Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects

Konrad J. Karczewski[1,2,3,*], Rahul Gupta[1,2,4,*], Masahiro Kanai[1,2,5,*], Wenhan Lu[1,2,6], Kristin Tsuo[1,2,4,6], Ying Wang[1,2,6], Raymond K. Walters[2,6], Patrick Turley[7,8], Shawneequa Callier[9,10], Nikolas Baya[1,2,6], Duncan S. Palmer[1,2,6], Jacqueline I. Goldstein[1,2,6], Gopal Sarma[1,2,6], Matthew Solomonson[1,2], Nathan Cheng[1,2,6], Sam Bryant[1,2,6], Claire Churchhouse[1,2,6], Caroline M. Cusick[1,2,6], Timothy Poterba[1,2,6], John Compitello[1,2,6], Daniel King[1,2,6], Wei Zhou[1,2,6], Cotton Seed[1,2,6], Hilary K. Finucane[1,2,6], Mark J. Daly[1,2,6,11], Benjamin M. Neale[1,2,3,6], Elizabeth G. Atkinson[12], Alicia R. Martin[1,2,6]

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
[3]Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[4]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA
[5]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA 02115, USA
[6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[7]Department of Economics, University of Southern California, Los Angeles, CA, 90089
[8]Center for Economic and Social Research, University of Southern California, Los Angeles, CA, 90089
[9]Department of Clinical Research and Leadership, The George Washington University, Washington, DC 20037, USA
[10]Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
[11]Institute for Molecular Medicine, Finland, Helsinki, Finland
[12]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
*These authors contributed equally

Summary

Large biobanks, such as the UK Biobank (UKB), enable massive phenome by genome-wide association studies that elucidate genetic etiology of complex traits. However, individuals from diverse genetic ancestry groups are often excluded from association analyses due to concerns about population structure introducing false positive associations. Here, we generate mixed model associations and meta-analyses across genetic ancestry groups, inclusive of a larger fraction of the UKB than previous efforts, to produce freely-available summary statistics for 7,271 traits. We build a quality control and analysis framework informed by genetic architecture. Overall, we identify 14,676 significant loci in the meta-analysis that were not found in the European genetic ancestry group alone, including novel associations for example between *CAMK2D* and triglycerides. We also highlight associations from ancestry-enriched variation, including a known pleiotropic missense variant in *G6PD* associated with several biomarker traits. We release these results publicly alongside FAQs that describe caveats for interpretation of results, enhancing available resources for interpretation of risk variants across diverse populations.

Introduction

Paired genetic and phenotypic data have grown explosively over the last decade, particularly with the maturity of massive global biobank efforts (Abul-Husn and Kenny 2019; Zhou et al. 2022). These data have led to the identification of over 275,000 associations between genetic loci and human traits and diseases to date (Buniello et al. 2019). However, genome-wide association studies (GWAS) tend to be vastly Eurocentric, limiting their generalizability to ancestrally and globally diverse populations (Sirugo, Williams, and Tishkoff 2019; Martin, Kanai, et al. 2019). Imbalanced data generation is the primary cause of this issue, but a secondary contributor is that GWAS tend to analyze the largest ancestry group in a dataset and exclude minority groups to avoid potential false positives arising due to population stratification. Using the ancestrally diverse data already available is critical for several reasons, including increasing the applicability of genetic findings across populations, as well as increasing power for gene discovery due to increased genetic diversity. Underrepresented populations also disproportionately contribute to genomic discoveries: for example, African ancestry and Hispanic/Latin American groups only comprise 2.4% and 1.3% of individuals represented in the GWAS catalog, respectively, but contribute 7% and 4.3% of associations overall (Morales et al. 2018). In comparison, 78% of individuals have primarily European ancestry but contribute only 54% of associations.

Box 1 | Genetic ancestry in the Pan-UKB

In this paper, we use principal component analysis to split the UK Biobank cohort into six groups where people within a group have similar genetic ancestries to each other. To increase the generalizability of our results, our groups are based on genetic ancestry groups that were defined in a pair of well-known reference panels (see Supplementary Information). Throughout this manuscript, we also refer to these groups using the ancestry labels assigned by those panels: EUR (European), CSA (Central/South Asian), AFR (African), EAS (East Asian), MID (Middle Eastern), and AMR (Admixed American). Dividing our sample into subgroups is necessary to reduce bias from population stratification; however, it may create the illusion of discrete, biological ancestral populations. This concern is augmented since the ancestry labels are generally based on geographic regions that may be associated with race or ethnicity. For this reason, we highlight four points, in accordance with recent reports (National Academies of Sciences, Engineering, and Medicine 2023; Meyer et al. 2023):

1. Genetic ancestry is a continuum and the human population cannot be divided into discrete, biologically-meaningful categories.
2. The groups used in this paper are informed by statistics, but they ultimately arise from social and historical factors.
3. There is more variation within than between our defined ancestry groups (Witherspoon et al. 2007).
4. Genetic ancestry is a distinct concept to identity- and geography-based descriptors (Bamshad et al. 2004)

More details on these and related points can be found in the Frequently Asked Questions section of the Supplementary Information.

Many trait- and disease-specific consortia have conducted multi-ancestry GWAS to increase sample sizes and investigate generalizability, which have yielded deep insights into biology (Sinnott-Armstrong et al. 2021). For most traits and diseases, a large number of variants--each with a small effect size--contribute to phenotypic variation. As a general rule, causal variant effect sizes tend to be largely consistent across populations (Lam et al. 2019; Chen et al. 2021; Hou et al. 2023), showing very low evidence of heterogeneity when accounting for differences in allele frequency and linkage disequilibrium (LD) across populations. Multi-ancestry studies, however, have highlighted some clear examples where population-enriched variants provide power for genetic discovery that would not be readily identified only in European ancestry studies. Some examples include associations between *HNF1A* and type 2 diabetes identified in Latin American populations (SIGMA Type 2 Diabetes Consortium et al. 2014); loss-of-function variant associations in *PCSK9* and low LDL cholesterol identified in African Americans (Cohen et al. 2005); associations between inflammatory bowel disease and variants enriched in East Asian ancestries (Z. Liu et al. 2023); associations between the Duffy-null allele and malaria identified in sub-Saharan African populations (Miller et al. 1976); and associations between *APOL1* and resistance to trypanosomes but also chronic kidney disease in African and African diaspora populations (Genovese et al. 2010; Ross 2019; Genovese, Friedman, and Pollak 2013). Several of these associations have clinical implications that benefit individuals from all backgrounds, such as the *PCSK9* association which led to one of the first genetically informed therapies to prevent heart disease.

In addition to genomic discovery, multi-ancestry genetic analyses are also critical for resolving, interpreting, generalizing, and translating GWAS results. For example, diverse GWAS provide greater resolution into the identification of putative causal variation via fine-mapping due to: 1) a joint analysis of different patterns of LD across diverse populations (Mägi et al. 2017; Asimit et al. 2016), and 2) larger sample sizes with more diverse ancestral recombination history (Mahajan et al. 2020; Huang et al. 2017; Schaid, Chen, and Larson 2018; Graff et al. 2021; Luo et al. 2020), improving identification of actionable targets using GWAS results. Another key application of GWAS is the construction of polygenic risk scores, with numerous potential clinical implications including disease risk stratification (Polygenic Risk Score Task Force of the International Common Disease Alliance 2021). Genetic prediction accuracy closely reflects the composition of

the study cohorts from which such models are derived, leading to the widely-replicated observation that Eurocentric discovery GWAS produce polygenic scores whose predictive powers are reduced by several fold in poorly represented genetic ancestry groups (Martin, Kanai, et al. 2019; Martin et al. 2017; Scutari, Mackay, and Balding 2016; Wang et al. 2020; Ding et al. 2023). Multi-ancestry studies have already begun improving genetic prediction accuracy in underrepresented populations for some phenotypes (Chen et al. 2021; Lam et al. 2019; Conti et al. 2021; Bigdeli et al. 2019; Zhou et al. 2022). Biobank-scale genetic correlation analyses and phenome-wide association studies (PheWAS) are additional approaches that have aided our understanding of molecular and epidemiological relationships across a wide variety of traits, however these have also tended to be Eurocentric in representation due to imbalances in GWAS (Bastarache 2021; Denny et al. 2013; Bastarache et al. 2018; Zheng et al. 2017); more diverse representation is needed particularly to disentangle population-enriched genetic and environmental factors that contribute to disease risk.

The UK Biobank (UKB) (Ben-Eghan et al. 2020; Bycroft et al. 2018) is one of the most impactful biobanks to date, due to its large number of participants, depth and breadth of phenotyping, consistency in data generation, and uniquely open and straightforward access model; despite this, analyses have largely focused on only European ancestry participants. While most (95%) UKB participants fall into the EUR group (the genetic ancestry group with predominantly European ancestries), more than 20,000 participants have primarily non-European ancestries (**Box 1**). The UKB therefore provides opportunities to conduct among the largest genetic studies to date of thousands of phenotypes in diverse continental ancestries.

Here, we describe the Pan-UK Biobank Project (https://pan.ukbb.broadinstitute.org/), a multi-ancestry genetic analysis of thousands of phenotypes. We extend previous phenome-wide association resources, adding 14,676 independent associations using meta-analysis of multiple ancestry groups rather than only the EUR subset. We highlight discoveries enabled by multi-ancestry analysis, including an association between *CAMK2D* and triglycerides. We also show how ancestry-enriched variation highlights interesting biology, including a pleiotropic association between *G6PD* and a number of biomarker traits, primarily accessible in the AFR genetic ancestry group. We describe and release pipelines for a robust analytic framework to facilitate future multi-phenotype multi-ancestry genetic analyses to improve gene discovery.

Results

A resource of multi-ancestry association results from 7,271 phenotypes in the UKB

To maximize genomic discovery in the UKB, we performed a multi-ancestry analysis of 7,271 phenotypes, with analysis of up to 441,331 total individuals from up to 6 genetic ancestry groups followed by meta-analysis. We assigned each individual to a genetic ancestry group by conducting principal components analysis (PCA) on a diverse reference panel consisting of the Human Genome Diversity Panel (HGDP) and 1000 Genomes Project genotype data (1000 Genomes Project Consortium et al. 2015; Li et al. 2008) (**Supplementary Table 1**), then projected UKB individuals into this space using their genotype data using a random forest (probability > 0.5) to partition the dataset into six genetic ancestry groups (**Extended Data Fig. 1, Supplementary Figs. 1-13, Supplementary Tables 2-5**). After initial assignments, to further reduce stratification, we removed ancestry outliers based on multidimensional centroid distances from average PC values (**Supplementary Information**). The ancestry groups follow expected trends based on self-reported ethnicity, continental birthplaces, and country of birth (**Supplementary Tables 6-7, Supplementary Fig. 14**), however we emphasize that genetic ancestry is a distinct concept from these other identity- and geography-based descriptors (see **Box 1** and **Supplementary Information**, FAQ). For example, a person assigned to some genetic ancestry group may or may not report having a corresponding ethnic identity (Mathieson and Scally 2020; Lewis et al. 2022; National Academies of Sciences, Engineering, and Medicine et al. 2023). From these data, we performed sample and variant quality control (QC) to remove sample outliers and low-quality and ultra-rare variants (**Methods**).

We used a two-step approach for genetic association testing, first performing GWAS within each genetic ancestry group for a given trait using a generalized mixed model approach (SAIGE; (Zhou et al. 2018)), and then performing inverse-variance weighted meta-analysis across all within-ancestry GWAS performed for that trait. We observed that our approach reduced stratification and type 1 error rate for several phenotypes compared to an alternative single-step "mega-analysis" approach that included all individuals in a single mixed model (**Extended Data Table 1**). Specifically, we observed genomic control statistics closer to 1 using our

approach, as well as comparable and in some cases improved statistical power with discovery of additional genome-wide significant loci using our approach.

Using this two-step approach, we performed association testing across 10-23 million SNPs (**Supplementary Table 8**) for all ancestry-trait pairs for quantitative phenotypes, and all binary traits with at least 50 cases observed in an ancestry group (except EUR, 100 cases, given the larger sample size; **Fig. 1a**; **Supplementary Fig. 15**). Altogether, this resulted in an analysis of 16,554 ancestry-trait pairs across 7,271 traits, of which 924 were run in all six populations (**Fig. 1b**). These traits include thousands of newly-analyzed phenotypes, including aggregate disease combinations (i.e., phecodes (Denny et al. 2013)), prescription drug status, and continuous updates to the COVID-19 phenotypes (COVID-19 Host Genetics Initiative 2021). We developed a summary statistics QC protocol to remove low-confidence variants and associations (**Supplementary Figs. 16-19**).

