nature genetics

ARTICLES

https://doi.org/10.1038/s41588-020-00757-z

Genetics of 35 blood and urine biomarkers in the UK Biobank

Nasa Sinnott-Armstrong[1,2,3,16] ✉, Yosuke Tanigawa[4,16] ✉, David Amar[4,5], Nina Mars[2], Christian Benner[2], Matthew Aguirre[4], Guhan Ram Venkataraman[4], Michael Wainberg[6], Hanna M. Ollila[2,7,8], Tuomo Kiiskinen[2,9], Aki S. Havulinna[2,9], James P. Pirruccello[10,11], Junyang Qian[12], Anna Shcherbina[2,5], FinnGen[*,**], Fatima Rodriguez[5], Themistocles L. Assimes[3,5], Vineeta Agarwala[5], Robert Tibshirani[4,12], Trevor Hastie[4,12], Samuli Ripatti[2,11,13], Jonathan K. Pritchard[1,14], Mark J. Daly[2,11,15] and Manuel A. Rivas[4,16] ✉

Clinical laboratory tests are a critical component of the continuum of care. We evaluate the genetic basis of 35 blood and urine laboratory measurements in the UK Biobank ($n = 363{,}228$ individuals). We identify 5,794 independent loci associated with at least one trait ($p < 5 \times 10^{-9}$), containing 3,374 fine-mapped associations and additional sets of large-effect (>0.1 s.d.) protein-altering, human leukocyte antigen (HLA) and copy number variant (CNV) associations. Through Mendelian randomization (MR) analysis, we discover 51 causal relationships, including previously known agonistic effects of urate on gout and cystatin C on stroke. Finally, we develop polygenic risk scores (PRSs) for each biomarker and build 'multi-PRS' models for diseases using 35 PRSs simultaneously, which improved chronic kidney disease, type 2 diabetes, gout and alcoholic cirrhosis genetic risk stratification in an independent dataset (FinnGen; $n = 135{,}500$) relative to single-disease PRSs. Together, our results delineate the genetic basis of biomarkers and their causal influences on diseases and improve genetic risk stratification for common diseases.

Serum and urine biomarkers are frequently measured to diagnose and monitor chronic disease conditions. Understanding the genetic predisposition to particular biomarker states and the factors that confound them may have implications for disease treatment. While the genetics of some biomarkers have been extensively studied, most notably lipids[1–3], glycemic traits[4–6] and measurements of kidney function[7–9], the genetic basis of most biomarkers has not been queried in large population-scale datasets.

To this end, the UK Biobank performed laboratory testing of >30 commonly measured biomarkers in serum and urine on a cohort of >480,000 individuals with extensive phenotypic and genome-wide genotypic data, including the unrelated individuals in this study (Supplementary Fig. 1)[10].

Here, we (1) performed a systematic analysis of the genetic architecture and detailed fine-mapping of biomarker-associated loci in 363,228 individuals including protein-altering variants (PAVs), protein-truncating variants (PTVs), non-coding variants, HLA variants and CNVs; (2) built phenome-wide associations for implicated genetic variants; (3) evaluated causal relationships between biomarkers and 40 medically relevant phenotypes; and (4) constructed polygenic prediction models (Fig. 1).

Results

Biomarker phenotype distributions. We first examined the consistency of the biomarker measurements[10,11]. After adjusting for statin usage (Supplementary Table 1a–c), we fit a regression model with multiple covariates (Methods). For each biomarker, we measured the proportion of phenotypic variance explained by these covariates; this ranged from 1.7% (rheumatoid factor) to 90% (testosterone), depending on the biomarker (Supplementary Fig. 2a–c and Supplementary Table 2). We evaluated body mass index as a confounder in associations, and there were minimal differences in genetic effects under this model (Supplementary Tables 3 and 4). Taking all the 35 laboratory phenotypes together, we recovered several previously estimated phenotypic correlations (Fig. 2)[12,13].

Genetics of biomarkers. We performed association analysis between directly genotyped and imputed autosomal genetic variants, CNVs and HLA allelotypes and 35 biomarkers in unrelated individuals in the UK Biobank across five population groups ($n = 318{,}953$ individuals for white British, 23,582 for non-British white, 6,019 for African, 7,338 for South Asian and 1,082 for East Asian), followed by meta-analysis of all but the East Asian

[1]Department of Genetics, School of Medicine, Stanford University, Stanford, CA, USA. [2]Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland. [3]VA Palo Alto Health Care System, Palo Alto, CA, USA. [4]Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, CA, USA. [5]Division of Cardiovascular Medicine and the Cardiovascular Institute, School of Medicine, Stanford University, Stanford, CA, USA. [6]Department of Computer Science, Stanford University, Stanford, CA, USA. [7]Department of Psychiatry and Behavioral Sciences, Stanford University, Palo Alto, CA, USA. [8]Center for Genomic Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [9]Finnish Institute for Health and Welfare, Helsinki, Finland. [10]Massachusetts General Hospital Division of Cardiology, Boston, MA, USA. [11]Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [12]Department of Statistics, Stanford University, Stanford, CA, USA. [13]Department of Public Health, Clinicum, University of Helsinki, Helsinki, Finland. [14]Department of Biology, Stanford University, Stanford, CA, USA. [15]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [16]These authors contributed equally: Nasa Sinnott-Armstrong, Yosuke Tanigawa, Manuel A. Rivas. *A list of authors and their affiliations appears at the end of the paper. **A full list of members and their affiliations appears in the Supplementary Information. ✉e-mail: nasa@stanford.edu; ytanigaw@stanford.edu; mrivas@stanford.edu

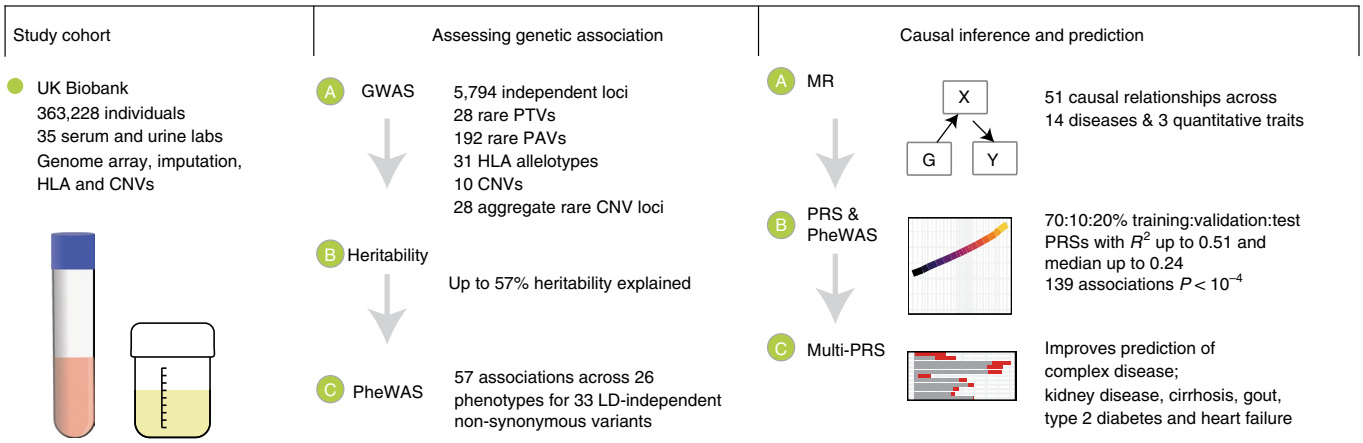

Fig. 1 | Schematic overview of the study. We prepared a dataset of 35 serum and urine biomarkers in the UK Biobank from 363,228 individuals. We analyzed the genetic basis of these biomarkers, assessed their relationship to medically relevant phenotypes and generated predictive models of disease outcomes from genome-wide data.

population (n in meta-analysis, 355,891; Methods, Fig. 2 and Supplementary Fig. 3). We stratified the genetic variants into three bins: (1) protein-truncating (27,816), (2) protein-altering (87,430) and (3) synonymous and non-coding variants (minor allele frequency (MAF) > 0.1% and estimated imputation quality (INFO) score > 0.3; imputed variants present in Haplotype Reference Consortium, 9,444,561 (ref. [14])) (Fig. 2). Comparison of effect sizes estimated across 42 other previously published study cohorts for 25 of the biomarkers showed overall high levels of agreement (Supplementary Fig. 4 and Supplementary Table 5). This was true when comparing to previous studies of lipids[1,2,15,16], glycemic traits[17,18], kidney function tests[19,20], liver function tests[17] and other biomarker measurements[21,22].

We adjusted the nominal association P values for multiple hypothesis testing and identified over 10,000 significant associations (Bonferroni-corrected meta-analysis $P < 5 \times 10^{-9}$ for assayed and imputed variants; Bonferroni-corrected $P < 1 \times 10^{-6}$ for non-rare (MAF > 0.1%) CNVs and CNV burden test for 23,598 genes; and Benjamini–Yekutieli (BY) adjusted $P < 0.05$ for HLA alleles; Methods, Supplementary Fig. 5 and Supplementary Tables 6–10). Linkage disequilibrium (LD) score intercepts for single-variant association results were between 0.999 and 1.137 for all 35 phenotypes, which is consistent with anthropometric traits in the UK Biobank and suggests that population structure in our analysis is well controlled[23] (Supplementary Table 11a).

