https://www.ukbiobank.ac.uk). The final release of the exome sequencing dataset of UK Biobank is available only through the DNAnexus Research Analysis Platform (https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform). Additional information about registration for access to the data is available at http://www.ukbiobank.ac.uk/register-apply/. Use of UK Biobank data was performed under application number 17488. Access to individual phenotypic and genetic data from All of Us is currently available to bona fide researchers within the United States through the All of Us Researcher Workbench, a cloud-based computing platform (https://www.researchallofus.org/register/). A publicly available data browser is provided by the research program: https://databrowser.researchallofus.org/. Access to individual level data for participants from the Mass General Brigham Biobank is currently not publicly available.

Other datasets used in this manuscript include: the dbNSFP database v.4.2a and v.4.3a (https://sites.google.com/site/jpopgen/dbNSFP); gnomAD exomes v.2.1 (https://gnomad.broadinstitute.org/downloads); the Online Mendelian Inheritance in Man (OMIM) database (omim.org) accessed on August 25th 2022; and Ensembl release 105 (https://www.ensembl.org/info/data/index.html); the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) was accessed in December 2022.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were based on the number of samples for which phenotypic and genetic (WGS and WES) data were available in the UK Biobank, All of Us and Massachusetts General Brigham Biobank datasets at the time. No power calculations were performed to pre-determine the required sample size.
Data exclusions	For the UK Biobank WES dataset, we used samples that passed central quality controls, and then we removed individuals who revoked their consent, duplicated samples, sex-mismatched samples, samples with overall call rates < 90%, and individuals who were outside 8 standard deviations from the mean for various other metrics (Ti/Tv ratio, het/hom ratio, SNV/INDEL ratio, and the number of singletons). For the All of Us WGS dataset, most QC was performed centrally and consisted of per-sample QC, including fingerprint concordance (array vs. WGS data), sex concordance (genetically determined vs. self-reported), cross-individual contamination rate and coverage to detect major errors, such as sample swaps or contamination. Participants who failed these tests were removed from the release. We removed flagged participants (population outliers) and possible duplicates from the current study For the Massachusetts General Brigham Biobank WES dataset, we removed duplicated samples, sex-mismatched samples, samples with discordance between sequence and genotyping array data, samples with overall call rates < 90%, and individuals who were outside 8 standard deviations from the mean for various other metrics (Ti/Tv ratio, het/hom ratio, SNV/INDEL ratio, and the number of singletons).
Replication	In the present study, the goal was to assemble a large meta-analysis of available sequencing biobanks, as to develop a strong discovery dataset of large size across hundreds of outcomes. Therefore, for the gene-based rare variant associations, no separate dataset was sought a priori to replicate the significant findings. After querying significant gene-phenotype findings, we did identify a small list of interesting novel findings for which we attempted replication: YLPM1 with bipolar disorder and personality disorders; UBR3 with hypertension, diabetes type 2, and obesity; MIB1 with diabetes; SYTL1 with hypothyroidism. In particular, we i) queried large-scale sequencing association studies if available and did not include UK Biobank, All of Us or MGB, and ii) we queried LOF results from FinnGen version 10 (if imputed LOF variants were available for the gene). For YLPM1, we queried published results from a large exome sequencing study on bipolar disorder (https://bipex.broadinstitute.org/gene/ENSG00000119596) where ultra-rare LOF and missense variants reached nominal significance (OR 3.4, P=0.01; one-sided Fisher exact test); nevertheless, we cannot exclude some sample overlap with our discovery datasets. For UBR3 and MIB1 with type 2 diabetes, we used a published large-scale exome sequencing study, where UBR3 LOFs could not be replicated due to only a single carrier in the dataset (OR 3.1, one-sided P=0.31, https://hugeamp.org/gene.html?gene=UBR3) and MIB1 LOFs showed the same effect size as in discovery but also did not reach significance (OR 1.3, one-sided P=0.08, https://hugeamp.org/gene.html?gene=MIB1). In FinnGen, imputed LOFs could only be found for SYTL1 and MIB1; both showed consistent effects when compared to our discovery findings: SYTL1 and "Hypothyroidism, strict autoimmune" reached significance (rs1238817269, OR 1.22, one-sided P=0.005) as did MIB1 with "Type 2 diabetes, wide definition" (rs200545301, OR 2.1, one-sided P=0.01).