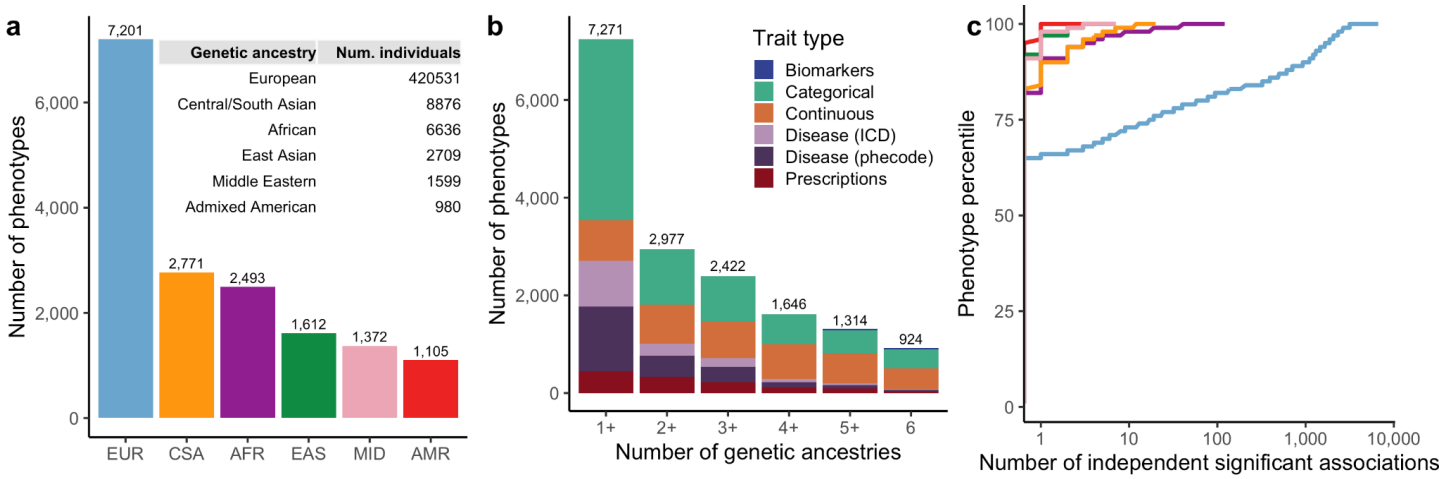


Figure 1 | Pan-UK Biobank GWAS resource facilitates multi-ancestry multi-trait analyses a-b, The number of phenotypes with GWAS computed across: **a**, genetic ancestry groups and **b**, number of genetic ancestry groups stratified by trait type; within each bar, trait types are ordered by total number of traits. **c**, A cumulative distribution function showing the number of independent genome-wide significant ($p < 5 \times 10^{-8}$) associations across all phenotypes analyzed within each genetic ancestry group. Independence was defined using clumping in plink (Purcell et al. 2007), including an r^2 threshold of 0.1 for ancestry-matched reference panels (Supplementary Information). Colors are consistent in **a** and **c**.

To avoid double counting variants in LD, we determined a set of LD-independent loci within each ancestry-trait pair using in-sample reference panels (**Supplementary Figure 20**). We extended this to our meta-analysis by constructing reference panels of 5,000 individuals matched by genetic ancestry proportions. For 924 traits for which association analysis was performed for all six ancestry groups, we discovered a mean

of 2.26 independent loci (sd=9.2) in non-EUR genetic ancestry groups, demonstrating that we have sufficient power for novel discovery in these understudied groups for which GWAS have not previously been run at scale in the UKB. Owing to the substantially larger sample size in EUR, we find more genome-wide significant associations per phenotype in EUR: 25% of phenotypes have >18 associated loci in EUR while fewer than 10% of phenotypes have 3 or more associated loci across all non-EUR populations (**Fig. 1c**). Despite the lower sample size but consistent with previous GWAS catalog summaries (Morales et al. 2018), we find a higher mean number of significant regionally-independent associations in AFR compared to CSA, potentially suggesting increased power from higher heterozygosity (**Extended Data Table 2**).

A framework for identifying high-quality phenotypes in diverse ancestries with massively imbalanced sample sizes

Previous phenome-wide studies of the UK Biobank have performed association testing of thousands of traits; however, these have used linear models (Howrigan 2017) and/or only analyzed a single ancestry group (Zhou et al. 2018; Howrigan 2017). In our analysis of data from multiple ancestry groups, we identified extensive challenges due to extreme imbalances of sample sizes, leading to unreliability of some association tests. Here, we propose a framework for testing the reliability of GWAS of multiple ancestries.

Specifically, we summarized properties of genotype-phenotype associations, including genomic control (λ_{GC}), heritability, and residual population stratification. We estimated SNP-heritability (h^2_{SNP}) for each ancestry-trait pair through several strategies. For a set of pilot phenotypes in EUR (**Supplementary Table 9**), we confirmed a high concordance for two methods (**Extended Data Fig. 2**; S-LDSC (Bulik-Sullivan et al. 2015; Finucane et al. 2015) and RHE-mc (Pazokitoroudi et al. 2020)), as well as between our results and previously reported results (**Supplementary Figs. 21-24**). To balance computational considerations with power gains, we used S-LDSC for all traits in EUR, and RHE-mc for all other genetic ancestry groups (**Extended Data Fig. 2b; Supplementary Information**). Across groups, the number of traits with significant heritability ($z \geq 4$) correlates with sample size of genetic ancestry group (**Extended Data Fig. 3a**; **Supplementary Fig. 25**). We identified traits with significant out-of-bounds (i.e. outside 0-1) heritability point estimates, deflated λ_{GC} estimates, and/or

elevated S-LDSC ratio statistics, particularly in traits that are prone to population stratification (e.g. country of birth, dietary preferences; **Supplementary Fig. 26**).

To increase confidence in analyses across traits and ancestries, we devised a quality control strategy to systematically flag traits with potentially problematic GWAS results while retaining GWAS of heritable traits passing QC in two or more populations (**Fig. 2a, Supplementary Figs. 27-29**, and Supplementary Information). Overall, we pruned 16,518 ancestry-trait pairs with available GWAS to 1,091 ancestry-trait pairs that passed all filters spanning 452 traits (**Fig. 2a**). Of the ancestry-trait GWAS pairs that passed, the majority were shared between the two largest ancestry groups (EUR and CSA), with 147 phenotypes found in three or more genetic ancestry groups (**Supplementary Fig. 29**). As many phenotypes in the biobank are correlated with each other, we pruned to a maximal set of independent phenotypes with pairwise $r^2 < 0.1$, resulting in a set of 151 phenotypes (**Supplementary Fig. 30**), and computed polygenicity estimates for all high quality traits (**Supplementary Fig. 31**). We note that these genome-wide summaries aid in prioritizing phenotypes with broad heritable components, although the failure of a phenotype in this framework does not necessarily preclude true individual SNP-level associations.

a

Phenotype QC filters	Total	EUR	CSA	AFR	EAS	MID
Significant and bounded heritability	2,210	993	745	193	84	195
Calibrated λ_{GC}	2,186	988	742	187	81	188
Controlled S-LDSC ratio	2,049	926	699	173	77	174
Passes all filters in EUR and ≥1 other ancestry group	1,091	452	349	92	62	136

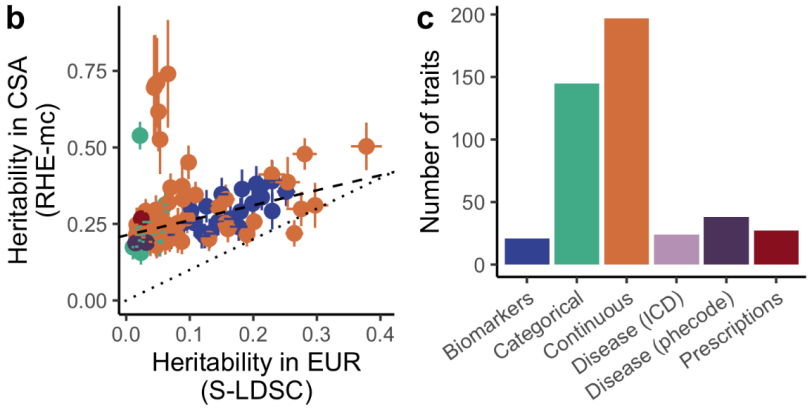


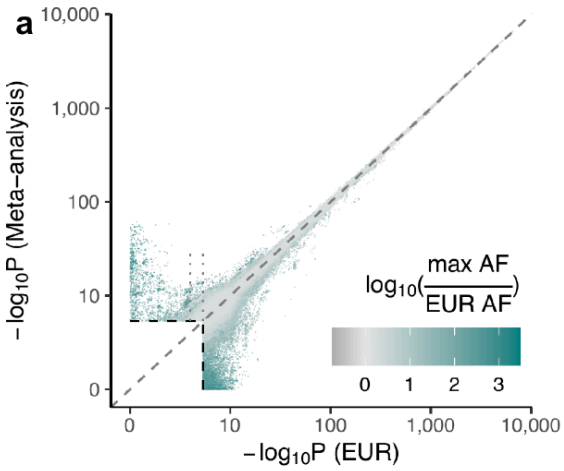
Figure 2 | Heritability informs robustness of GWAS across ancestry-trait pairs. a, to balance large differences in sample size across genetic ancestry groups, we developed a stepwise series of phenotype QC filters applied based on heritability estimates, genomic control (λ_{GC}), evidence of residual stratification (S-LDSC ratio), and high-quality data in multiple ancestry groups (see also **Supplementary Figs. 25-27** for more detail). **b**, comparison of heritability estimates across EUR and CSA genetic ancestry groups. Note that two different heritability estimation methods are used to balance computational efficiency (S-LDSC in EUR) versus precision in small sample sizes (RHE-mc in CSA). Binary phenotype heritability estimates are reported on the observed scale due to highly variable prevalences and liability scaling at smaller sample sizes. The dotted line shows y=x, while the dashed line is a fitted York regression (slope = 0.49, intercept = 0.21, $p = 5 \times 10^{-12}$). A version with less filtering is shown in Extended Data Fig. 2c. **c**, The number of traits passing final QC filters by trait type. Colors are consistent in **b, c**.

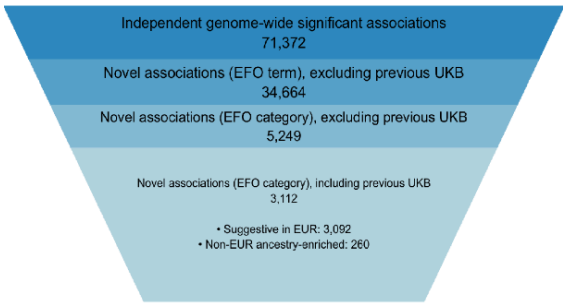
Among high-quality GWAS results that passed our broad-scale QC of GWAS results, we identified consistency in the relative magnitude of heritability estimates among populations. For example, 113 independent phenotypes passed QC in both EUR and CSA genetic ancestry groups, and the heritability estimates for these traits had a significant positive correlation (York regression $p = 5 \times 10^{-12}$; **Fig. 2b, Extended Data Fig. 2c**); the heritability estimates in this set were systematically higher in the CSA, likely reflecting a combination of winner's curse from selection of phenotypes with significant estimates in CSA and of residual population stratification. Biomarkers and continuous phenotypes tended to have the highest heritability estimates (EUR average $h^2 = 0.19$ and 0.16), whereas disease and prescription phenotypes tended to have the lowest heritability estimates (EUR average $h^2 = 0.021$ and 0.016, **Extended Data Fig. 3b**).

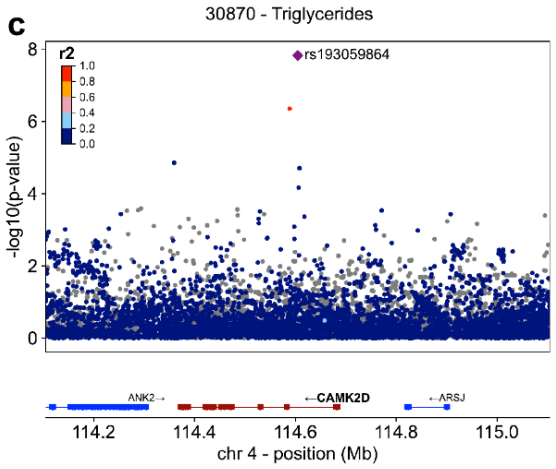
Genomic discoveries powered by Pan-UK Biobank analysis

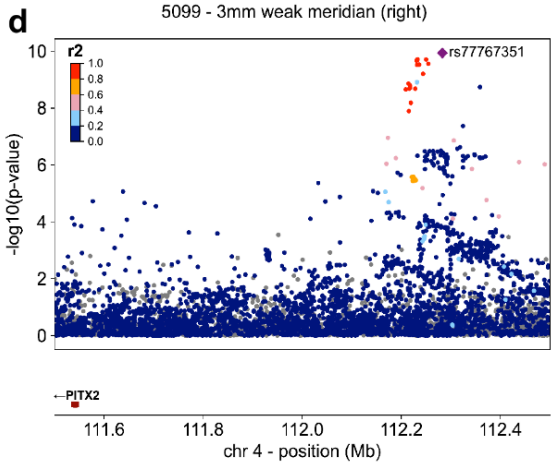
We next investigated the extent to which broader inclusion of ancestrally diverse participants identifies specific biological signals through meta-analysis, compared to EUR-only analysis. In particular, for 452 high-quality phenotypes, we compared p-values for the multi-ancestry meta-analysis to those from the EUR GWAS, which were overall highly correlated (**Fig. 3a**; **Supplementary Fig. 32**; $r^2 = 0.999$; $p < 10^{-100}$). We identified 237,360 significant ($p < 5 \times 10^{-8}$), LD-independent associations in the largest meta-analysis available across 431 phenotypes (**Extended Data Fig. 4a**). Of these, 14,676 (6.2%) were not significant in EUR, representing new biology discovered solely by analyzing already-available data from diverse populations.

In order to compare to prior trait-specific and systematic analyses in the UK Biobank (Howrigan 2017; Zhou et al. 2018) and other datasets, we created a framework to quantify associations added by the Pan-UK Biobank analysis compared to existing associations in Open Targets Genetics (Ghoussaini et al. 2020). Two challenges with phenome-wide comparisons is the pervasiveness of pleiotropy coupled with mapping phenotypes across GWAS, which may be coded slightly differently and have subtle differences in phenotype definitions (Solovieff et al. 2013). Through semi-manual curation, we mapped 3,047 (42%) of our traits to terms in the Experimental Factor Ontology (EFO), of which 2,566 matched a study in the GWAS Catalog; our ability to cross-reference phenotypes varied by trait type (**Supplementary Table 10**). In order to assess novelty conservatively, we computed distance-based independent associations, and compared to known associations









accordingly. Specifically, we defined a variant as novel only if no known associations for the same EFO broad category are present within 1 Mb. We identified 71,372 regionally-independent associations that mapped to an EFO category, of which 68,260 (96%) were previously reported for the same EFO category. 3,112 (4%) were not previously identified (**Fig. 3b**), with this rate varying by EFO category (**Extended Data Fig. 5**). The X chromosome contributes an outsized proportion of this novelty, with 573 of 2,448 (23%) associations not previously reported for the same EFO category. This disproportionate contribution is likely due to exclusion of the X chromosome in many previous GWAS.