Global and local heritability of biomarkers. To characterize the heritability of the 35 biomarkers, we first applied LD Score regression[24] to stratify heritability into ten tissue types and 53 general genomic features (for example, coding variants and regulatory variants) and further applied the Heritability Estimator from Summary Statistics (HESS)[25,26]. We found that both LD Score regression and HESS indicated common single-nucleotide polymorphisms (SNPs) that provide an explanation for the substantial heritability of some but not all biomarkers (0.6% (lipoprotein A, also referred to as lipoprotein(a)) to 23.9% (insulin-like growth factor (IGF)-1) using LD Score regression and 3.2% (microalbumin in urine) to 57% (total bilirubin) using HESS across the studied continuous phenotypes, Supplementary Table 11a,b). Estimates were lower for LD Score regression than for HESS for traits with lower polygenicity (for example, lipoprotein A, $h^2_{\text{LD Score}} = 0.6\%$ and $h^2_{\text{HESS}} = 24\%$), as LD Score regression estimates polygenic heritability[24]. We compared the polygenicity of all 35 biomarkers by computing the fraction of total SNP heritability attributable to loci by the top 1% of SNPs. We found that more than 50% of the SNP heritability for three biomarkers was explained by the top 1% of loci (lipoprotein A, 67.7%; total bilirubin, 60.9%; direct bilirubin, 57.5%), while the remaining 32 phenotypes showed patterns of moderate to high polygenicity (Supplementary Table 11b).

Associated variants prioritize therapeutic targets. We found 58 (43 rare, MAF < 1% and 55 not reported in the comparison study, Methods) PTV associations and 1,323 (306 rare, 1,079 not reported in comparison studies) PAV associations outside the major histocompatibility complex (MHC) region (hg19 chr6:25,477,797–36,448,354; meta-analyzed $P < 5 \times 10^{-9}$). We found 19 non-MHC PTV associations (17 rare (MAF < 1%)) with large estimated biomarker-lowering effects (>0.1 s.d.) and 26 (24 rare) with biomarker-raising effects (>0.1 s.d.) across 31 (27 rare) PTVs and at least one biomarker phenotype, for which the same PTV may have both increasing and decreasing associations across different biomarkers (Fig. 2 and Supplementary Table 6). Similarly, there were 240 (161 rare) and 182 (125 rare) non-MHC PAV associations with large estimated biomarker-lowering and -raising effects (>0.1 s.d.) across 241 (179 rare) PAVs and at least one biomarker phenotype, respectively (Fig. 2 and Supplementary Table 7). To assess whether the variants associated with biomarkers impact medically relevant phenotypes, we performed a phenome-wide association analysis (PheWAS) across 166 traits in the UK Biobank, compared our findings with previously published literature and sought independent replication in the FinnGen R2 cohort (Supplementary Tables 12 and 13 and Methods). We found 57 phenotypic associations (33 and 24 for increasing and decreasing disease risk, respectively) across 26 medically relevant phenotypes for two PTVs and 31 PAVs ($P < 1 \times 10^{-7}$), of which 31 associations were previously reported and 26 were new (Supplementary Table 13a and Methods).

For eight cardiovascular biomarkers (Supplementary Table 4a), we identified a stop-gain variant in *PDE3B* with documented protection against high cholesterol and a range of effects, including increasing levels of high-density lipoprotein (HDL) cholesterol and apolipoprotein A (ApoA; 0.40, 0.27 s.d.) and decreasing levels of triglycerides and apolipoprotein B (ApoB; 0.43, 0.27 s.d.)[2,27]; a stop-gain variant in *ANGPTL8*, with which we replicated a previously reported effect on HDL cholesterol (0.06 s.d. in our dataset) and discovered a triglyceride-lowering effect (0.06 s.d.)[28]; two PTVs in *LPA* with lowering effects on lipoprotein A levels (0.37, 0.42 s.d.),

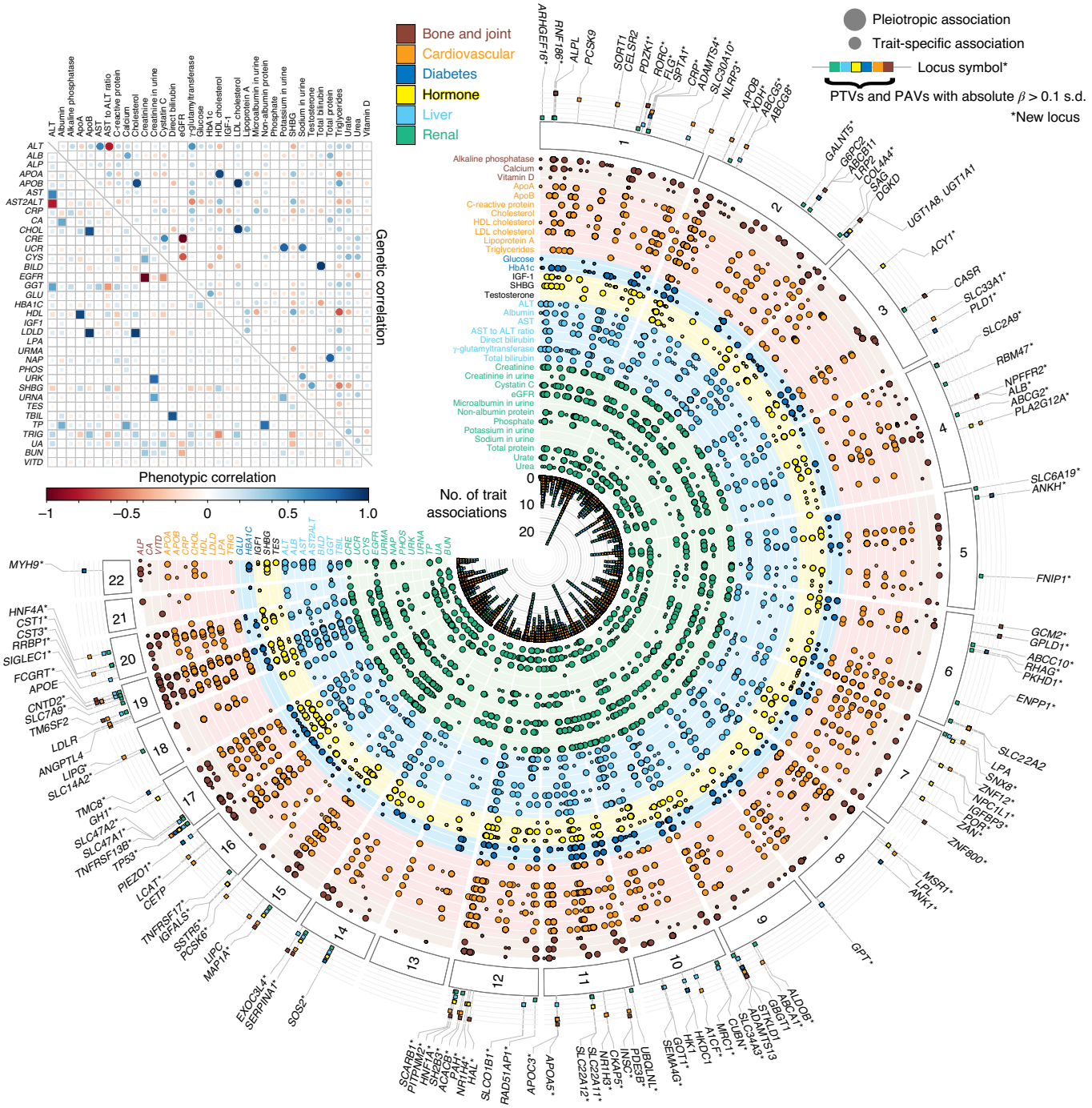

Fig. 2 | Genetics of 35 biomarkers. Top left inset, correlation plot of phenotypic (lower triangular matrix) and genetic (upper triangular matrix) effects between the 35 laboratory phenotypes, estimated using LD Score regression. The absolute heritability estimates with s.e.m. are in Supplementary Table 11a. Main panel, Fuji plot of lab phenotypes across the six categories provided by the UK Biobank, with genetic variant associations shown for LD-independent variants with meta-analysis, $P < 5 \times 10^{-9}$. Large-effect PTVs and PAVs (labeled when absolute $\beta \geq 0.1$ s.d.) are annotated with the displayed category of association (filled colored boxes) and highlighted if the loci were not previously reported in the comparison studies (Methods). Pleiotropic association and trait-specific association are shown by circles of different sizes. *P* values were from two-sided tests and were not corrected for multiple hypothesis testing.

of which one is known to be associated with decreased risk of coronary artery disease ($P = 3 \times 10^{-11}$; odds ratio (OR) = 0.89 (95% confidence interval (CI), 0.86–0.92))[29]; a missense allele (MAF = 0.2%) in *ACACB* associated with low-density lipoprotein (LDL), triglyceride, ApoB and alkaline phosphatase[30]; two independent missense alleles in *PLA2G12A* with increasing effects on levels of triglycerides, sex hormone-binding globulin (SHBG) and testosterone and lowering effects on HDL cholesterol, ApoA and hemoglobin (Hb)A1c levels (Supplementary Table 6); a splice region variant in *CPT1A* with lowering effects on triglyceride levels; and a missense variant in *PCSK6* with ApoB- and LDL-lowering effects (Supplementary Table 7).