For analyses of rare variant effect size consistency across ancestries, we assessed two different rare variant masks (namely LOF variants and ultra-rare missense variants) separately. Here, the two orthogonal rare variant masks were chosen to represent cross-validation of trans-ancestry effect size correlation.

Randomization

Samples were not experimentally randomized, given that the exposure in our analysis is genetic variation.

Blinding

No formal blinding was performed during analysis of the data. We note, however, that the main discovery analyses represented exome-by-phenome-wide association tests, where all genes/variants reaching minimum allele count criteria were tested for association with 601 disease endoints. At the same time, group allocation was performed systematically through billing codes (phenotype) and genetic variation (genotype), both of which were not influenced in any way by knowledge of the analysts; therefore blinding was not relevant to the study design.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

We combined large-scale exome sequence data from the UK Biobank (UKB) and the Mass General Brigham Biobank (MGB), with whole-genome sequencing data from All of Us (AoU). After QC procedures, we had a total of 748,879 individuals including 454,162 from UKB, 242,902 from AoU and 51,815 from MGB.

The mean age in UKB was 57.0 (SD=8.1) years, while 207,859 had a genetically-determined male sex (45.8%); in AoU the mean age was 51.7 (SD=16.9) and 94,882 (39.1%) had genetic male sex; in MGB the mean age was 57.7 (SD=17.6) and 22,946 (44.2%) had genetic male sex.

As expected, the ancestral diversity was greatest in AoU with 49.9% of participants having an ancestry other than European (most notably 21.0% African and 16.6% Admixed-American ancestry). In contrast, 94.4% and 83.5% of samples from UKB and MGB were determined to be of European ancestry. Across the three datasets, 119,660 individuals (16.0%) were of a defined ancestry other than European, and another 35,576 samples (4.7%) were of undefined/admixed ancestry, totaling 155,236 non-European samples (20.7%).

Recruitment

For UKB, prospective participants were invited to visit an assessment centre, at which they completed an automated questionnaire and were interviewed about lifestyle, medical history and nutritional habits; basic variables such weight, height, blood pressure etc. were measured; and blood and urine samples were taken. These samples were preserved so that it was possible to later extract DNA and measure other biologically important substances. During the whole duration of the study it was intended that all disease events, drug prescriptions and deaths of the participants are recorded in a database, taking advantage of the centralized UK National Health Service.

For MGB (formerly known as Partners Biobank) samples were prospectively recruited - in an ongoing observational design - from a multicenter health system in Eastern Massachusetts. In MGB, participants are enrolled with broad-based consent collected by local research coordinators, either as part of a collaborative research study or electronically through a patient portal. Demographic data, blood samples and surveys are collected at baseline and linked to electronic health record data.

For AoU, samples were enrolled in a longitudinal cohort study (with aim of including 1 million racially, ancestrally and demographically diverse participants) from across the United States. Data is prospectively collected, combining phenotypic data from various sources including patient-derived information and electronic health record linkage. One of the goals set by AoU was to recruit individuals that have been and continue to be underrepresented in biomedical research because of limited access to health care.

Ethics oversight

The UKB resource was approved by the UK Biobank Research Ethics Committee and all participants provided written informed consent to participate. Use of UKB data was performed under application number 17488 and was approved by the local Massachusetts General Hospital Institutional Review Board.

Use of AoU data was approved under a data use agreement between the Massachusetts General Hospital and the AoU

program.

All adult patients provided informed consent to participate. A small number of children were enrolled with IRB-approved assent forms; upon reaching 18 years of age all enrolled children had to provide consent or were removed from the study. The Human Research Committee of MGB approved the Biobank protocol (2009P002312).

Note that full information on the approval of the study protocol must also be provided in the manuscript.