Figure 3 | Biobank-wide analysis improves genetic discovery. **a**, Comparison of GWAS significance in EUR GWAS only versus meta-analysis across ancestries. Variants with $p < 10^{-10}$ in either analysis are shown. Colors highlight that variants with higher non-EUR allele frequencies (teal) are more likely to be identified in the meta-analysis. Dashed gray line shows y=x, and the scale is logarithmic for p from 1 to 10^{-10}, and log-log for $p < 10^{-10}$. Dotted lines indicate variants suggestive in EUR, but significant in meta-analysis ($5 \times 10^{-8} < p < 10^{-6}$; see Fig. 3b). **b,** Summary of regionally-independent genome-wide significant associations. The total number, number of novel associations (at the EFO term and category level), excluding and not excluding previous UK Biobank multi-phenotype analyses (Howrigan 2017; Zhou et al. 2018). The final category shows the number of associations that are suggestive in EUR alone ($p < 10^{-6}$), and where the frequency of the variant is enriched in a non-EUR genetic ancestry group by at least 10-fold (322 that were both suggestive and enriched). **c-d**, Locuszoom plots of (**c**) *CAMK2D* (rs193059864) and triglycerides ($p = 1.5 \times 10^{-8}$; N = 416,764; allele frequency in AFR = 0.016, EUR = 1.4×10^{-4}) and (**d**) *PITX2* (rs77767351) of keratometry (3mm weak meridian - right; $p = 1.2 \times 10^{-10}$; N = 89,664), with lead variants indicated by the purple diamond and LD (r^2) for neighboring variants derived from a weighted reference panel (see Supplementary Information). Both lead variants (**c-d**) have info score > 0.9 and heterogeneity p-value > 0.1.

Newly-significant associations arise due to a combination of factors: bolstered associations that crossed the genome-wide significance threshold with mixed models from previously sub-threshold associations in standard regression, increased sample size of EUR participants, and inclusion of participants with non-EUR ancestries that either added support or contributed outsized significance due to ancestry-enriched variation. Of the 3,112 novel meta-analysis associations, 3,092 show suggestive signals (e.g. $p < 10^{-6}$) in EUR-only analyses where the multi-ancestry meta-analysis results in genome-wide significance ($p < 5 \times 10^{-8}$). Allele frequencies were enriched by at least 10-fold in at least one genetic ancestry group over EUR in at least one genetic ancestry group for 260 associations (247 both suggestive in EUR and enriched outside EUR; **Fig. 3b**). For instance, we find a significant association between triglycerides and *CAMK2D* (meta-analysis $p = 1.5 \times 10^{-8}$; EUR $p = 0.0017$; **Fig. 3c**), at a variant (rs193059864) that is 114-fold enriched in AFR (AFR frequency 1.6%, EUR frequency 1.4×10^{-4}). Another variant in this gene has recently been implicated in heart failure (Rasooly et al. 2023); however, rs193059864 is in low LD with the variant identified in this study (rs17620390; $r^2 = 0.001$), indicating these represent independent associations.

We investigated broad patterns of gene function closest to each association to assess biological relevance. We find that 66% (124/187) of haploinsufficient genes are near a novel significant association (**Extended Data Fig. 6a**; all hits shown in **Supplementary Fig. 33**), compared to 34% of all genes (5,969/17,428). The associations near haploinsufficient genes often correspond to broadly similar phenotype categories as the OMIM annotation of the gene. For instance, we find associations between SNP rs77767351 (near *PITX2*) and several keratometry measurements, including 3mm weak meridian-right (meta-analysis p-value = 1.2×10^{-10}; **Fig. 3d**). Previous studies have identified a crucial role for *PITX2* in embryonic development and tissue formation (Gage, Suh, and Camper 1999) and implicated mutations in this gene with rare Mendelian eye-related diseases, such as Axenfeld-Rieger syndrome (Tümer and Bach-Holm 2009; Berry et al. 2006). To our knowledge, this SNP has not been significantly associated in any GWAS, and indicates a potential allelic series, in which the intermediate eye phenotype associated with the common variant can provide context for the molecular mechanisms underlying the dominant condition caused by high-impact variants in *PITX2*.

Inclusion of multiple genetic ancestries improves genetic discovery

To further explore potential reasons for associations only significant in the meta-analysis, we compared effect sizes to allele frequencies by ancestry. For quantitative traits where effect sizes are directly comparable, we broadly observed the characteristic inverse relationship between minor allele frequency (MAF) and effect size, in which rarer variants tend to have larger effect sizes than common variants (Gibson 2018; Martin, Daly, et al. 2019) (**Fig. 4a**). We identified genome-wide significant associations through meta-analysis and not in the EUR GWAS in some cases due to higher ancestry-specific MAFs, which rendered some variants with larger effect sizes accessible to discovery despite the smaller sample size (**Fig. 4b**), although winner's curse may inflate some effect size estimates (D. J. Liu and Leal 2012).

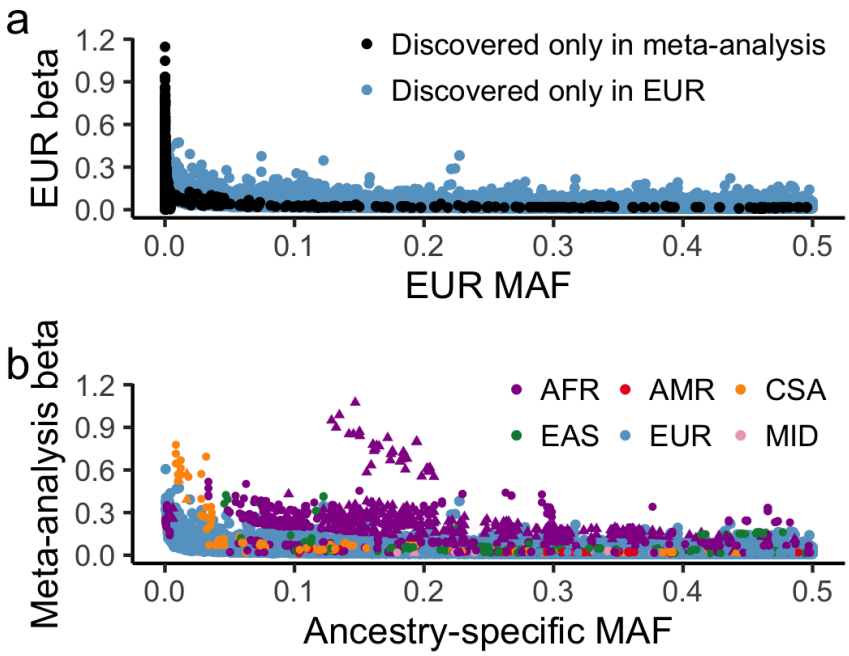


Figure 4 | Differences in allele frequencies across ancestries yield novel genetic discoveries. a, Significantly associated variants for a set of 140 quantitative traits identified in the EUR genetic ancestry group (blue) versus those discovered only in the meta-analysis (black), with allele frequencies and effect sizes in EUR shown. **b**, The same significantly associated variants as shown in **a**, but with ancestry-specific frequencies and effect sizes estimated from the multi-ancestry meta-analysis. Associations on the X chromosome (e.g. *G6PD*) are denoted with triangles. Contrasting **a-b** highlights the importance of higher allele frequencies in underrepresented ancestry groups for empowering associations.

We thus investigated the most extreme differences (i.e. associations significant in the meta-analysis, but not in the EUR GWAS, i.e. those in the upper left quadrant of **Fig. 3a**). We find that these associations

were more likely to be found at high-quality variants (**Supplementary Information**) and were 6-fold enriched for variants more common in AFR (4-fold in any non-EUR ancestry; **Extended Data Fig. 7**), indicating that these associations are likely enriched for universally tagging or putatively causal variants in multiple populations. The most extreme differences in p-values stemmed from extreme frequency differences, such as between variants in/near *G6PD* and a number of traits (**Fig. 5**), where a higher frequency in AFR increased power for association. However, perhaps paradoxically, variants with p-values more significant in EUR-only GWAS compared to meta-analysis were also 2.4-fold enriched for those with higher AFR frequencies (**Extended Data Fig. 7**). These variants are most likely not causal, but tagging variants with differential LD patterns across ancestries leading to apparent heterogeneity in meta-analysis. For instance, a variant in LD with a causal variant in EUR, but only partial LD in AFR with increased frequency will be identified as heterogeneous with little or no effect in AFR. Indeed, heterogeneity (Cochran's Q $p < 0.01$) was 2.5-fold enriched for variants with reduced significance in the meta-analysis compared to the EUR GWAS alone (**Extended Data Fig. 7; Supplementary Fig. 32**).

Refining biological signals through ancestry enriched variation

Among the most extreme differences we identified between the meta-analysis and EUR-only results was a missense variant (rs1050828) in *G6PD*, which had significant associations with five phenotypes in the AFR group (allele frequency = 16%; **Fig. 5a**) but not in EUR (allele frequency = 1.5×10^{-4}), including glycated hemoglobin (HbA1c; **Fig. 5b**), high light scatter reticulocyte count, red blood cell count, red blood cell distribution width, and mean sphered cell volume (**Fig. 5c, Supplementary Fig. 34, Supplementary Table 11**). Our results replicate previous associations between variants in/near *G6PD* and HbA1c (Sarnowski et al. 2019), and further validate the pleiotropic effect of this missense variant. We performed ancestry-specific fine-mapping of these signals (**Extended Data Fig. 8a**), which confirm the likely functional nature of this variant; however, fine-mapping of meta-analyzed summary statistics presented a significant challenge (**Extended Data Fig. 8b**), particularly due to highly imbalanced sample sizes.

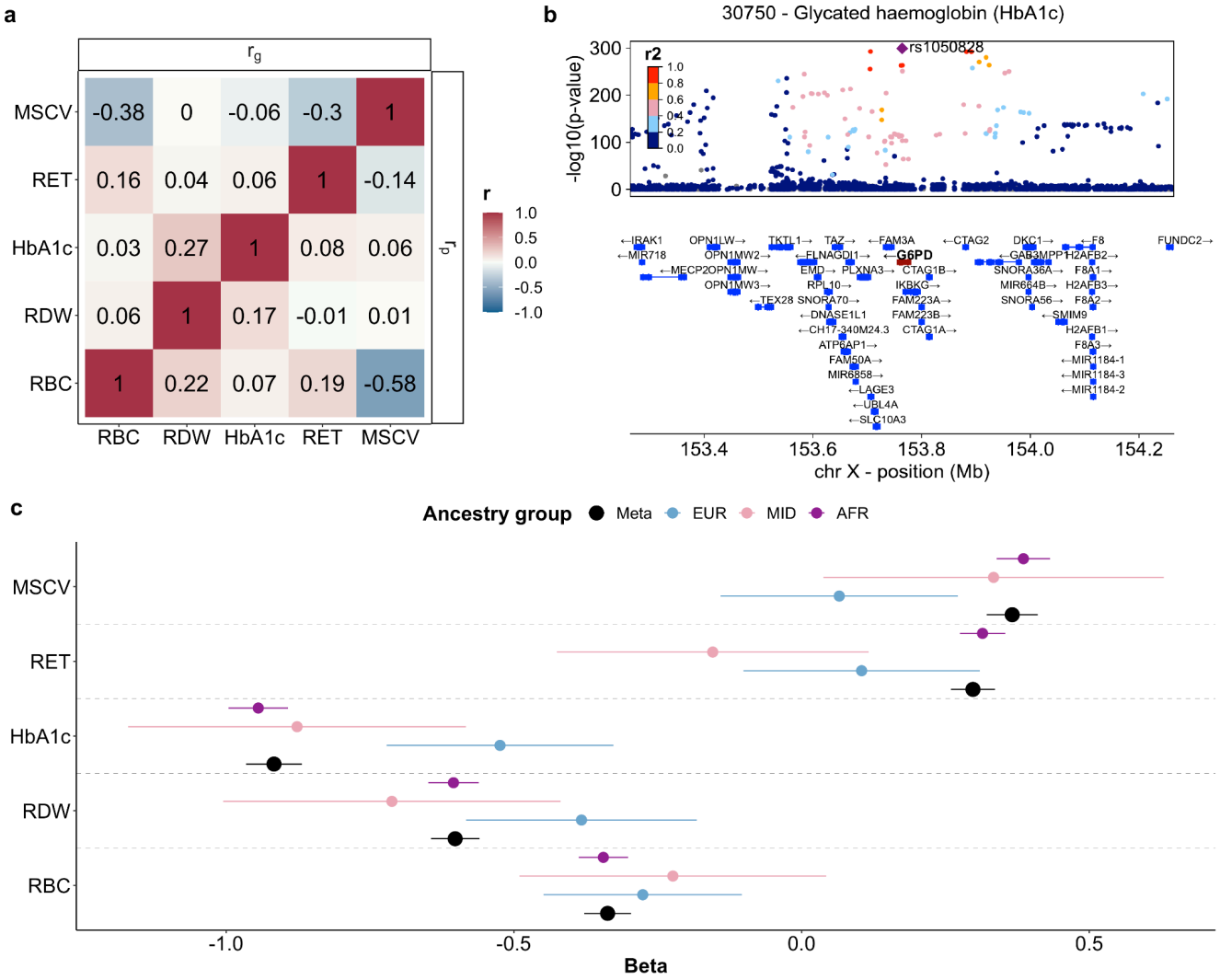


Figure 5 | Meta-analysis identifies pleiotropic signals from non-European populations. a, correlation matrix of the top five phenotypes significantly (meta-analysis $p < 5 \times 10^{-8}$) associated with rs1050828 (chrX:153764217), a missense variant in *G6PD* (all p-values are shown in Supplementary Table 11). The upper and lower triangles of the matrix represent the genetic and phenotypic correlations, respectively. (RET: High light scatter reticulocyte count; RDW: Red blood cell (erythrocyte) distribution width; RBC: Red blood cell (erythrocyte) count; MSCV: Mean sphered cell volume; HbA1c: Glycated hemoglobin) **b**, Locuszoom plot of a 1Mb region around the lead SNP rs1050828 (purple diamond) for the meta-analysis result of glycated hemoglobin ($p = 1.1 \times 10^{-299}$; N = 408,539), as in **Fig. 3c-d**. **c**, Forest plot showing association beta for each phenotype for rs1050828, including a meta-analysis across all available ancestry groups (full results shown in **Supplementary Fig. 34**). rs1050828 was low-frequency in CSA and EAS and thus, association statistics were not computed. Error bars correspond to 95% confidence intervals.