For seven liver biomarkers (Supplementary Table 4a), we found an in-frame deletion (MAF = 0.05%) in *GOT1* with a lowering effect on aspartate aminotransferase (AST) levels (2.6 s.d.); a missense

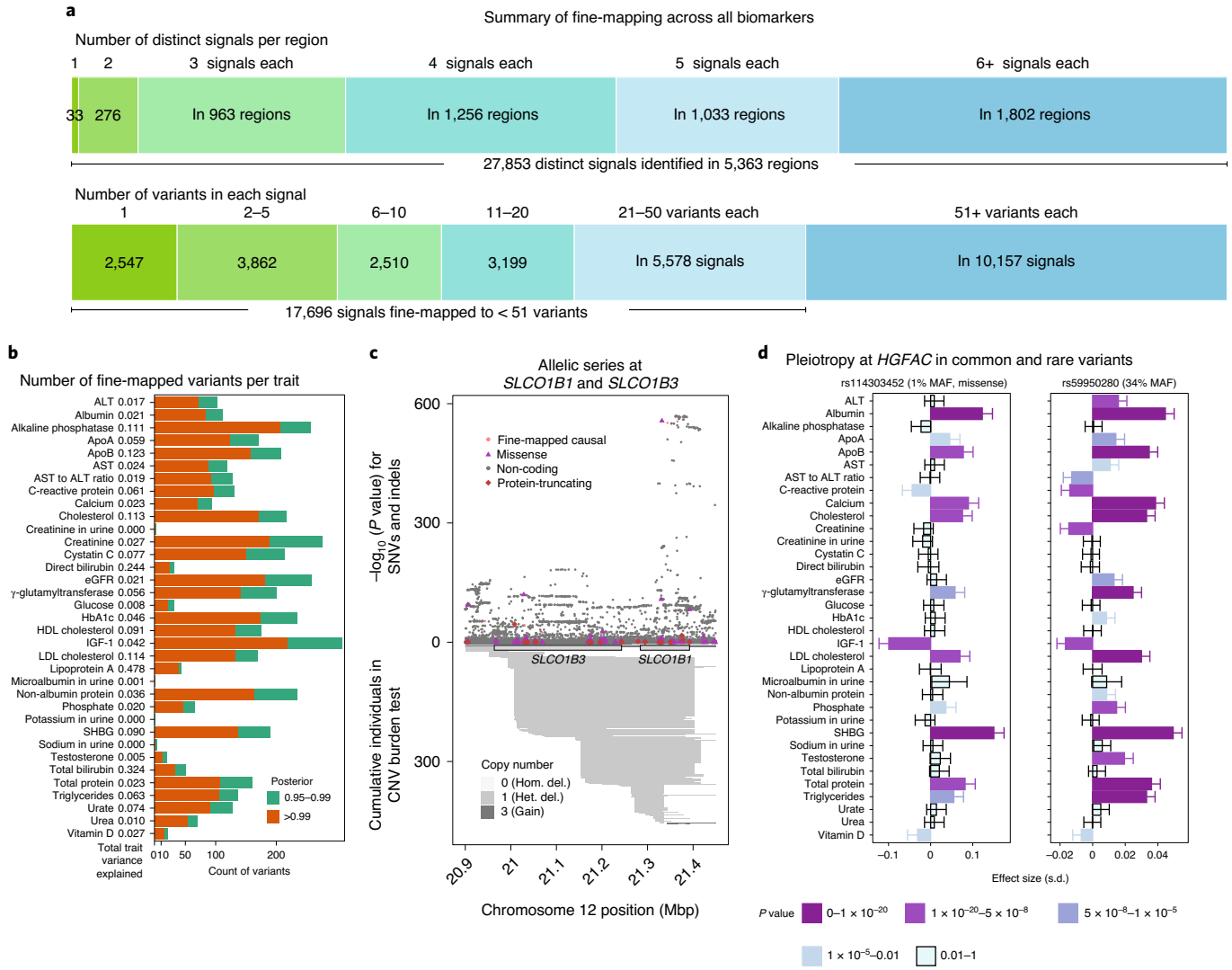

Fig. 3 | Summary of fine-mapped associations across 35 biomarker traits. a, FINEMAP analysis summary. Top, the number of identified distinct association signals (color gradient from green to blue) in each region with at least one genome-wide significant (UK Biobank meta-analysis $P < 5 \times 10^{-9}$) association, and the number of regions are shown, such as a single signal at 33 regions and 2–40 signals at 5,330 regions across 35 traits. Bottom, the number of identified candidate causal variants in the credible set with ≥99% posterior probability (color gradient from green to blue), and the number of signals are shown; for example, 2,547 signals were mapped to a single variant in the credible set across 35 traits. **b**, Breakdown of the number of fine-mapped associations with posterior probability greater than 0.95 or 0.99 across all biomarkers. Orange, posterior probability greater than 0.99; green, posterior probability between 0.95 and 0.99. The total variance explained for each trait is shown and is also detailed in Supplementary Table 14b. **c**, Allelic series showing combined missense variants, non-coding variants and rare CNVs with effects on total bilirubin levels at the *SLCO1B1* and *SLCO1B3* locus. CNVs are annotated below the axis, and SNPs and short indels are annotated above the axis. SNV, single-nucleotide variant; hom., homozygous; del., deletion; het., heterozygous. **d**, Pleiotropic mean effect estimates of fine-mapped rare coding (rs114303452, left) and common non-coding (rs59950280, right) variants at the *HGFAC* locus. Darker shades of purple indicate more significant associations. The *P* values were from two-sided tests and were not corrected for multiple hypothesis testing. The error bars represent s.d.

allele (MAF = 0.1%) in *SLC30A10* with increasing effects on alanine aminotransferase (ALT) and AST levels; four missense alleles in *GPT* with ALT-lowering effects; a missense variant in *ABCB4* associated with an increasing effect on ALT levels and an increased risk of gallstones in the UK Biobank ($P = 1.2 \times 10^{-8}$, OR = 1.38 (95% CI, 1.23–1.38)); an allelic series of three missense variants in *SERPINA1* with pleiotropic increasing effects on levels of albumin, AST, direct bilirubin and γ-glutamyltransferase and lowering effects on the AST to ALT ratio, and one of these missense alleles was associated with an increased risk of gallstones ($P = 8.1 \times 10^{-17}$, OR = 1.36 (95% CI, 1.27–1.47)) and cholecystitis ($P = 1.6 \times 10^{-8}$, OR = 1.26 (95% CI, 1.16–1.37)) in the UK Biobank; and two missense alleles in *DGKD*, with raising and lowering effects, respectively, on direct and total bilirubin levels (Supplementary Tables 7 and 13a).

For 12 renal biomarkers (Supplementary Table 4a), we found a PTV in *COL4A4* associated with an increasing effect on microalbumin levels in urine (0.77 s.d.) and with an increased risk of kidney disease ($P = 6.7 \times 10^{-13}$, OR = 6.9 (95% CI, 4.06–11.6)) in the UK Biobank, which is defined using a combination of hospital inpatient records (International Classification of Diseases (ICD)-10 code Q60 (renal agenesis and other reduction defects of kidney) and its subconcepts) and self-reported kidney diseases (coded as 1405 (other renal or kidney problem) in the UK Biobank)[31]; a frameshift variant in *SLC22A2* with strong lowering effects on the estimated

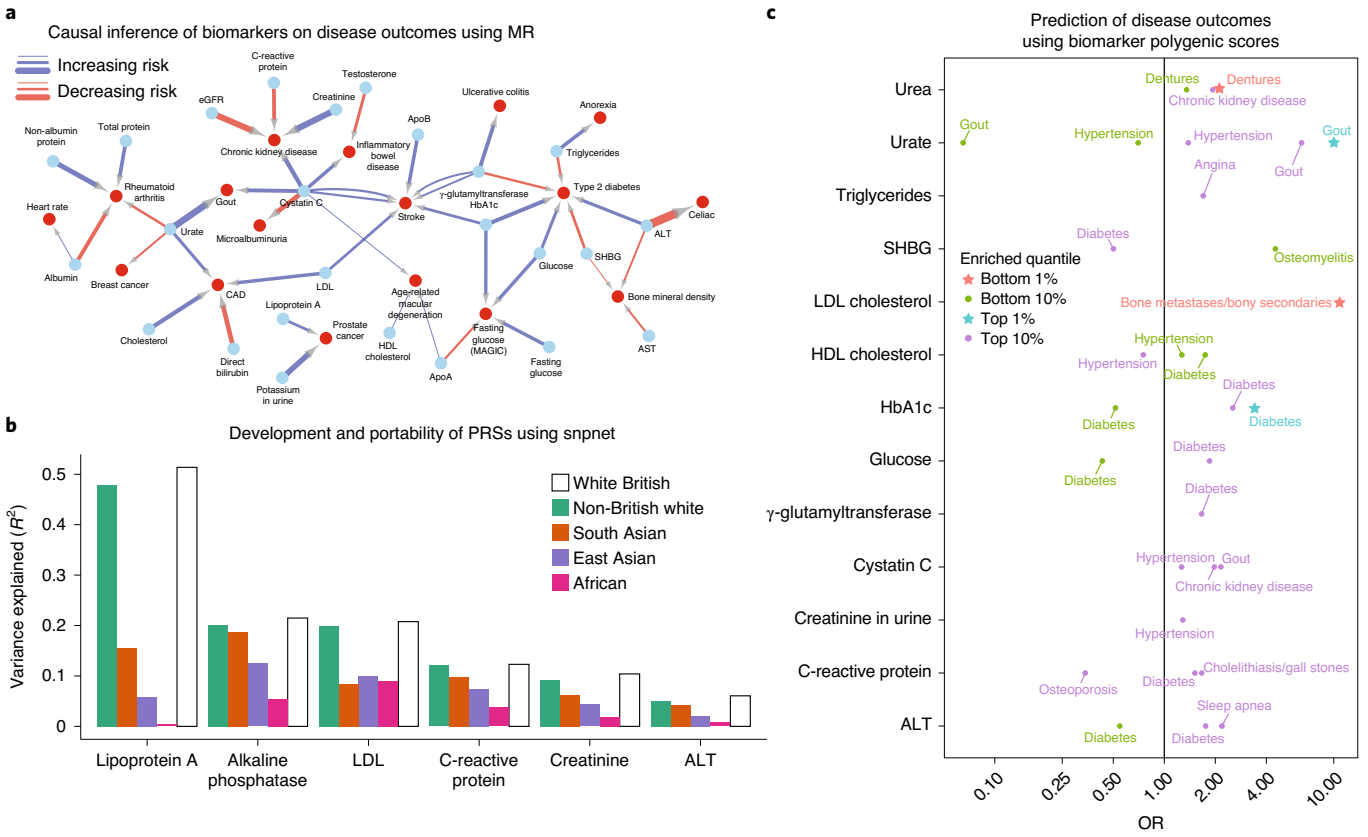

Fig. 4 | Causal inference, transferability of PRSs and complex trait association in polygenic risk tails. a, MR estimates causal links between biomarkers (blue nodes) and selected complex traits (red nodes). Association arrows were drawn based on effect direction (red, decreasing; blue, increasing). Associations were adjusted with a cutoff of 5% FDR across all tests (Methods and Supplementary Table 16). Edge width is proportional to the absolute causal effect size (log (odds per s.d.)). **b**, Summary of prediction accuracy of the snpnet polygenic scores across traits, evaluated on a held-out test set in the white British population, as well as in four other populations in the UK Biobank cohort. **c**, Biomarker PRSs (*x* axis) for the top 1%, top 10%, bottom 1% and bottom 10% of individuals and their associations with different diseases in the UK Biobank cohort, represented as the OR of the disease in each group relative to the 40–60% quantiles. Traits without rows did not have any outcomes with FDR-adjusted significant associations. CAD, coronary artery disease; MAGIC, Meta-Analyses of Glucose and Insulin-related traits Consortium.

glomerular filtration rate (eGFR) (0.52 s.d.) and an increasing effect on creatinine levels (0.52 sd); a stop-gain variant in *SLC22A11* with raising effects on urate levels (0.14 s.d.; Supplementary Tables 6 and 13a); a 0.1% rare missense allele in *SLC34A3* with strong eGFR- and phosphate-lowering effects and serum creatinine-, cystatin C- and urea-raising effects; and missense alleles in *SLC6A19*, *LRP2*, *ALDOB* and *SLC7A9* and two missense variants in *SLC25A45*, all associated with creatinine-lowering and eGFR-raising effects, among other examples (Supplementary Table 7). Notably, the majority of these genes are known to have high expression levels in renal tissue (https://www.gtexportal.org/home/).

For three bone and joint biomarkers (Supplementary Table 4a), we found an allelic series of two frameshift variants and a missense variant in *GPLD1* (ref. [32]), in addition to an allelic series of missense variants in *ALPL*. Similarly, we found an allelic series of variants in *CASR* that was associated with both calcium-increasing and -lowering effects (Supplementary Tables 6 and 7).

For glucose and HbA1c (biomarkers for diabetes, Supplementary Table 4a), we found a known missense variant association in *ANKH* (−0.11 and −0.17 s.d. for glucose and HbA1c, respectively), which also replicated the documented protective effects for diabetes ($P=1.2\times10^{-8}$, OR = 0.66 (95% CI, 0.57–0.76)). We also found a splice-donor variant in *RHAG* that was strongly associated with lower HbA1c levels (0.80 s.d.) and allelic series containing four missense variants each in *G6PC2* and *TMC8* (Fig. 2 and Supplementary Tables 6 and 7).

For three hormone biomarkers (Supplementary Table 4a), we found a PTV in *ADH1C*, *MSR1* and *NUBP2* affecting serum IGF-1 levels and an allelic series including the hepatocyte growth factor (HGF) genes *HGFAC*, *HGF* and *HNF4A* with effects on SHBG. Among these, we identified new associations with *HNF4A* alleles: a missense variant with MAF = 0.02% was associated with increased risk for diabetic eye disorders ($P=3.1\times10^{-8}$, OR = 9.60 (95% CI, 4.30–21.4)) and diabetes ($P=4.7\times10^{-8}$, OR = 3.8 (95% CI, 2.34–6.09)), and another missense variant (MAF = 3.1%) was associated with an increased risk for cholecystitis ($P=2.2\times10^{-13}$, OR = 1.27 (95% CI, 1.22–1.38) in the UK Biobank and was also replicated in FinnGen R2 data ($P=2.9\times10^{-17}$, OR = 1.46 (95% CI, 1.34–1.60)) (Fig. 2 and Supplementary Tables 6, 7 and 13a).

These results suggest that the genetic underpinning of biomarker levels could aid in prioritizing and better understanding the mechanisms of disease-associated variants.

CNVs and HLA allelotypes influencing biomarkers. CNVs constitute a substantial fraction of all base pair differences between individuals. We found 13 unique associations across ten individual CNVs (Bonferroni $P<1\times10^{-6}$, MAF > 0.01%; Supplementary Table 10a)[33]. We performed aggregate rare (MAF < 0.1%) CNV burden tests,