Discussion

We present a multi-ancestry genetic study and resource that extends previous analyses focused only on EUR participants to a wider set of UKB participants from understudied genetic ancestry groups. Here, we build a new framework, including pipelines and best practices for multi-phenotype multi-ancestry analysis, that

we recommend be adopted by groups performing similar analyses for the UK Biobank and other diverse biobanks, such as the *All of Us* project. Our quality control framework provides a blueprint to identify phenotypes with robust GWAS, particularly when sample sizes are imbalanced and/or with stratification. Our results show that diverse analysis is critical for maximizing biological discovery even with imbalanced sample sizes, and that full data release is critical to broad comparisons across datasets. We provide summary statistics for 16,554 GWAS in scalable and per-phenotype form, as well as reference data (including LD scores/matrices and sample metadata returned to the UK Biobank) to reduce the barrier for future multi-ancestry analyses of these existing and new phenotypes, respectively. Indeed, this work contributed continuously updated summary statistics to the COVID-19 host genetics initiative (COVID-19 Host Genetics Initiative 2021), as well as broader efforts such as the Global Biobank Meta-analysis Initiative (Zhou et al. 2021).

In this work, we highlight challenges in performing multi-phenotype analyses of thousands of phenotypes, particularly around quantifying novelty and fine-mapping in the presence of imbalanced sample sizes. We identify 3,112 genome-wide significant associations that were not previously found in Open Targets Genetics (Ghoussaini et al. 2020). These discoveries arise from ancestry-enriched variants (e.g. *CAMK2D* rs193059864 with frequency enriched in AFR) as well as previously sub-significant signals ($5 \times 10^{-8} < p < 10^{-6}$; e.g. rs77767351 in *PITX2*), which are significantly associated here due to a combination of larger sample sizes in the European ancestry group, inclusion of participants with primarily non-European ancestries, and use of more advanced mixed model analysis that increase statistical power for discovery. Some associations demonstrate allelic series, in which variants in the same gene have different levels of effects on related phenotypes: common variants in *PITX2* are associated with variation in traits (keratometry traits), whereas rare pLoF variants in ClinVar are related to severe diseases (e.g. Axenfeld-Rieger syndrome).

We additionally highlight a set of pleiotropic associations for a common missense variant in *G6PD* in the AFR genetic ancestry group that is rare and thus inaccessible in other genetic ancestry groups. While some of these phenotypes were filtered out by our QC pipeline in AFR, we note that these broad filters do not preclude individual true associations. Fine-mapping of this locus in AFR shows sensible credible sets, whereas

fine-mapping across a meta-analyzed cohort leads to instability, particularly in this case where a smaller ancestry group contributes an outsized fraction of the variance explained (i.e. $2pq*\beta^2$). These results demonstrate a clear need for cohorts with more balanced sample sizes across ancestry groups, such as the burgeoning *All of Us* Research Program cohort (All of Us Research Program Genomics Investigators 2024). More diverse cohorts will require care when running GWAS to retain high quality association data, particularly as statistical methods often assume homogeneity that can be easily violated when analyzing genomes with high degrees of recent admixture. Scaling methods that consider recent admixture (Atkinson et al. 2021) may improve these analyses further. While analyzing increasingly diverse GWAS data can be challenging, it is scientifically imperative to accurately identify novel associations, resolve which variants are most likely to be causal, and increase accuracy in polygenic score analysis for all.

Conducting systematic analyses across a massive range of traits raises sensitivities that users of this resource will attempt to make inappropriate comparisons across ancestry groups. Previous work has shown that comparing polygenic score distributions across traits is not scientifically meaningful, and we discourage use of this resource to make race- and ethnicity-based comparisons. We encourage consulting a resource of FAQs we carefully developed as part of this project (https://pan.ukbb.broadinstitute.org/) when using this resource to make any population-based comparisons and evaluating risks versus benefits.

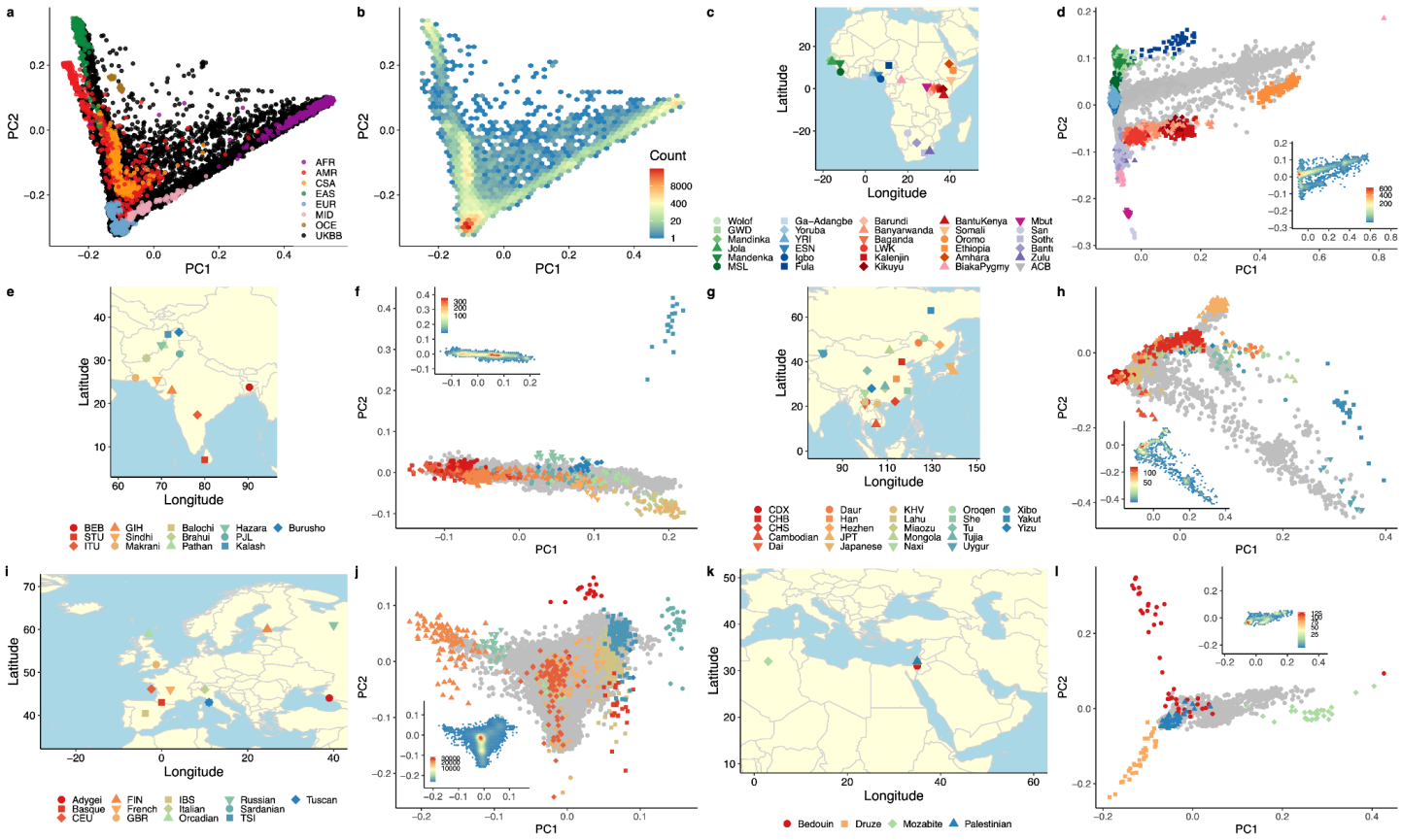
Acknowledgments

This work was supported by the Novo Nordisk Foundation (NNF21SA0072102), NIH grants R37MH107649, R00MH117229, K01MH121659, F31HL167378, and F30AG074507, and BroadIgnite funding. This study used data from the UK Biobank. UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval. This approval means that researchers do not require separate ethical clearance and can operate under the RTB approval.

Data availability

All data are available at https://pan.ukbb.broadinstitute.org/, and sample metadata is available in the UK Biobank showcase under return number 2442: https://biobank.ndph.ox.ac.uk/ukb/dset.cgi?id=2442.

Extended Data Figures



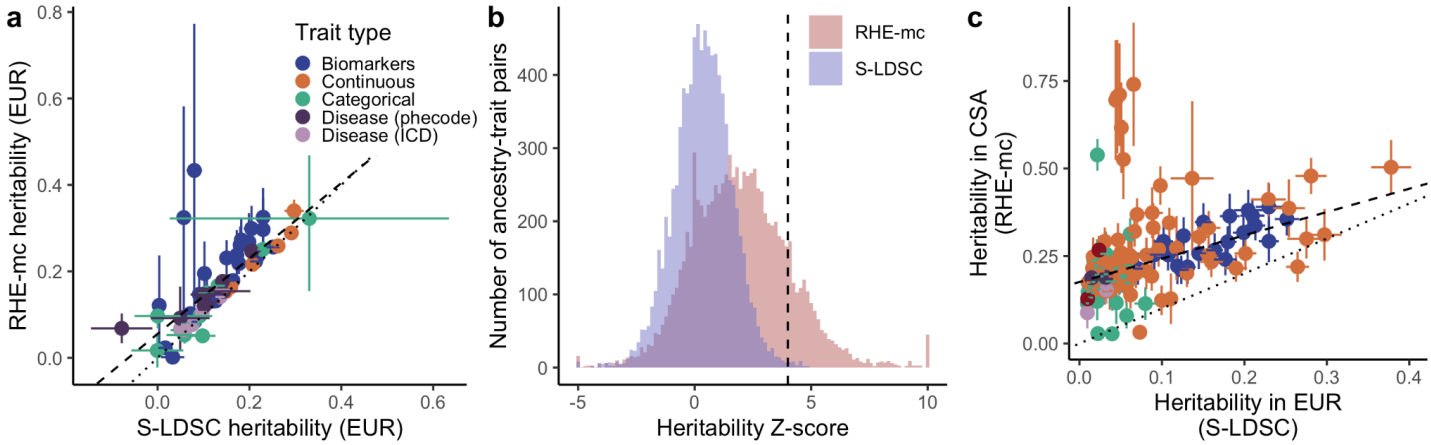
Extended Data Figure 1 | Global and subcontinental PCA. a, Global PCA projection of UKBB into PCs 1-2 defined by HGDP and 1000 Genomes Project reference panel, which are shown in colored dots on top of UKBB in black. **b**, Global PCA density plot of UKBB points only, excluding reference panel. **c**, Map of HGDP, 1000 Genomes Project, and AGVP reference used to define AFR PC space. **d**, PCs 1-2 within AFR. Inset shows density of UKBB samples assigned to AFR using a random forest. **c-d**, colors and shapes are consistent across panels. **e**, Map of HGDP and 1000 Genomes Project reference used to define CSA PC space. **f**, PCs 1-2 within CSA. Inset shows density of UKBB samples assigned to CSA using a random forest. **e-f**, colors and shapes are consistent across panels. **g**, Map of HGDP and 1000 Genomes Project reference used to define EAS PC space. **h**, PCs 1-2 within EAS. Inset shows density of UKBB samples assigned to EAS using a random forest. **g-h**, colors and shapes are consistent across panels. **i**, Map of HGDP and 1000 Genomes Project reference used to define EUR PC space. **j**, PCs 1-2 within EUR. Inset shows density of UKBB samples assigned to EUR using a random forest. **i-j**, colors and shapes are consistent across panels. **k**, Map of HGDP and 1000 Genomes Project reference used to define MID PC space. **l**, PCs 1-2 within MID. Inset shows density of UKBB samples assigned to MID using a random forest. **k-l**, colors and shapes are consistent across panels.

Extended Data Table 1 | Comparison of Lambda 1000 and Lambda GC for five phenotypes across three association study paradigms. Lambda GC is the genomic inflation factor, and Lambda 1000 is the genomic inflation factor for an equivalent study of 1,000 cases and 1,000 controls. We show these metrics for EUR (the European genetic ancestry group alone), mega-analysis (a single association test across all samples), and meta-analysis (across all available population-specific association results). The number of independent GWAS significant loci is computed with an LD r^2 cutoff of 0.1 and a GWAS significance threshold of 5×10^{-8}.

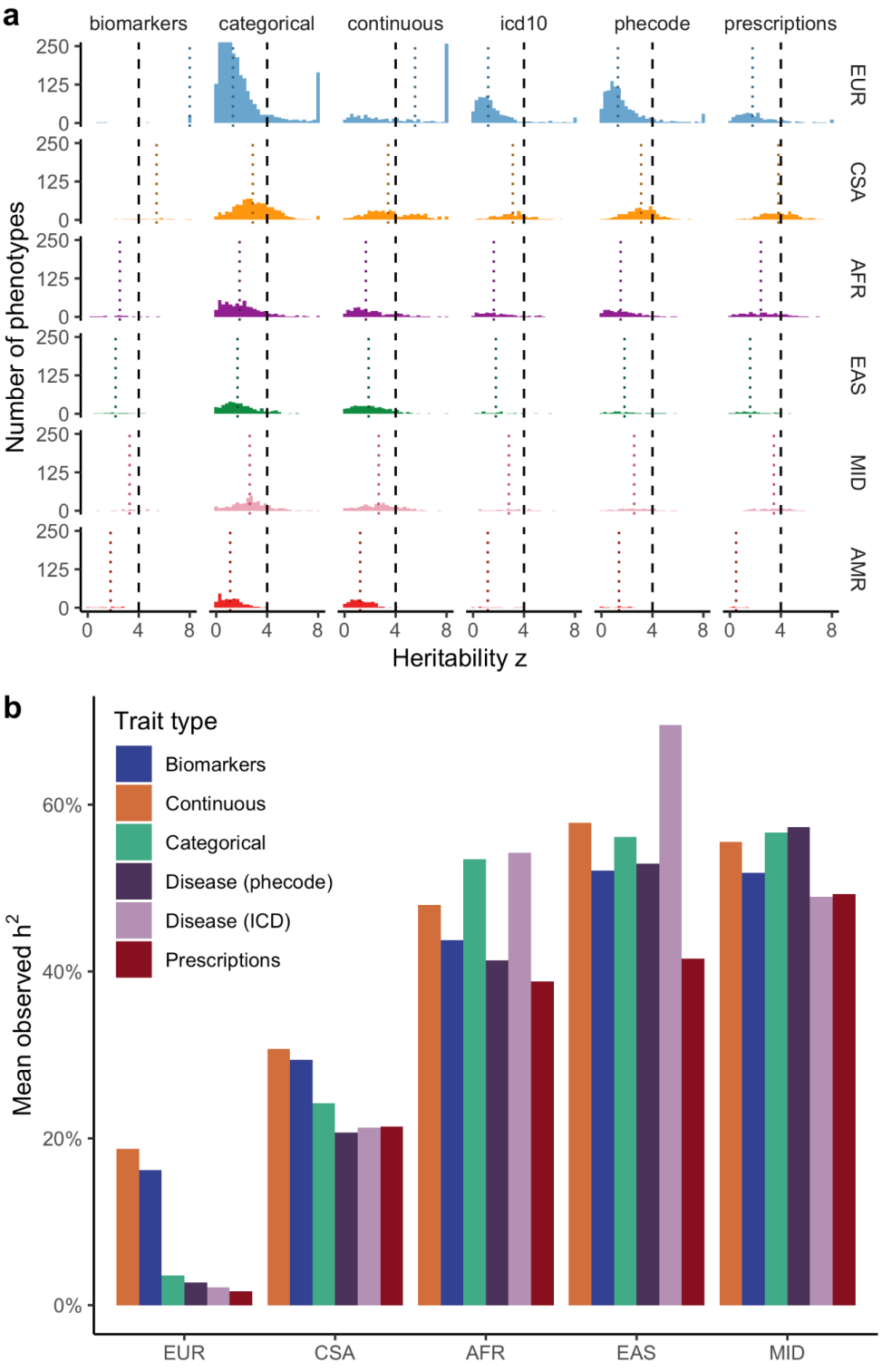
Phenocode		30000	30060	250.2	411	495
Phenotype		White blood cell count	Mean corpuscular haemoglobin concentration	Type 2 Diabetes	Ischemic Heart Disease	Asthma
Lambda 1000	EUR	1.0016	1.0010	1.0038	1.0018	1.0025
	Mega Analysis	1.0016	1.0014	1.0028	1.0023	1.0021
	Meta Analysis	1.0012	1.0006	1.0019	1.0012	1.0014
Lambda GC	EUR	1.3360	1.2037	1.1655	1.1225	1.1414
	Mega Analysis	1.3368	1.3009	1.1337	1.1650	1.1318
	Meta Analysis	1.2582	1.1327	1.0896	1.0867	1.0837
Number of independent GWAS significant loci	EUR	1,375	393	107	74	111
	Mega Analysis	1,580	523	128	86	116
	Meta Analysis	1,551	469	124	82	125

Extended Data Table 2 | Number of significant associations for height in AFR and CSA. The number of variants associated with height at $p < 10^{-4}$ for AFR and CSA are shown (Total), as well as filtered to independent SNPs based on distance-based windows (1 Mb). Despite the smaller sample size, the AFR ancestry group identified 15% more significant associations than CSA (1,261 vs 1,089).

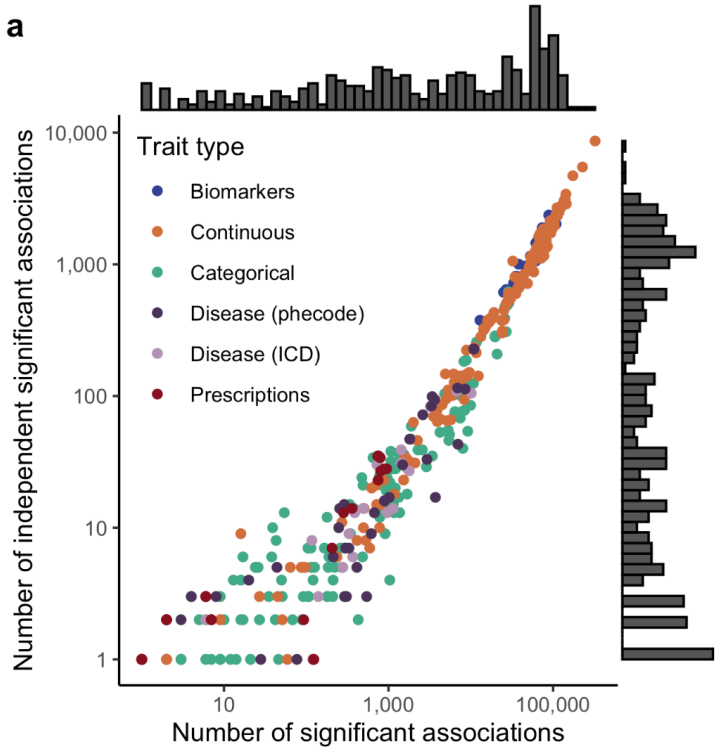
	AFR (N: 6,556)		CSA (N: 8,657)	
	Total	**Regionally-independent**	**Total**	**Regionally-independent**
Significant variants	4,682	1,380	4,284	1,212
High-quality significant variants	3,016	1,261	2,695	1,089



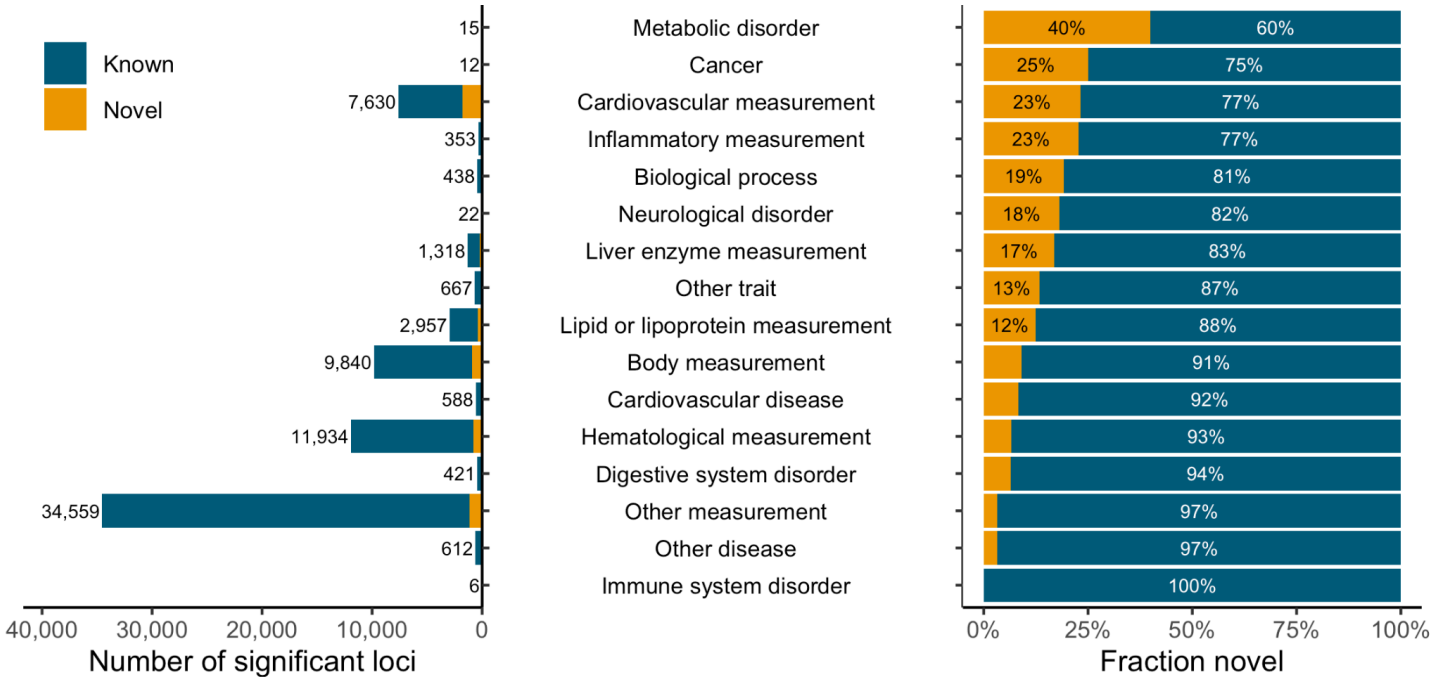
Extended Data Figure 2 | Heritability informs the robustness of GWAS across ancestry-trait pairs. a, heritability estimates are generally concordant in the EUR genetic ancestry group across 64 pilot phenotypes (**Supplementary Table 9**) and two statistical methods. RHE-mc uses a randomized multi-component version of classical Haseman-Elston regression with a genetic relatedness matrix (Pazokitoroudi et al. 2020), whereas S-LDSC uses GWAS summary statistics (Finucane et al. 2015). For binary phenotypes, heritability estimates are reported on the liability scale. All pilot phenotypes are shown, except for sepsis which had negative heritability estimates by both methods. The dotted line shows y=x, while the dashed line is a fitted linear regression (slope = 0.87, intercept = 0.05, $p = 7 \times 10^{-13}$). **b**, Across the same non-EUR ancestry-trait pairs, heritability estimated with RHE-mc have higher z-scores due to the smaller standard errors compared to S-LDSC. Dashed line at Z=4 was used as a QC filter. **c**, As in Fig. 2b, without filtering to phenotypes passing QC, but instead only filtering to EUR z > 4 and defined heritability in both genetic ancestry groups. Dotted line shows y = x and dashed line shows York regression fit (slope = 0.66, intercept = 0.17, $p < 10^{-100}$).



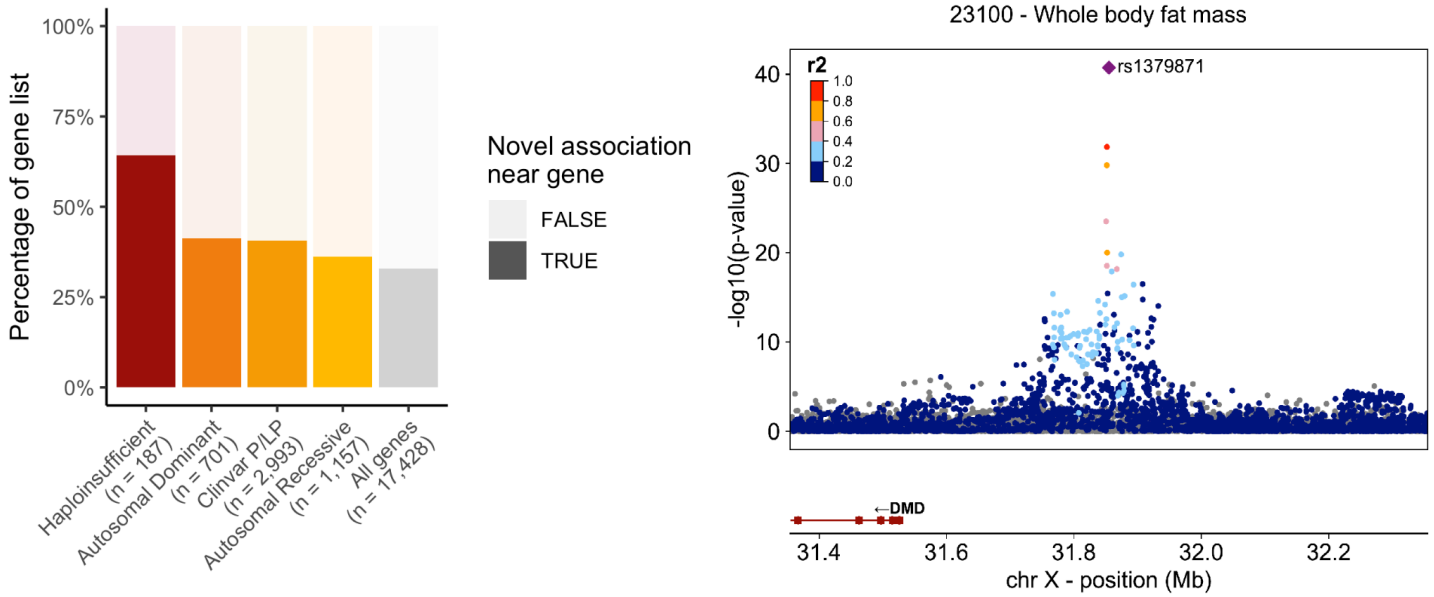
Extended Data Figure 3 | Heritability summaries across trait types and genetic ancestry groups. (a) The confidence metrics (heritability *z* score) across traits (columns) and ancestry groups (rows) are shown for the final heritability metrics used (S-LDSC for EUR, otherwise RHE-mc). Dashed line indicates inclusion criteria ($z \geq 4$). (**b**) The mean observed heritability (h^2) is plotted by ancestry group and trait type. For ancestry groups with smaller sample sizes, heritabilities are likely inflated due to a combination of winner's curse, as only significantly heritable phenotypes in each ancestry group are shown, and residual stratification.



Extended Data Figure 4 | Summary of associations per trait in Pan-UKB. The number of raw significant associations is shown, compared to the number of LD-independent (clumped) associations for each trait. Axes are log-scaled axes for increased resolution.