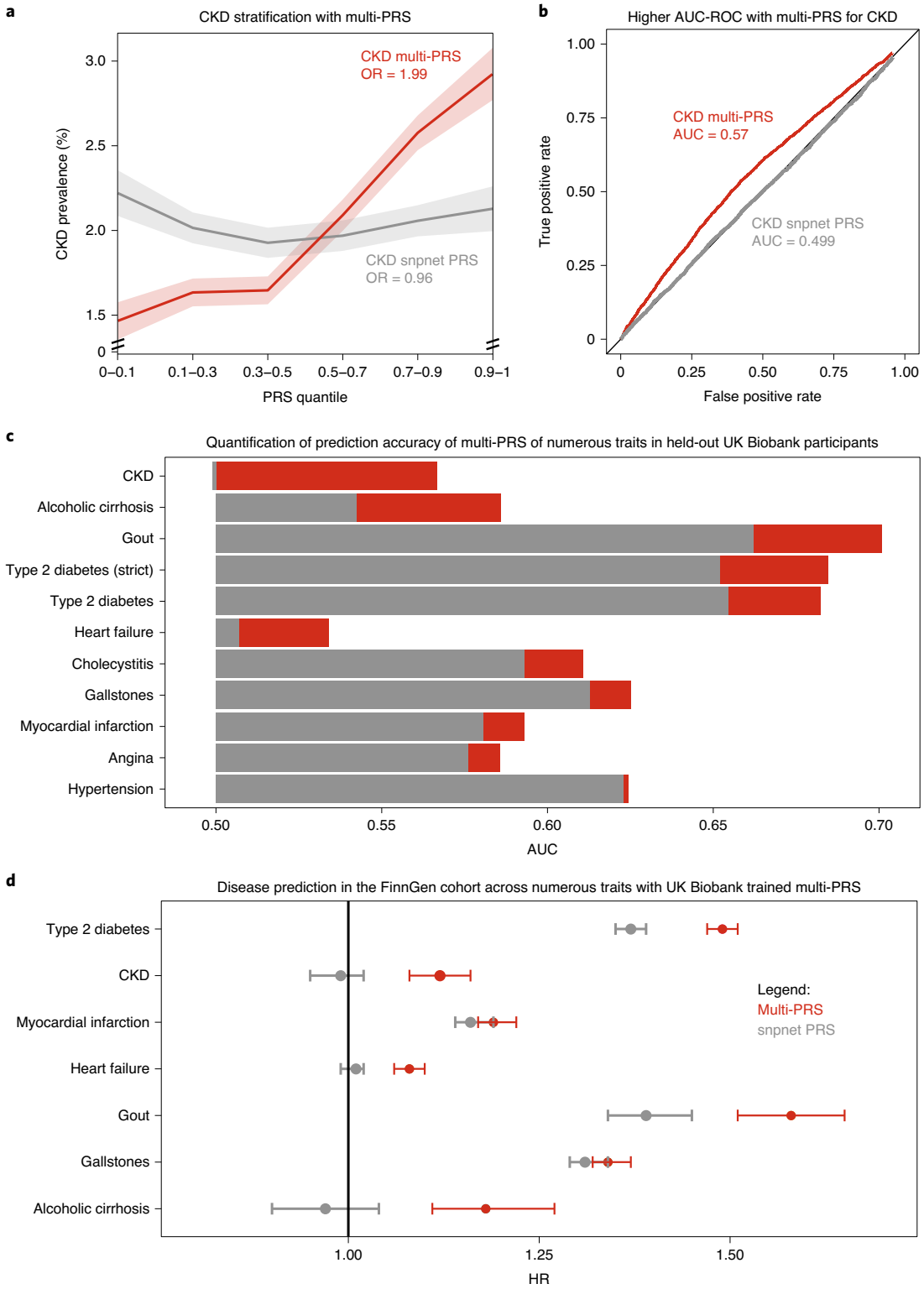

Fig. 5 | Multiple regression with biomarker polygenic scores improves prevalent and incident disease prediction. a, Quantiles of PRS (*x* axis), spaced to linearly represent the mean of the corresponding bin of scores. Prevalence of chronic kidney disease (CKD) (*y* axis; $n = 2{,}780$ cases and $n = 89{,}409$ total, defined by verbal questionnaires and hospital inpatient record ICD code data) within each quantile bin of the PRS. Error bars represent the s.e.m. around each measurement, and evaluated individuals were held-out individuals of European ancestry in the UK Biobank cohort. **b**, Receiver operating characteristic (ROC) curve with AUC for chronic kidney disease, comparing the snpnet-derived polygenic score to a multi-PRS model also trained across biomarkers. Evaluated individuals were held-out individuals of European ancestry in the UK Biobank cohort. **c**, Area under the ROC curve (AUC-ROC) estimates for the prediction of ten disease outcomes in a held-out test set of the UK Biobank. The diabetes analysis was performed using both a strict definition (excluding individuals with HbA1c > 39 from the control) and the complete sample (Methods). **d**, HRs for the incidence of type 2 diabetes ($n = 17{,}519$), chronic kidney disease ($n = 3{,}058$), myocardial infarction ($n = 7{,}913$), heart failure ($n = 13{,}965$), gout ($n = 1{,}936$), gallstones ($n = 11{,}629$) and cirrhosis ($n = 845$) in the FinnGen cohort using the standard single-disease PRSs trained on the UK Biobank dataset using snpnet versus the multi-PRSs, including both biomarker PRSs and the trait PRS. The strict definition of type 2 diabetes is shown. Error bars represent 95% CIs, and points represent HR estimates per s.d.

pooling CNVs in each gene for 23,598 genes. We found 28 gene-level associations (Bonferroni meta-analysis $P<1\times10^{-6}$; Supplementary Table 10a) including a burden of rare CNVs overlapping *HNF1B* that were associated with serum urea, eGFR, creatinine and cystatin C (the largest P value was 8.8×10^{-13} in white British subset analysis) and estimated to have large effects ($\beta=0.77$, -0.90, 0.93, 0.98 s.d., respectively; Supplementary Fig. 6a). Previous studies associated mutations in *HNF1B* with maturity onset diabetes of the young and altered kidney function[34]. The rare CNVs overlapping *HNF1B* were associated with chronic kidney disease ($P=1\times10^{-7}$; OR = 4.94, s.e.m. = 0.30; Supplementary Fig. 6a)[35,36] in a diabetes-dependent fashion (Supplementary Table 10b). We found a rare duplication in the *CST3* gene that was associated with increased levels of cystatin C, the protein that it encodes; this duplication had the opposite effect of a rare PTV at the same locus (Supplementary Fig. 6b). These results highlight the value of CNV analysis with potentially large effects on laboratory measurements.

To identify HLA allelotype associations that are not driven by pervasive LD structure in the *HLA* region, we applied Bayesian model averaging (Methods) to the significant allelotype–trait pairs (BY-adjusted P value, <0.05). We found 58 associations across 28 biomarker traits and allelotypes (Supplementary Table 9).

Fine-mapping of common associated variants. To nominate potentially causal variants at loci with common (MAF > 1%) variant associations, we performed fine-mapping analysis. Specifically, focusing on the summary statistics from white British individuals, we applied FINEMAP software[37,38]. From over 9,000 biomarker-associated loci, we identified 27,853 distinct signals in 5,363 regions across 35 traits. In the identified credible sets, 17,696 signals were fine-mapped to 50 or fewer variants with posterior probability of including the causal variant $P>0.99$; at 2,547 biomarker-associated loci, we resolved the signal to a single nominated causal variant (Fig. 3a and Supplementary Table 14a). Moreover, we identified 3,374 unique trait–variant associations with a posterior probability $P>0.99$ of being the causal variant (Fig. 3b and Supplementary Table 14a). These associations explained between 0% (urine potassium) and 48% (lipoprotein A) of the residual trait variance (Supplementary Table 14b).

Glycemic trait fine-mapping. We discovered fine-mapped associations for glycemic traits, including multiple variants at the *TGFB1* and *AXL* loci; rare missense variants in *PFN1* and *GYPC* (previously implicated in a small genome-wide association study (GWAS) of Mexican Americans)[39]; an intronic variant in a gene encoding a cytokine receptor (*IL6R*); a downstream variant at *VEGFA*[40]; a missense variant at *HFE*, the gene responsible for hemochromatosis[41,42]; and an intronic variant at *CD33* and a variant in the 3′ untranslated region of *CD36* (Supplementary Table 14a). *CD36* encodes a well-studied fatty acid receptor and biomarker for type 2 diabetes[43,44], and CD33 levels are known to be perturbed in type 2 diabetes cases[45].

Allelic series at the *SLCO1B* locus. We discovered several alleles implicated in the genetic control of bilirubin levels at the *SLCO1B* locus (Fig. 3c). We found several heterozygous deletion events and single-nucleotide variants that we fine-mapped to two main signals: a missense variant in *SLCO1B1* (rs34671512, marginal $\beta=-0.11$ s.d., $P=1.25\times10^{-95}$) and a non-coding association in an intron of *SLCO1B3* (rs11045598, marginal $\beta=0.076$ s.d., $P=1.31\times10^{-139}$). Despite the presence of two PTVs (one in *SLCO1B1* and one in *SLCO1B3*) at the locus, neither PTV had a conditionally independent effect on bilirubin levels. The diversity of variant types at these critical bilirubin and drug transporter loci suggests that large-effect loci can harbor variants with multiple independent genetic mechanisms that contribute to their trait associations.

HGFAC pleiotropy. We scanned for loci with large effects across multiple biomarkers. The most prominent of these was *HGFAC*, the gene encoding HGF activator. At this locus, we discovered two independent fine-mapped variants, rs114303452 (a missense variant with MAF = 1%) and rs59950280 (a non-coding variant with MAF = 34%). These two variants showed significant associations with a number of diverse biomarker traits, including lipids, IGF-1, albumin and calcium (Fig. 3d). In addition, rs114303452 was previously associated with serum HGF levels[46], supporting the role of HGFAC in controlling a number of other serum biomarkers through the regulation of HGF.

Targeted phenome-wide association analysis. We conducted PheWAS of the fine-mapped imputed variants across 166 UK Biobank phenotypes and identified 14 and 263 coding and non-coding associations, respectively, of which 109 were not previously reported in the literature ($P<10^{-7}$, Supplementary Tables 12 and 13b,c and Methods). For example, a common (MAF = 33%) intronic variant in *DPEP1* had protective effects against skin cancers (OR = 0.88 (95% CI, 0.86–0.90), 0.81 (0.77–0.84) and 0.89 (0.87–0.91) for skin cancer, malignant melanoma and non-melanoma skin cancer, respectively), a result that was replicated in the FinnGen R2 cohort ($P=3.1\times10^{-5}$, OR = 0.81 (95% CI, 0.74–0.90) for malignant neoplasm of skin). An allelic series of two intronic variants in *ABCG2* was identified that was associated with increasing and decreasing urate levels and had risk-increasing ($P=2.8\times10^{-67}$, OR = 1.38 (95% CI, 1.33–1.44)) and protective (OR = 0.72 (95% CI, 0.69–0.74)) associations with gout, respectively. Both of these associations with gout were also replicated in the FinnGen R2 cohort ($P=6.3\times10^{-6}$, OR = 1.25 (95% CI, 1.13–1.37) and $P=8.4\times10^{-5}$, OR = 0.84 (95% CI, 0.78–0.92)). These two variants ($r^2=0.47$) had low linkage with a known common PAV in *ABCG2* ($r^2=0.22$ and 0.11 in the UK Biobank white British population for rs2231142 (Q141K)), which contributes to risk of gout[47]. These results indicate that variants with effects on biomarkers may have pleiotropic effects across medically relevant phenotypes.