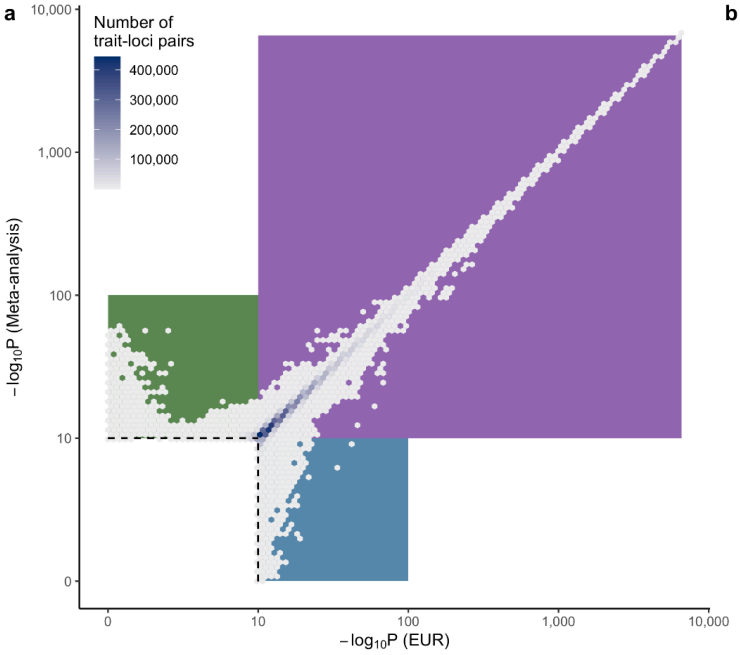


Extended Data Figure 5 | Improved identification of associations by EFO category. Number (left) and percentage (right) of known and novel variants identified in this study compared to the GWAS catalog across EFO categories.



Extended Data Figure 6 | GWAS hits near haploinsufficient genes. a, The percentage of novel associations by gene category. 66% of haploinsufficient genes have a novel significant hit nearby, compared to 34% of all genes. **b**, Locuszoom plots of a 1Mb region around rs1379871 (purple diamond; *DMD*), for whole body fat mass ($p = 1.84 \times 10^{-41}$; N = 431,792). The -log10(p-value) is plotted along chromosomal position, with neighboring variants colored by sample-size weighted LD (with lead SNP) for ancestries included in meta-analysis (gray: LD not defined for at least one ancestry group). This variant has recently been identified in a larger study of BMI (Zhang et al. 2023).

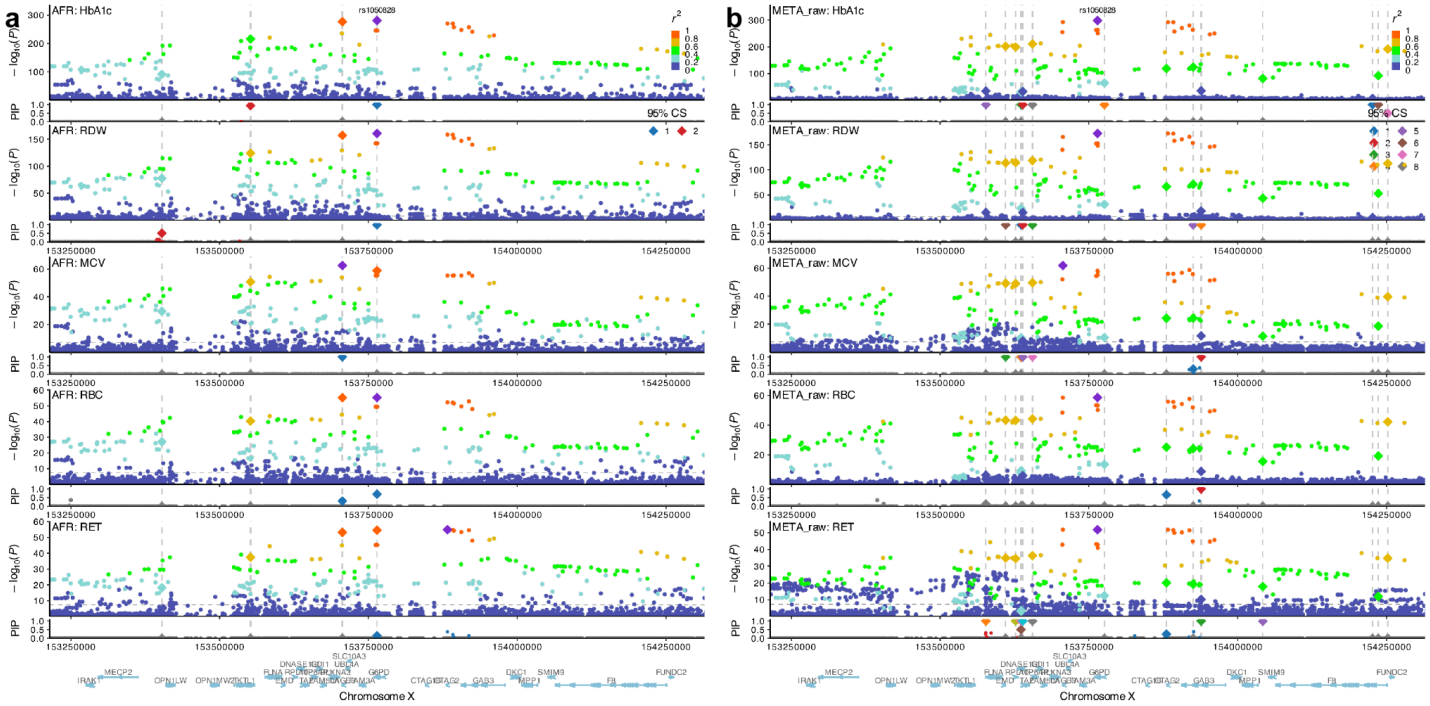
a



b

	Meta-only	Both	EUR-only
Number of variants (raw)	282101	7512124	298994
Percent heterogeneous	1.54	6.78	17.12
Percent low INFO score	2.33	3.11	3.06
Percent low quality	11.43	14.34	13.75
Number of variants (filtered)	246377	6009071	214829
Percent common in AFR (rare in EUR)	0.72	0.12	0.29
Percent common outside EUR (rare in EUR)	0.88	0.23	0.44

Extended Data Figure 7 | Comparison of meta-analysis and EUR summary statistics. (**a**) As in Fig. 3c, the p-value in EUR is plotted compared to the p-value in the meta-analysis, as a density plot to indicate the relative number of points in each region of the plot. Three quadrants are highlighted for significant in meta-analysis only (green), both meta-analysis and EUR (purple), and EUR-only (blue). (**b**) Summaries and meta-data of the variants in each of these three quadrants are shown. Heterogeneous is defined as Cochran's Q $p < 0.01$, low INFO score is defined as INFO < 0.9, and low quality is defined as failing quality filters from gnomAD or allele frequency significantly differing between gnomAD and Pan-UKB in at least one ancestry group (see Supplementary Information, QC of summary statistics). Common is defined as frequency > 1%.



Extended Data Figure 8 | Fine-mapping of the G6PD locus. Fine-mapping results for rs1050828 (*G6PD*) in (**a**) AFR and (**b**) meta-analysis. (**a**) AFR fine-mapping results highlight the missense variant (rs1050828) in a credible set, with a second independent signal for some phenotypes. (**b**) Meta-analysis fine-mapped results show instability as the major signal at rs1050828 is discovered in a group with a relatively small sample size, which results in a small contribution to the LD panel and thus, poor performance in fine-mapping.

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Abul-Husn, Noura S., and Eimear E. Kenny. 2019. “Personalized Medicine and the Power of Electronic Health Records.” *Cell* 177 (1): 58–69.

All of Us Research Program Genomics Investigators. 2024. “Genomic Data in the All of Us Research Program.” *Nature*, February. https://doi.org/10.1038/s41586-023-06957-x.

Asimit, Jennifer L., Konstantinos Hatzikotoulas, Mark McCarthy, Andrew P. Morris, and Eleftheria Zeggini. 2016. “Trans-Ethnic Study Design Approaches for Fine-Mapping.” *European Journal of Human Genetics: EJHG* 24 (9): 1330–36.

Atkinson, Elizabeth G., Adam X. Maihofer, Masahiro Kanai, Alicia R. Martin, Konrad J. Karczewski, Marcos L. Santoro, Jacob C. Ulirsch, et al. 2021. “Tractor Uses Local Ancestry to Enable the Inclusion of Admixed Individuals in GWAS and to Boost Power.” *Nature Genetics* 53 (2): 195–204.

Bamshad, Michael, Stephen Wooding, Benjamin A. Salisbury, and J. Claiborne Stephens. 2004. “Deconstructing the Relationship between Genetics and Race.” *Nature Reviews. Genetics* 5 (8): 598–609.

Bastarache, Lisa. 2021. “Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS,” July. https://doi.org/10.1146/annurev-biodatasci-122320-112352.

Bastarache, Lisa, Jacob J. Hughey, Scott Hebbring, Joy Marlo, Wanke Zhao, Wanting T. Ho, Sara L. Van Driest, et al. 2018. “Phenotype Risk Scores Identify Patients with Unrecognized Mendelian Disease Patterns.” *Science* 359 (6381): 1233–39.

Ben-Eghan, Chief, Rosie Sun, Jose Sergio Hleap, Alex Diaz-Papkovich, Hans Markus Munter, Audrey V. Grant, Charles Dupras, and Simon Gravel. 2020. “Don’t Ignore Genetic Data from Minority Populations.” *Nature*. https://doi.org/10.1038/d41586-020-02547-3.

Berry, Fred B., Matthew A. Lines, J. Martin Oas, Tim Footz, D. Alan Underhill, Philip J. Gage, and Michael A. Walter. 2006. “Functional Interactions between FOXC1 and PITX2 Underlie the Sensitivity to FOXC1 Gene Dose in Axenfeld-Rieger Syndrome and Anterior Segment Dysgenesis.” *Human Molecular Genetics* 15 (6): 905–19.

Bigdeli, Tim B., Giulio Genovese, Penelope Georgakopoulos, Jacquelyn L. Meyers, Roseann E. Peterson, Conrad O. Iyegbe, Helena Medeiros, et al. 2019. “Contributions of Common Genetic Variants to Risk of Schizophrenia among Individuals of African and Latino Ancestry.” *Molecular Psychiatry*, October. https://doi.org/10.1038/s41380-019-0517-y.

Bulik-Sullivan, Brendan, Hilary K. Finucane, Verneri Anttila, Alexander Gusev, Felix R. Day, Po-Ru Loh, ReproGen Consortium, et al. 2015. “An Atlas of Genetic Correlations across Human Diseases and Traits.” *Nature Genetics* 47 (11): 1236–41.

Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. “The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019.” *Nucleic Acids Research* 47 (D1): D1005–12.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. “The UK Biobank Resource with Deep Phenotyping and Genomic Data.” *Nature* 562 (7726): 203–9.

Chen, Ji, Cassandra N. Spracklen, Gaëlle Marenne, Arushi Varshney, Laura J. Corbin, Jian ’an Luan, Sara M. Willems, et al. 2021. “The Trans-Ancestral Genomic Architecture of Glycemic Traits.” *Nature Genetics* 53 (6): 840–60.

Cohen, Jonathan, Alexander Pertsemlidis, Ingrid K. Kotowski, Randall Graham, Christine Kim Garcia, and Helen H. Hobbs. 2005. “Low LDL Cholesterol in Individuals of African Descent Resulting from Frequent Nonsense Mutations in PCSK9.” *Nature Genetics* 37 (2): 161–65.

Conti, David V., Burcu F. Darst, Lilit C. Moss, Edward J. Saunders, Xin Sheng, Alisha Chou, Fredrick R. Schumacher, et al. 2021. “Trans-Ancestry Genome-Wide Association Meta-Analysis of Prostate Cancer Identifies New Susceptibility Loci and Informs Genetic Risk Prediction.” *Nature Genetics* 53 (1): 65–75.

COVID-19 Host Genetics Initiative. 2021. “Mapping the Human Genetic Architecture of COVID-19.” *Nature*, July. https://doi.org/10.1038/s41586-021-03767-x.

Denny, Joshua C., Lisa Bastarache, Marylyn D. Ritchie, Robert J. Carroll, Raquel Zink, Jonathan D. Mosley, Julie R. Field, et al. 2013. “Systematic Comparison of Phenome-Wide Association Study of Electronic Medical Record Data and Genome-Wide Association Study Data.” *Nature Biotechnology* 31 (12): 1102–10.

Ding, Yi, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, Kristin Boulier, Florian Privé, Bjarni J. Vilhjálmsson, Loes M. Olde Loohuis, and Bogdan Pasaniuc. 2023. “Polygenic Scoring Accuracy Varies across the Genetic Ancestry Continuum.” *Nature* 618 (7966): 774–81.

Finucane, Hilary K., Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, et al. 2015. “Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics.” *Nature Genetics* 47 (11): 1228–35.

Gage, P. J., H. Suh, and S. A. Camper. 1999. “Dosage Requirement of Pitx2 for Development of Multiple Organs.” *Development* 126 (20): 4643–51.

Genovese, Giulio, David J. Friedman, and Martin R. Pollak. 2013. “APOL1 Variants and Kidney Disease in People of Recent African Ancestry.” *Nature Reviews. Nephrology* 9 (4): 240–44.

Genovese, Giulio, David J. Friedman, Michael D. Ross, Laurence Lecordier, Pierrick Uzureau, Barry I. Freedman, Donald W. Bowden, et al. 2010. “Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans.” *Science* 329 (5993): 841–45.

Ghoussaini, Maya, Edward Mountjoy, Miguel Carmona, Gareth Peat, Ellen M. Schmidt, Andrew Hercules, Luca Fumis, et al. 2020. “Open Targets Genetics: Systematic Identification of Trait-Associated Genes Using Large-Scale Genetics and Functional Genomics.” *Nucleic Acids Research* 49 (D1): D1311–20.