Causal inference. Given the relevance of several of the biomarkers studied to disease conditions, we estimated causal effects of biomarker levels on 40 medically relevant phenotypes (including 32 diseases; Supplementary Table 15) using two-sample MR with the genome-wide significant variants for each biomarker as instrumental variables[48–51] (Methods). We identified 51 significant causal relationships at a false discovery rate (FDR) of 5% (Fig. 4a and Supplementary Table 16). Many of these and their causal effects are well described. We found genetic evidence supporting the protective effect of SHBG on diabetes (0.7 OR per s.d.), which is consistent with existing reports[52,53]; the effect of ApoA on fasting glucose levels (0.84 OR per s.d., FDR-adjusted $P=0.02$)[54–56]; and a risk effect of ALT on diabetes (1.53 OR per s.d., FDR-adjusted $P=0.0018$)[57,58]. There was a consistent effect of cystatin C on stroke risk (1.2 OR per s.d., FDR-adjusted $P=8.7\times10^{-4}$ for any stroke and 1.21 OR per s.d., FDR-adjusted $P=2.8\times10^{-3}$ for ischemic stroke)[59,60]. Finally, both HDL and ApoA were associated with increased risk of age-related macular degeneration[61], as was cystatin C[62–64].

We estimated a causal protective effect of testosterone on inflammatory bowel disease (0.70 OR per s.d., FDR-adjusted $P=3.86\times10^{-3}$)[65,66] and a protective effect of urate on breast cancer risk (0.87 OR per s.d., FDR-adjusted $P=0.033$)[67].

Polygenic prediction of biomarkers. The vast size of the UK Biobank cohort affords the opportunity to build predictive polygenic risk models of biomarkers from genotypic data alone[68]. We constructed PRSs for all 35 biomarkers using batch screening iterative lasso (BASIL) implemented in the R package snpnet[69,70]. Specifically, we split the white British individuals into 70% train-

ing, 10% validation (to identify the optimal sparsity parameter) and 20% test sets and evaluated the predictive performance (R^2) in the held-out test set ($n=63,818$) as well as in four populations in the UK Biobank cohort (Methods). We found that the mean predictive performances relative to the white British test set for these four populations were 93%, 70%, 51% and 24%, respectively, suggesting that these polygenic models have limited generalizability across populations (Fig. 4b, Supplementary Fig. 7 and Supplementary Table 17)[71]. In an external validation, we found that the PRSs had high portability to self-identified white individuals from the MESA cohort (Supplementary Table 18)[72].

Multiple regression with PRSs. We hypothesized that the 35 biomarker PRSs might improve the prediction of higher-level traits and diseases in combination with the PRS for the trait or disease itself. To this end, we constructed multi-PRS models for traits by using multiple regression analysis to predict the trait or disease from (1) its own PRS, (2) the PRSs for each of the 35 biomarkers and (3) the covariates age, sex and principal components (Methods).

We selected disease endpoints for multi-PRS analysis by considering the enrichment of disease prevalence at the tails of the distribution of the single-trait biomarker PRSs (Fig. 4c, Supplementary Table 19 and Supplementary Fig. 8). We focused on traits with three or more associated biomarkers (Supplementary Fig. 9), as we reasoned that these would benefit most from the combination of multiple biomarker PRSs.

For chronic kidney disease, the multi-PRS analysis stratified individuals according to disease status better than did the snpnet PRS analysis (Fig. 5a,b and Supplementary Table 20). In contrast, the myocardial infarction snpnet PRS analysis was equally stratifying as compared to the multi-PRS analysis, with both explaining a substantial portion of trait heritability (area under the curve (AUC), 0.58–0.59; Fig. 5c). This trend held after including additional existing polygenic scores for type 2 diabetes and also for myocardial infarction (AUC, 0.594 and 0.611 respectively; Supplementary Fig. 10 and Supplementary Tables 20–23). This suggests that the genetic basis of myocardial infarction, as previously reported[73], already captures the majority of the genetic component of serum lipids and other biomarkers. Similar weak effects of biomarkers were estimated for hypertension, angina and gallstones, while alcoholic cirrhosis, gout, type 2 diabetes and heart failure were better predicted with multi-PRS models (Fig. 5c and Supplementary Table 20). Improved predictions relied on relevant and variable biomarkers across these traits (Supplementary Fig. 11), including eGFR, creatinine, cystatin C and bilirubin for chronic kidney disease; creatinine, bilirubin, total and LDL cholesterol, cystatin C and eGFR for heart failure; and bilirubin, GGT, eGFR, HDL cholesterol and IGF-1 for alcoholic cirrhosis.

Encouraged by these findings, we evaluated the potential of these improved polygenic scores to identify disease cases by applying both trait-specific PRSs and combined PRSs in an independent replication cohort, FinnGen (R3, $n=135,500$, Supplementary Tables 24–26). Here, we found evidence that the combination of PRSs increased the effect size in chronic kidney disease (hazard ratio (HR) = 0.99, $P=0.46$ for snpnet PRS and HR = 1.12, $P=2.09\times10^{-10}$ for multi-PRS; Fig. 5d, Supplementary Fig. 12 and Supplementary Table 24), type 2 diabetes (HR = 1.37, $P<2\times10^{-16}$ for snpnet PRS and HR = 1.49 for multi-PRS), gout (HR = 1.39, $P<2\times10^{-16}$ for snpnet PRS and HR = 1.58 for multi-PRS), heart failure (HR = 1.01, $P=0.38$ for snpnet PRS and HR = 1.08, $P<2\times10^{-16}$ for multi-PRS) and alcoholic cirrhosis (HR = 0.97, $P=0.35$ for snpnet PRS and HR = 1.18, $P=1.04\times10^{-6}$ for multi-PRS; Supplementary Table 24). Results similar to those from the UK Biobank were found in models including existing polygenic scores (Supplementary Table 26 and Supplementary Fig. 13) with the integrated type 2 diabetes model, including both pre-existing PRSs and biomarker PRSs, resulting in an HR change of 1.67 per s.d. This suggests that multiple regression of polygenic risk for biomarkers might capture multiple underlying disease states and/or underlying causes, and that these multiple states are predictive of disease.

Discussion

Using data from 35 biomarkers in ~363,000 UK Biobank samples, we provide an assessment of genetic associations with biomarker levels, the relevance of these associations in disease phenotypes and their utility in risk stratification.

PAVs that modify biomarker levels and disease risk can provide in vivo validation of therapeutic targets[74,75]. Here, we found multiple PAVs that directly implicate genes associated with the studied biomarkers, and we hypothesize that some of these genes may provide potential therapeutic targets.

To assess the translatability of our findings, we built predictive models aggregating trait PRSs with those of the biomarkers, improving the predictive accuracy of multiple disease outcomes both overall and especially at the extremes of genetic risk. Given that biomarker values are already routinely collected in structured data formats, we anticipate that multi-PRS methods could inform clinical practice in the coming years, as a larger fraction of the population is genotyped and sequenced.

In addition to the discovery of multiple individual loci and candidate causal variants, we can also draw some general conclusions across the traits evaluated with our multi-PRS models. Traits and diseases were predicted best when they had individually predictive biomarkers and a complex etiology (for example, chronic kidney disease) but underpowered genetic studies. We believe that a large number of disease cases is typically most useful in developing well-powered models, as it helps both with the baseline polygenic score and fitting of the multi-PRS components. Further exploration of the conditions under which multi-PRS models perform particularly well is an area of future study. Numerous limitations to this work are present. We assigned individuals to ancestry groups based on self-reported ancestry categories and the top two principal components of the genotype matrix. We included many technical covariates to reduce bias in the measurements of the biomarkers, but doing so has the potential to reduce power. We fine-mapped based on imputed genotypes and summary statistics, and both of these reduce power to detect true causal variants. In addition, the large and complex linkage present at some loci, including notably the *LPA* locus, might result in spurious fine-mapped and rare coding variant associations, although conditional analyses (for example, of a rare coding variant in *SLC22A2*) are inconclusive (Supplementary Fig. 14). Similarly, causal inference using individual-level data[76] can increase power and reduce bias, and we recommend it for future studies. Lastly, we anticipate that including other genetic risk scores will fit well into the multi-PRS framework to further improve prediction of common complex diseases.

The genome-wide resource made available with this study, including the association summary statistics, fine-mapped regions and polygenic prediction models (Data availability)[77], provides a starting point for causal mapping of genetic variants affecting the 35 biomarkers and their relevance to medical phenotypes. These results highlight the benefits of direct measurements of biomarkers in population cohorts for interpreting the genetic basis of biomarkers and improved prediction of multiple common diseases.

Online content

Any methods, additional references, Nature Research 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/s41588-020-00757-z.

Received: 15 July 2020; Accepted: 1 December 2020;
Published online: 18 January 2021

References

1. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
2. Klarin, D. et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
3. Liu, D. J. et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
4. Wheeler, E. et al. Impact of common genetic determinants of hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: a transethnic genome-wide meta-analysis. *PLoS Med.* **14**, e1002383 (2017).
5. Spracklen, C. N. et al. Identification and functional analysis of glycemic trait loci in the China Health and Nutrition Survey. *PLoS Genet.* **14**, e1007275 (2018).
6. Scott, R. A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
7. Graham, S. E. et al. Sex-specific and pleiotropic effects underlying kidney function identified from GWAS meta-analysis. *Nat. Commun.* **10**, 1847 (2019).
8. Wuttke, M. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
9. Okada, Y. et al. Meta-analysis identifies multiple loci associated with kidney function-related traits in East Asian populations. *Nat. Genet.* **44**, 904–909 (2012).
10. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
11. Fry, D., Almond, R., Moffat, S., Gordon, M. & Singh, P. *UK Biobank Biomarker Project: Companion Document to Accompany Serum Biomarker Data* (UK Biobank Document Showcase, 2019); https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/serum_biochemistry.pdf
12. Kathiresan, S. et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med. Genet.* **8**, S17 (2007).
13. Snell-Bergeon, J. K. et al. Evaluation of urinary biomarkers for coronary artery disease, diabetes, and diabetic kidney disease. *Diabetes Technol. Ther.* **11**, 1–9 (2009).
14. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
15. Lu, X. et al. Exome chip meta-analysis identifies novel loci and East Asian-specific coding variants that contribute to lipid levels and coronary artery disease. *Nat. Genet.* **49**, 1722–1730 (2017).
16. Kettunen, J. et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7**, 11122 (2016).
17. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
18. Horikoshi, M. et al. Discovery and fine-mapping of glycaemic and obesity-related trait loci using high-density imputation. *PLoS Genet.* **11**, e1005230 (2015).
19. Teumer, A. et al. Genome-wide association studies identify genetic loci associated with albuminuria in diabetes. *Diabetes* **65**, 803–817 (2016).
20. Gorski, M. et al. 1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function. *Sci. Rep.* **7**, 45040 (2017).
21. Jiang, X. et al. Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. *Nat. Commun.* **9**, 260 (2018).
22. Köttgen, A. et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.* **45**, 145–154 (2012).
23. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
24. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
25. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).
26. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
27. Tanigawa, Y. et al. Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.* **10**, 4064 (2019).
28. Peloso, G. M. et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and Blacks. *Am. J. Hum. Genet.* **94**, 223–232 (2014).
29. van der Harst, P. & Verweij, N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* **122**, 433–443 (2018).
30. Myocardial Infarction Genetics and CARDIoGRAM Exome Consortia Investigators et al. Coding variation in ANGPTL4, LPL, and SVEP1 and the risk of coronary disease. *N. Engl. J. Med.* **374**, 1134–1144 (2016).
31. DeBoever, C. et al. Assessing digital phenotyping to enhance genetic studies of human diseases. *Am. J. Hum. Genet.* **106**, 611–622 (2020).
32. McComb, R. B., Bowers, G. N. & Posen, S. *Alkaline Phosphatase* (Springer, 1979).
33. Aguirre, M., Rivas, M. A. & Priest, J. Phenome-wide burden of copy-number variation in the UK Biobank. *Am. J. Hum. Genet.* **105**, 373–383 (2019).
34. Horikawa, Y. et al. Mutation in hepatocyte nuclear factor-1β gene (TCF2) associated with MODY. *Nat. Genet.* **17**, 384–385 (1997).
35. Iwasaki, N. et al. Liver and kidney function in Japanese patients with maturity-onset diabetes of the young. *Diabetes Care* **21**, 2144–2148 (1998).
36. Nishigori, H. et al. Frameshift mutation, A263fsinsGG, in the hepatocyte nuclear factor-1β gene associated with diabetes and renal dysfunction. *Diabetes* **47**, 1354–1355 (1998).
37. Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).
38. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
39. Hayes, M. G. et al. Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. *Diabetes* **56**, 3033–3044 (2007).
40. Abhary, S. et al. Common sequence variation in the *VEGFA* gene predicts risk of diabetic retinopathy. *Invest. Ophthalmol. Vis. Sci.* **50**, 5552–5558 (2009).
41. Barton, J. C. & Acton, R. T. Diabetes in *HFE* hemochromatosis. *J. Diabetes Res.* **2017**, 9826930 (2017).
42. Raju, K. & Venkataramappa, S. M. Primary hemochromatosis presenting as type 2 diabetes mellitus: a case report with review of literature. *Int. J. Appl. Basic Med. Res.* **8**, 57–60 (2018).
43. Marks, J. D. et al. Pressure indices of myocardial oxygen consumption during pulsatile ventricular assistance. *ASAIO Trans.* **35**, 436–439 (1989).
44. Alkhatatbeh, M. J., Enjeti, A. K., Acharya, S., Thorne, R. F. & Lincz, L. F. The origin of circulating CD36 in type 2 diabetes. *Nutr. Diabetes* **3**, e59 (2013).
45. Gonzalez, Y. et al. High glucose concentrations induce TNF-α production through the down-regulation of CD33 in primary human monocytes. *BMC Immunol.* **13**, 19 (2012).
46. Larson, N. B. et al. Trans-ethnic meta-analysis identifies common and rare variants associated with hepatocyte growth factor levels in the Multi-Ethnic Study of Atherosclerosis (MESA). *Ann. Hum. Genet.* **79**, 264–274 (2015).
47. Woodward, O. M. et al. Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. *Proc. Natl Acad. Sci. USA* **106**, 10338–10342 (2009).
48. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
49. Bowden, J. et al. Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *Int. J. Epidemiol.* **47**, 2100 (2018).
50. Rücker, G., Schwarzer, G., Carpenter, J. R., Binder, H. & Schumacher, M. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics* **12**, 122–142 (2011).
51. Bowden, J. et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).
52. Curhan, G. C., Willett, W. C., Rimm, E. B. & Stampfer, M. J. A prospective study of dietary calcium and other nutrients and the risk of symptomatic kidney stones. *N. Engl. J. Med.* **328**, 833–838 (1993).
53. Ruth, K. S. et al. Using human genetics to understand the disease impacts of testosterone in men and women. *Nat. Med.* **26**, 252–258 (2020).
54. Wang, F. et al. Apolipoprotein A-IV improves glucose homeostasis by enhancing insulin secretion. *Proc. Natl Acad. Sci. USA* **109**, 9641–9646 (2012).
55. Pietzsch, J., Julius, U., Nitzsche, S. & Hanefeld, M. In vivo evidence for increased apolipoprotein A-I catabolism in subjects with impaired glucose tolerance. *Diabetes* **47**, 1928–1934 (1998).
56. Zhang, P., Gao, J., Pu, C. & Zhang, Y. Apolipoprotein status in type 2 diabetes mellitus and its complications (Review). *Mol. Med. Rep.* **16**, 9279–9286 (2017).
57. Vozarova, B. et al. High alanine aminotransferase is associated with decreased hepatic insulin sensitivity and predicts the development of type 2 diabetes. *Diabetes* **51**, 1889–1895 (2002).
58. De Silva, N. M. G. et al. Liver function and risk of type 2 diabetes: bidirectional Mendelian randomization study. *Diabetes* **68**, 1681–1691 (2019).
59. Huang, G.-X., Ji, X.-M., Ding, Y.-C. & Huang, H.-Y. Association between serum cystatin C levels and the severity or potential risk factors of acute ischemic stroke. *Neurol. Res.* **38**, 518–523 (2016).

60. van der Laan, S. W. et al. Cystatin C and cardiovascular disease: a Mendelian randomization study. *J. Am. Coll. Cardiol.* **68**, 934–945 (2016).
61. Colijn, J. M. et al. Increased high-density lipoprotein levels associated with age-related macular degeneration: evidence from the eye-risk and European Eye Epidemiology consortia. *Ophthalmology* **126**, 393–406 (2019).
62. Kay, P. et al. Age-related changes of cystatin C expression and polarized secretion by retinal pigment epithelium: potential age-related macular degeneration links. *Invest. Ophthalmol. Vis. Sci.* **55**, 926–934 (2014).
63. Klein, R., Knudtson, M. D., Lee, K. E. & Klein, B. E. K. Serum cystatin C level, kidney disease markers, and incidence of age-related macular degeneration: the Beaver Dam Eye Study. *Arch. Ophthalmol.* **127**, 193–199 (2009).
64. Zurdel, J., Finckh, U., Menzer, G., Nitsch, R. M. & Richard, G. *CST3* genotype associated with exudative age related macular degeneration. *Br. J. Ophthalmol.* **86**, 214–219 (2002).
65. Khalili, H. et al. Endogenous levels of circulating androgens and risk of Crohn's disease and ulcerative colitis among women: a nested case–control study from the nurses' health study cohorts. *Inflamm. Bowel Dis.* **21**, 1378–1385 (2015).
66. Nasser, M. et al. Testosterone therapy in men with Crohn's disease improves the clinical course of the disease: data from long-term observational registry study. *Horm. Mol. Biol. Clin. Investig.* **22**, 111–117 (2015).
67. Kühn, T. et al. Albumin, bilirubin, uric acid and cancer risk: results from a prospective population-based study. *Br. J. Cancer* **117**, 1572–1579 (2017).
68. Wray, N. R., Kemper, K. E., Hayes, B. J., Goddard, M. E. & Visscher, P. M. Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans. *Genetics* **211**, 1131–1141 (2019).
69. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
70. Qian, J. et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.* **16**, e1009141 (2020).
71. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
72. Bild, D. E. et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am. J. Epidemiol.* **156**, 871–881 (2002).
73. Inouye, M. et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
74. Cohen, J. C., Boerwinkle, E., Mosley, T. H. Jr & Hobbs, H. H. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
75. DeBoever, C. et al. Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).
76. Amar, D., Ashley, E. & Rivas, M. A. Constraint-based analysis for causal discovery in population-based biobanks. Preprint at *bioRxiv* https://doi.org/10.1101/566133 (2019).
77. Tanigawa, Y., Sinnott-Armstrong, N., Benner, C. & Rivas, M. A. *Datasets described in 'Genetics of 35 blood and urine biomarkers in the UK Biobank'* (2020); https://doi.org/10.35092/yhjc.c.5043872.v1

FinnGen

Nina Mars[2], Tuomo Kiiskinen[2,9], Aki S. Havulinna[2,9], Samuli Ripatti[2,11,13] and Mark J. Daly[2,11,15]

A full list of members and their affiliations appears in the Supplementary Information.