Gibson, Greg. 2018. “Population Genetics and GWAS: A Primer.” *PLoS Biology*.

Graff, Mariaelisa, Anne E. Justice, Kristin L. Young, Eirini Marouli, Xinruo Zhang, Rebecca S. Fine, Elise Lim, et al. 2021. “Discovery and Fine-Mapping of Height Loci via High-Density Imputation of GWASs in Individuals of African Ancestry.” *American Journal of Human Genetics* 108 (4): 564–82.

Hou, Kangcheng, Yi Ding, Ziqi Xu, Yue Wu, Arjun Bhattacharya, Rachel Mester, Gillian M. Belbin, et al. 2023. “Causal Effects on Complex Traits Are Similar for Common Variants across Segments of Different Continental Ancestries within Admixed Individuals.” *Nature Genetics* 55 (4): 549–58.

Howrigan, D. 2017. “Details and Considerations of the UK Biobank GWAS.”

Huang, Hailiang, Ming Fang, Luke Jostins, Maša Umićević Mirkov, Gabrielle Boucher, Carl A. Anderson, Vibeke Andersen, et al. 2017. “Fine-Mapping Inflammatory Bowel Disease Loci to Single-Variant Resolution.” *Nature* 547 (7662): 173–78.

Lam, Max, Chia-Yen Chen, Zhiqiang Li, Alicia R. Martin, Julien Bryois, Xixian Ma, Helena Gaspar, et al. 2019. “Comparative Genetic Architectures of Schizophrenia in East Asian and European Populations.” *Nature Genetics*, November, 1–9.

Lewis, Anna C. F., Santiago J. Molina, Paul S. Appelbaum, Bege Dauda, Anna Di Rienzo, Agustin Fuentes, Stephanie M. Fullerton, et al. 2022. “Getting Genetic Ancestry Right for Science and Society.” *Science* 376 (6590): 250–52.

Li, Jun Z., Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, et al. 2008. “Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation.” *Science* 319 (5866): 1100–1104.

Liu, Dajiang J., and Suzanne M. Leal. 2012. “Estimating Genetic Effects and Quantifying Missing Heritability Explained by Identified Rare-Variant Associations.” *American Journal of Human Genetics* 91 (4): 585–96.

Liu, Zhanju, Ruize Liu, Han Gao, Seulgi Jung, Xiang Gao, Ruicong Sun, Xiaoming Liu, et al. 2023. “Genetic Architecture of the Inflammatory Bowel Diseases across East Asian and European Ancestries.” *Nature Genetics* 55 (5): 796–806.

Luo, Yang, Masahiro Kanai, Wanson Choi, Xinyi Li, Kenichi Yamamoto, Kotaro Ogawa, Maria Gutierrez-Arcelus, et al. 2020. “A High-Resolution HLA Reference Panel Capturing Global Population Diversity Enables Multi-Ethnic Fine-Mapping in HIV Host Response.” *medRxiv*. https://www.medrxiv.org/content/10.1101/2020.07.16.20155606v1.full-text.

Mägi, Reedik, Momoko Horikoshi, Tamar Sofer, Anubha Mahajan, Hidetoshi Kitajima, Nora Franceschini, Mark

I. McCarthy, COGENT-Kidney Consortium, T2D-GENES Consortium, and Andrew P. Morris. 2017. "Trans-Ethnic Meta-Regression of Genome-Wide Association Studies Accounting for Ancestry Increases Power for Discovery and Improves Fine-Mapping Resolution." *Human Molecular Genetics* 26 (18): 3639–50.

Mahajan, Anubha, Cassandra N. Spracklen, Weihua Zhang, Maggie C. Y. Ng, Lauren E. Petty, Hidetoshi Kitajima, Z. Yu Grace, et al. 2020. "Trans-Ancestry Genetic Study of Type 2 Diabetes Highlights the Power of Diverse Populations for Discovery and Translation." *medRxiv*. https://www.medrxiv.org/content/10.1101/2020.09.22.20198937v1.full-text.

Martin, Alicia R., Mark J. Daly, Elise B. Robinson, Steven E. Hyman, and Benjamin M. Neale. 2019. "Predicting Polygenic Risk of Psychiatric Disorders." *Biological Psychiatry* 86 (2): 97–109.

Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. 2017. "Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations." *American Journal of Human Genetics* 100 (4): 635–49.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. "Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities." *Nature Genetics* 51 (4): 584–91.

Mathieson, Iain, and Aylwyn Scally. 2020. "What Is Ancestry?" *PLoS Genetics* 16 (3): e1008624.

Meyer, Michelle N., Paul S. Appelbaum, Daniel J. Benjamin, Shawneequa L. Callier, Nathaniel Comfort, Dalton Conley, Jeremy Freese, et al. 2023. "Wrestling with Social and Behavioral Genomics: Risks, Potential Benefits, and Ethical Responsibility." *The Hastings Center Report* 53 Suppl 1 (Suppl 1): S2–49.

Miller, L. H., S. J. Mason, D. F. Clyde, and M. H. McGinniss. 1976. "The Resistance Factor to Plasmodium Vivax in Blacks. The Duffy-Blood-Group Genotype, FyFy." *The New England Journal of Medicine* 295 (6): 302–4.

Morales, Joannella, Danielle Welter, Emily H. Bowler, Maria Cerezo, Laura W. Harris, Aoife C. McMahon, Peggy Hall, et al. 2018. "A Standardized Framework for Representation of Ancestry Data in Genomics Studies, with Application to the NHGRI-EBI GWAS Catalog." *Genome Biology* 19 (1): 21.

National Academies of Sciences, Engineering, and Medicine. 2023. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. National Academies Press.

National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education, Health and Medicine Division, Committee on Population, Board on Health Sciences Policy, and Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research. 2023. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. National Academies Press.

Pazokitoroudi, Ali, Yue Wu, Kathryn S. Burch, Kangcheng Hou, Aaron Zhou, Bogdan Pasaniuc, and Sriram Sankararaman. 2020. "Efficient Variance Components Analysis across Millions of Genomes." *Nature Communications* 11 (1): 4020.

Polygenic Risk Score Task Force of the International Common Disease Alliance. 2021. "Responsible Use of Polygenic Risk Scores in the Clinic: Potential Benefits, Risks and Gaps." *Nature Medicine* 27 (11): 1876–84.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3): 559–75.

Rasooly, Danielle, Gina M. Peloso, Alexandre C. Pereira, Hesam Dashti, Claudia Giambartolomei, Eleanor Wheeler, Nay Aung, et al. 2023. "Genome-Wide Association Analysis and Mendelian Randomization Proteomics Identify Drug Targets for Heart Failure." *Nature Communications* 14 (1): 3826.

Ross, Michael J. 2019. "New Insights into APOL1 and Kidney Disease in African Children and Brazilians Living With End-Stage Kidney Disease." *Kidney International Reports* 4 (7): 908–10.

Sarnowski, Chloé, Aaron Leong, Laura M. Raffield, Peitao Wu, Paul S. de Vries, Daniel DiCorpo, Xiuqing Guo, et al. 2019. "Impact of Rare and Common Genetic Variants on Diabetes Diagnosis by Hemoglobin A1c in Multi-Ancestry Cohorts: The Trans-Omics for Precision Medicine Program." *American Journal of Human Genetics* 105 (4): 706–18.

Schaid, Daniel J., Wenan Chen, and Nicholas B. Larson. 2018. “From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping.” *Nature Reviews. Genetics* 19 (8): 491–504.

Scutari, Marco, Ian Mackay, and David Balding. 2016. “Using Genetic Distance to Infer the Accuracy of Genomic Prediction.” *PLoS Genetics* 12 (9): e1006288.

SIGMA Type 2 Diabetes Consortium, Karol Estrada, Ingvild Aukrust, Lise Bjørkhaug, Noël P. Burtt, Josep M. Mercader, Humberto García-Ortiz, et al. 2014. “Association of a Low-Frequency Variant in HNF1A with Type 2 Diabetes in a Latino Population.” *JAMA: The Journal of the American Medical Association* 311 (22): 2305–14.

Sinnott-Armstrong, Nasa, Yosuke Tanigawa, David Amar, Nina Mars, Christian Benner, Matthew Aguirre, Guhan Ram Venkataraman, et al. 2021. “Genetics of 35 Blood and Urine Biomarkers in the UK Biobank.” *Nature Genetics* 53 (2): 185–94.

Sirugo, Giorgio, Scott M. Williams, and Sarah A. Tishkoff. 2019. “The Missing Diversity in Human Genetic Studies.” *Cell* 177 (1): 26–31.

Solovieff, Nadia, Chris Cotsapas, Phil H. Lee, Shaun M. Purcell, and Jordan W. Smoller. 2013. “Pleiotropy in Complex Traits: Challenges and Strategies.” *Nature Reviews. Genetics* 14 (7): 483–95.

Tümer, Zeynep, and Daniella Bach-Holm. 2009. “Axenfeld-Rieger Syndrome and Spectrum of PITX2 and FOXC1 Mutations.” *European Journal of Human Genetics: EJHG* 17 (12): 1527–39.

Wang, Ying, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher, and Loic Yengo. 2020. “Theoretical and Empirical Quantification of the Accuracy of Polygenic Scores in Ancestry Divergent Populations.” *bioRxiv*. https://doi.org/10.1101/2020.01.14.905927.

Witherspoon, D. J., S. Wooding, A. R. Rogers, E. E. Marchani, W. S. Watkins, M. A. Batzer, and L. B. Jorde. 2007. “Genetic Similarities within and between Human Populations.” *Genetics* 176 (1): 351–59.

Zhang, Xinruo, Jennifer A. Brody, Mariaelisa Graff, Heather M. Highland, Nathalie Chami, Hanfei Xu, Zhe Wang, et al. 2023. “WHOLE GENOME SEQUENCING ANALYSIS OF BODY MASS INDEX IDENTIFIES NOVEL AFRICAN ANCESTRY-SPECIFIC RISK ALLELE.” *medRxiv : The Preprint Server for Health Sciences*, August. https://doi.org/10.1101/2023.08.21.23293271.

Zheng, Jie, A. Mesut Erzurumluoglu, Benjamin L. Elsworth, John P. Kemp, Laurence Howe, Philip C. Haycock, Gibran Hemani, et al. 2017. “LD Hub: A Centralized Database and Web Interface to Perform LD Score Regression That Maximizes the Potential of Summary Level GWAS Data for SNP Heritability and Genetic Correlation Analysis.” *Bioinformatics* 33 (2): 272–79.

Zhou, Wei, Masahiro Kanai, Kuan-Han H. Wu, Rasheed Humaira, Kristin Tsuo, Jibril B. Hirbo, Ying Wang, et al. 2021. “Global Biobank Meta-Analysis Initiative: Powering Genetic Discovery across Human Diseases.” *bioRxiv*. https://doi.org/10.1101/2021.11.19.21266436.

Zhou, Wei, Masahiro Kanai, Kuan-Han H. Wu, Humaira Rasheed, Kristin Tsuo, Jibril B. Hirbo, Ying Wang, et al. 2022. “Global Biobank Meta-Analysis Initiative: Powering Genetic Discovery across Human Disease.” *Cell Genomics* 2 (10): 100192.

Zhou, Wei, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, et al. 2018. “Efficiently Controlling for Case-Control Imbalance and Sample Relatedness in Large-Scale Genetic Association Studies.” *Nature Genetics* 50 (9): 1335–41.

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” Nature 526 (7571): 68–74.

Abul-Husn, Noura S., and Eimear E. Kenny. 2019. “Personalized Medicine and the Power of Electronic Health Records.” Cell 177 (1): 58–69.

Asimit, Jennifer L., Konstantinos Hatzikotoulas, Mark McCarthy, Andrew P. Morris, and Eleftheria Zeggini. 2016. “Trans-Ethnic Study Design Approaches for Fine-Mapping.” European Journal of Human Genetics: EJHG 24 (9): 1330–36.

Bamshad, Michael, Stephen Wooding, Benjamin A. Salisbury, and J. Claiborne Stephens. 2004. “Deconstructing the Relationship between Genetics and Race.” Nature Reviews. Genetics 5 (8): 598–609.

Bastarache, Lisa. 2021. “Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS,” July. https://doi.org/10.1146/annurev-biodatasci-122320-112352.

Bastarache, Lisa, Jacob J. Hughey, Scott Hebbring, Joy Marlo, Wanke Zhao, Wanting T. Ho, Sara L. Van Driest, et al. 2018. “Phenotype Risk Scores Identify Patients with Unrecognized Mendelian Disease Patterns.” Science 359 (6381): 1233–39.

Ben-Eghan, Chief, Rosie Sun, Jose Sergio Hleap, Alex Diaz-Papkovich, Hans Markus Munter, Audrey V. Grant, Charles Dupras, and Simon Gravel. 2020. “Don’t Ignore Genetic Data from Minority Populations.” Nature. https://doi.org/10.1038/d41586-020-02547-3.

Berry, Fred B., Matthew A. Lines, J. Martin Oas, Tim Footz, D. Alan Underhill, Philip J. Gage, and Michael A. Walter. 2006. “Functional Interactions between FOXC1 and PITX2 Underlie the Sensitivity to FOXC1 Gene Dose in Axenfeld-Rieger Syndrome and Anterior Segment Dysgenesis.” Human Molecular Genetics 15 (6): 905–19.

Bigdeli, Tim B., Giulio Genovese, Penelope Georgakopoulos, Jacquelyn L. Meyers, Roseann E. Peterson, Conrad O. Iyegbe, Helena Medeiros, et al. 2019. “Contributions of Common Genetic Variants to Risk of Schizophrenia among Individuals of African and Latino Ancestry.” Molecular Psychiatry, October. https://doi.org/10.1038/s41380-019-0517-y.

Bulik-Sullivan, Brendan, Hilary K. Finucane, Verneri Anttila, Alexander Gusev, Felix R. Day, Po-Ru Loh, ReproGen Consortium, et al. 2015. “An Atlas of Genetic Correlations across Human Diseases and Traits.” Nature Genetics 47 (11): 1236–41.

Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. “The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019.” Nucleic Acids Research 47 (D1): D1005–12.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. “The UK Biobank Resource with Deep Phenotyping and Genomic Data.” Nature 562 (7726): 203–9.

Chen, Ji, Cassandra N. Spracklen, Gaëlle Marenne, Arushi Varshney, Laura J. Corbin, Jian ’an Luan, Sara M. Willems, et al. 2021. “The Trans-Ancestral Genomic Architecture of Glycemic Traits.” Nature Genetics 53 (6): 840–60.

Cohen, Jonathan, Alexander Pertsemlidis, Ingrid K. Kotowski, Randall Graham, Christine Kim Garcia, and Helen H. Hobbs. 2005. “Low LDL Cholesterol in Individuals of African Descent Resulting from Frequent Nonsense Mutations in PCSK9.” Nature Genetics 37 (2): 161–65.

Conti, David V., Burcu F. Darst, Lilit C. Moss, Edward J. Saunders, Xin Sheng, Alisha Chou, Fredrick R. Schumacher, et al. 2021. “Trans-Ancestry Genome-Wide Association Meta-Analysis of Prostate Cancer Identifies New Susceptibility Loci and Informs Genetic Risk Prediction.” Nature Genetics 53 (1): 65–75.

COVID-19 Host Genetics Initiative. 2021. “Mapping the Human Genetic Architecture of COVID-19.” Nature, July. https://doi.org/10.1038/s41586-021-03767-x.

Denny, Joshua C., Lisa Bastarache, Marylyn D. Ritchie, Robert J. Carroll, Raquel Zink, Jonathan D. Mosley, Julie R. Field, et al. 2013. “Systematic Comparison of Phenome-Wide Association Study of Electronic Medical Record Data and Genome-Wide Association Study Data.” Nature Biotechnology 31 (12): 1102–10.

Finucane, Hilary K., Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, et al. 2015. “Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics.” Nature Genetics 47 (11): 1228–35.

Gage, P. J., H. Suh, and S. A. Camper. 1999. “Dosage Requirement of Pitx2 for Development of Multiple Organs.” Development 126 (20): 4643–51.

Genovese, Giulio, David J. Friedman, and Martin R. Pollak. 2013. “APOL1 Variants and Kidney Disease in People of Recent African Ancestry.” Nature Reviews. Nephrology 9 (4): 240–44.

Genovese, Giulio, David J. Friedman, Michael D. Ross, Laurence Lecordier, Pierrick Uzureau, Barry I. Freedman, Donald W. Bowden, et al. 2010. “Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans.” Science 329 (5993): 841–45.

Ghoussaini, Maya, Edward Mountjoy, Miguel Carmona, Gareth Peat, Ellen M. Schmidt, Andrew Hercules, Luca Fumis, et al. 2020. “Open Targets Genetics: Systematic Identification of Trait-Associated Genes Using Large-Scale Genetics and Functional Genomics.” Nucleic Acids Research 49 (D1): D1311–20.

Gibson, Greg. 2018. “Population Genetics and GWAS: A Primer.” PLoS Biology.

Graff, Mariaelisa, Anne E. Justice, Kristin L. Young, Eirini Marouli, Xinruo Zhang, Rebecca S. Fine, Elise Lim, et al. 2021. “Discovery and Fine-Mapping of Height Loci via High-Density Imputation of GWASs in Individuals of African Ancestry.” American Journal of Human Genetics 108 (4): 564–82.

Hou, Kangcheng, Yi Ding, Ziqi Xu, Yue Wu, Arjun Bhattacharya, Rachel Mester, Gillian M. Belbin, et al. 2023. “Causal Effects on Complex Traits Are Similar for Common Variants across Segments of Different Continental Ancestries within Admixed Individuals.” Nature Genetics 55 (4): 549–58.

Howrigan, D. 2017. “Details and Considerations of the UK Biobank GWAS.”

Huang, Hailiang, Ming Fang, Luke Jostins, Maša Umićević Mirkov, Gabrielle Boucher, Carl A. Anderson, Vibeke Andersen, et al. 2017. “Fine-Mapping Inflammatory Bowel Disease Loci to Single-Variant Resolution.” Nature 547 (7662): 173–78.

Lam, Max, Chia-Yen Chen, Zhiqiang Li, Alicia R. Martin, Julien Bryois, Xixian Ma, Helena Gaspar, et al. 2019. “Comparative Genetic Architectures of Schizophrenia in East Asian and European Populations.” Nature Genetics, November, 1–9.

Lewis, Anna C. F., Santiago J. Molina, Paul S. Appelbaum, Bege Dauda, Anna Di Rienzo, Agustin Fuentes, Stephanie M. Fullerton, et al. 2022. “Getting Genetic Ancestry Right for Science and Society.” Science 376 (6590): 250–52.

Li, Jun Z., Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, et al. 2008. “Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation.” Science 319 (5866): 1100–1104.

Liu, Dajiang J., and Suzanne M. Leal. 2012. “Estimating Genetic Effects and Quantifying Missing Heritability Explained by Identified Rare-Variant Associations.” American Journal of Human Genetics 91 (4): 585–96.

Liu, Zhanju, Ruize Liu, Han Gao, Seulgi Jung, Xiang Gao, Ruicong Sun, Xiaoming Liu, et al. 2023. “Genetic Architecture of the Inflammatory Bowel Diseases across East Asian and European Ancestries.” Nature Genetics 55 (5): 796–806.

Luo, Yang, Masahiro Kanai, Wanson Choi, Xinyi Li, Kenichi Yamamoto, Kotaro Ogawa, Maria Gutierrez-Arcelus, et al. 2020. “A High-Resolution HLA Reference Panel Capturing Global Population Diversity Enables Multi-Ethnic Fine-Mapping in HIV Host Response.” medRxiv. https://www.medrxiv.org/content/10.1101/2020.07.16.20155606v1.full-text.

Mägi, Reedik, Momoko Horikoshi, Tamar Sofer, Anubha Mahajan, Hidetoshi Kitajima, Nora Franceschini, Mark I. McCarthy, COGENT-Kidney Consortium, T2D-GENES Consortium, and Andrew P. Morris. 2017. “Trans-Ethnic Meta-Regression of Genome-Wide Association Studies Accounting for Ancestry Increases Power for Discovery and Improves Fine-Mapping Resolution.” Human Molecular Genetics 26 (18): 3639–50.

Mahajan, Anubha, Cassandra N. Spracklen, Weihua Zhang, Maggie C. Y. Ng, Lauren E. Petty, Hidetoshi Kitajima, Z. Yu Grace, et al. 2020. “Trans-Ancestry Genetic Study of Type 2 Diabetes Highlights the Power of Diverse Populations for Discovery and Translation.” medRxiv. https://www.medrxiv.org/content/10.1101/2020.09.22.20198937v1.full-text.

Martin, Alicia R., Mark J. Daly, Elise B. Robinson, Steven E. Hyman, and Benjamin M. Neale. 2019. “Predicting Polygenic Risk of Psychiatric Disorders.” Biological Psychiatry 86 (2): 97–109.

Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. 2017. “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.” American Journal of Human Genetics 100 (4): 635–49.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. “Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities.” Nature Genetics 51 (4): 584–91.

Mathieson, Iain, and Aylwyn Scally. 2020. “What Is Ancestry?” PLoS Genetics 16 (3): e1008624.

Meyer, Michelle N., Paul S. Appelbaum, Daniel J. Benjamin, Shawneequa L. Callier, Nathaniel Comfort, Dalton Conley, Jeremy Freese, et al. 2023. “Wrestling with Social and Behavioral Genomics: Risks, Potential Benefits, and Ethical Responsibility.” The Hastings Center Report 53 Suppl 1 (Suppl 1): S2–49.

Miller, L. H., S. J. Mason, D. F. Clyde, and M. H. McGinniss. 1976. “The Resistance Factor to Plasmodium

Vivax in Blacks. The Duffy-Blood-Group Genotype, FyFy." The New England Journal of Medicine 295 (6): 302–4.

Morales, Joannella, Danielle Welter, Emily H. Bowler, Maria Cerezo, Laura W. Harris, Aoife C. McMahon, Peggy Hall, et al. 2018. "A Standardized Framework for Representation of Ancestry Data in Genomics Studies, with Application to the NHGRI-EBI GWAS Catalog." Genome Biology 19 (1): 21.

National Academies of Sciences, Engineering, and Medicine. 2023. Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field. National Academies Press.

National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education, Health and Medicine Division, Committee on Population, Board on Health Sciences Policy, and Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research. 2023. Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field. National Academies Press.

Pazokitoroudi, Ali, Yue Wu, Kathryn S. Burch, Kangcheng Hou, Aaron Zhou, Bogdan Pasaniuc, and Sriram Sankararaman. 2020. "Efficient Variance Components Analysis across Millions of Genomes." Nature Communications 11 (1): 4020.

Polygenic Risk Score Task Force of the International Common Disease Alliance. 2021. "Responsible Use of Polygenic Risk Scores in the Clinic: Potential Benefits, Risks and Gaps." Nature Medicine 27 (11): 1876–84.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." American Journal of Human Genetics 81 (3): 559–75.

Rasooly, Danielle, Gina M. Peloso, Alexandre C. Pereira, Hesam Dashti, Claudia Giambartolomei, Eleanor Wheeler, Nay Aung, et al. 2023. "Genome-Wide Association Analysis and Mendelian Randomization Proteomics Identify Drug Targets for Heart Failure." Nature Communications 14 (1): 3826.

Ross, Michael J. 2019. "New Insights into APOL1 and Kidney Disease in African Children and Brazilians Living With End-Stage Kidney Disease." Kidney International Reports 4 (7): 908–10.

Sarnowski, Chloé, Aaron Leong, Laura M. Raffield, Peitao Wu, Paul S. de Vries, Daniel DiCorpo, Xiuqing Guo, et al. 2019. "Impact of Rare and Common Genetic Variants on Diabetes Diagnosis by Hemoglobin A1c in Multi-Ancestry Cohorts: The Trans-Omics for Precision Medicine Program." American Journal of Human Genetics 105 (4): 706–18.

Schaid, Daniel J., Wenan Chen, and Nicholas B. Larson. 2018. "From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping." Nature Reviews. Genetics 19 (8): 491–504.

Scutari, Marco, Ian Mackay, and David Balding. 2016. "Using Genetic Distance to Infer the Accuracy of Genomic Prediction." PLoS Genetics 12 (9): e1006288.

SIGMA Type 2 Diabetes Consortium, Karol Estrada, Ingvild Aukrust, Lise Bjørkhaug, Noël P. Burtt, Josep M. Mercader, Humberto García-Ortiz, et al. 2014. "Association of a Low-Frequency Variant in HNF1A with Type 2 Diabetes in a Latino Population." JAMA: The Journal of the American Medical Association 311 (22): 2305–14.

Sinnott-Armstrong, Nasa, Yosuke Tanigawa, David Amar, Nina Mars, Christian Benner, Matthew Aguirre, Guhan Ram Venkataraman, et al. 2021. "Genetics of 35 Blood and Urine Biomarkers in the UK Biobank." Nature Genetics 53 (2): 185–94.

Sirugo, Giorgio, Scott M. Williams, and Sarah A. Tishkoff. 2019. "The Missing Diversity in Human Genetic Studies." Cell 177 (1): 26–31.

Solovieff, Nadia, Chris Cotsapas, Phil H. Lee, Shaun M. Purcell, and Jordan W. Smoller. 2013. "Pleiotropy in Complex Traits: Challenges and Strategies." Nature Reviews. Genetics 14 (7): 483–95.

Tümer, Zeynep, and Daniella Bach-Holm. 2009. "Axenfeld-Rieger Syndrome and Spectrum of PITX2 and FOXC1 Mutations." European Journal of Human Genetics: EJHG 17 (12): 1527–39.

Wang, Ying, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher, and Loic Yengo. 2020. "Theoretical and Empirical Quantification of the Accuracy of Polygenic Scores in Ancestry Divergent Populations." bioRxiv. https://doi.org/10.1101/2020.01.14.905927.

Witherspoon, D. J., S. Wooding, A. R. Rogers, E. E. Marchani, W. S. Watkins, M. A. Batzer, and L. B. Jorde. 2007. "Genetic Similarities within and between Human Populations." Genetics 176 (1): 351–59.

Zhang, Xinruo, Jennifer A. Brody, Mariaelisa Graff, Heather M. Highland, Nathalie Chami, Hanfei Xu, Zhe Wang, et al. 2023. “WHOLE GENOME SEQUENCING ANALYSIS OF BODY MASS INDEX IDENTIFIES NOVEL AFRICAN ANCESTRY-SPECIFIC RISK ALLELE.” medRxiv : The Preprint Server for Health Sciences, August. https://doi.org/10.1101/2023.08.21.23293271.

Zheng, Jie, A. Mesut Erzurumluoglu, Benjamin L. Elsworth, John P. Kemp, Laurence Howe, Philip C. Haycock, Gibran Hemani, et al. 2017. “LD Hub: A Centralized Database and Web Interface to Perform LD Score Regression That Maximizes the Potential of Summary Level GWAS Data for SNP Heritability and Genetic Correlation Analysis.” Bioinformatics 33 (2): 272–79.

Zhou, Wei, Masahiro Kanai, Kuan-Han H. Wu, Rasheed Humaira, Kristin Tsuo, Jibril B. Hirbo, Ying Wang, et al. 2021. “Global Biobank Meta-Analysis Initiative: Powering Genetic Discovery across Human Diseases.” bioRxiv. https://doi.org/10.1101/2021.11.19.21266436.

Zhou, Wei, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, et al. 2018. “Efficiently Controlling for Case-Control Imbalance and Sample Relatedness in Large-Scale Genetic Association Studies.” Nature Genetics 50 (9): 1335–41.