Methods

Genotypic and phenotypic data in the UK Biobank. We used genotype datasets from the UK Biobank (release version 2 for the directly genotyped variants and the imputed HLA allelotype datasets and release version 3 for the imputed genotype dataset), the CNV dataset[33] and the hg19 human genome reference for all analyses in the study[10]. To minimize the variability due to population structure in our dataset, we restricted our analyses to unrelated individuals based on the following four criteria reported by the UK Biobank in the sample quality control file 'ukb_sqc_v2.txt': (1) used to compute principal components ('used_in_pca_calculation' column), (2) not marked as outliers for heterozygosity and missing rates ('het_missing_outliers' column), (3) do not show putative sex chromosome aneuploidy ('putative_sex_chromosome_aneuploidy' column) and (4) have at most ten putative third-degree relatives ('excess_relatives' column).

Additionally, we used the 'in_white_British_ancestry_subset' column in the sample quality control file as a part of the population definition as shown below.

We used a combination of self-reported ancestry (UK Biobank field ID 21000) and principal component analysis and analyzed five subpopulations in the study: self-identified white British ($n=337{,}151$ individuals), African (6,498), East Asian (1,772), South Asian (7,962) and self-identified non-British white (24,909). We first used the genotype principal components of the genotyped variants from the UK Biobank and defined thresholds on principal component 1 and principal component 2 and further refined the population definition (described in the Supplementary Note). We subsequently focused on a subset of individuals with non-missing values for covariates and biomarkers as described below.

Variant annotation and quality control. Detailed information on variant annotation and quality control is described in the Supplementary Note.

Biomarker phenotype definition. Phenotypic and covariate quality control excluded rheumatoid factor and estradiol from further analyses, and fasting glucose (available for 17,439 self-reported fasting individuals) was used as a phenotype-level quality control for the glucose measurements; throughout the text, 'glucose' refers to glucose levels adjusted for fasting time rather than the GWAS among only fasting individuals (self-reporting more than 7 h and less than 24 h of fasting, $n=17{,}439$), unless otherwise noted. We focused on 32 biomarkers for genetic analysis and also defined three derived phenotypes, eGFR, non-albumin protein and AST to ALT ratio, for a total of 35 biomarkers (Supplementary Table 4a). The eGFR measurement is an indicator of renal function and is defined by the CKD-EPI equation[78]. We defined non-albumin protein levels as the difference between the levels of total protein and albumin. Then, after applying covariate correction (see Covariate correction below; Supplementary Table 4b), we additionally defined the AST to ALT ratio as the difference of the (log-transformed) estimates for AST and ALT levels.

Statin identification and LDL adjustment. Statin identification and LDL adjustment are described in the Supplementary Note.

Covariate correction. Covariate adjustment is described in the Supplementary Note.

Definition of type 2 diabetes. We used the definition of type 2 diabetes from our previous paper, including the removal of individuals with type 1 diabetes from both affected and control groups[79]. We use the terminology from Eastwood et al. throughout this description[80]. Type 2 diabetes was assigned case status for 'probable type 2 diabetes' and 'possible type 2 diabetes' and control status for 'type 2 diabetes unlikely'; in addition, individuals with 'probable type 1 diabetes', 'possible type 1 diabetes' or 'probable gestational diabetes' were excluded. Finally, for the 'strict' type 2 diabetes definition, we removed controls with HbA1c $\geq$ 39 mmol per mol.

Genome-wide association analyses. We performed genome-wide association analyses using the following four datasets: (1) the directly genotyped variants on the array (for PTVs and PAVs), (2) the imputed variants (version 3), (3) the imputed HLA alleles and (4) the CNVs and gene-level aggregated CNV burden[33]. All the *P* values from the association analyses are from two-sided tests. A detailed description of the association analysis is provided in the Supplementary Note.

Meta-analysis. Using the GWAS summary statistics for four analyzed populations (white British, non-British white, South Asian and African; East Asian GWASs were excluded) in the UK Biobank, we performed inverse-variance weighted (IVW) meta-analysis using METAL (version 2011-03-25) and included a heterogeneity of effects analysis.

For the summary statistics from the meta-analysis, we determined whether the A1 and A2 alleles matched with the alternate and reference alleles in the GRCh37/hg19 reference genome (fasta file) using the BEDTools getfasta subcommand[81] and canonicalized our association summary statistics so that the effect size was always reported with respect to the alternate allele in the reference genome.

Derivation of independent loci. Once we performed the GWAS, full summary statistics were clumped to $r^2 > 0.1$ using the following clump command.

plink1.9 --bfile <1000G Phase 3 European plink file> --clump <summary statistics> --clump-p1 1e-6 --clump-p2 1e-4 --clump-r2 0.1 --clump-kb 10000 --clump-field P --clump-snp-field ID

Then, to avoid calling very large signals as multiple associations, these were further filtered so that any SNPs within 0.1 cM of each other (as annotated by 1000 Genomes) were considered part of the same association signal, with the cM annotation derived from 1000 Genomes phase 3 European samples ($n=489$)[24]; variants within 0.1 cM were chosen to only have the minimum *P* value.

To report independent signals, we ran the following PLINK command and counted the number of independent SNPs that it reported.

plink1.9 --bfile <1000G Phase 3 European plink file> --extract <all unique hit SNPs, n = 6269> --indep 50 5 2

Comparison of effect sizes with published studies. Full summary statistics from comparison studies (PMID in Supplementary Table 5) were downloaded and overlapped with our GWAS summary statistics using the munging framework from LD Score regression to align alleles (modified to additionally report the unnormalized β). The observed correlation coefficients and linear effect regression coefficients across variants with $P < 1\times10^{-6}$ in either study (subthreshold) or $P < 5\times10^{-8}$ in our study (GWAS hits) are listed in Supplementary Table 5. Using the same set of comparison studies, we also determined whether the PTV and PAV associations were previously reported for a given trait by calling the association reported if the *P* value of the variant was less than 1×10^{-6} in any comparison study for a given trait.

Fine-mapping biomarker-associated regions. Independent loci were defined by clumping white British GWAS summary statistics (Derivation of independent loci). For each putative SNP, distance-independent regions were defined by collating all variants in LD with the following PLINK command.

plink1.9 --clump-p1 1e-3 --clump-p2 1e-3 --clump-r2 0.0001 --clump-kb 10000 --clump-field P-value --clump-snp-field MarkerName

In this manner, we defined the individual loci contributing to the fine-mapping. We identified putative causal SNPs in each locus by using FINEMAP software versions 1.3 and 1.4 (ref. [38]). The output from FINEMAP is (1) a list of potential causal configurations together with their posterior probabilities and Bayes factors, (2) for each SNP, the posterior probability and Bayes factor of being causal and (3) credible sets for each identified causal signal. We applied FINEMAP with its default settings while allowing for a maximum of 40 causal SNPs and by using pairwise correlations between SNPs computed from the original GWAS genotype data as previously recommended[37].

We performed fine-mapping on all associations with more than one variant for which the most significantly associated variant had a *P* value less than 1×10^{-3}. We filtered regions based on the unique variant ID (in the MFI file from the UK Biobank) for those regions for which at least one of the variants in the region was annotated as an association lead SNP in our analysis ($P < 5\times10^{-9}$).

Heritability estimates. Heritability analysis is described in the Supplementary Note.

Targeted phenome-wide association analysis. We curated a list of 166 medically relevant phenotypes from previously reported binary phenotypes in Global Biobank Engine[27,31,75]. Specifically, we selected phenotypes with at least 700 cases in the white British population and removed phenotypes that were likely to be duplicated (Supplementary Table 12). Those phenotypes include non-cancer disease-outcome endpoints derived from a combination of the ICD codes from hospital inpatient records as well as self-reported disease ascertainment status[31], family history phenotype (UK Biobank data category 100034), cancer phenotypes derived from a combination of the UK cancer registry data and questionnaire data[75] and an additional set of medically relevant phenotypes derived from the following data fields in UK Biobank: 2247, 2463, 2834, 3591, 6148, 6149, 6152, 6153, 20126, 20406, 20483 and 21068. For example, the chronic kidney disease phenotype was defined based on the combination of self-reported kidney disease (coded as '1192' in UK Biobank Data coding ID 6) and ICD-10 code (N17 ('acute kidney failure'), N18 ('chronic kidney disease (CKD)'), N19 ('unspecified kidney failure') and its subconcepts) from hospital inpatient data (Supplementary Fig. 8a), which was visualized with the R UpSetR package version 1.4.0. The data source for the phenotype definitions is described in the 'Source of the phenotype' column in Supplementary Table 12.

After performing LD pruning using PLINK with '--indep 50 5 2' as previously described[27,75], we prioritized (1) the 632 LD-independent PAVs or PTVs outside of the MHC region that showed significant associations ($P < 5\times10^{-9}$) on the genotyping array, as well as (2) 43 non-synonymous variants and (3) 2,442 synonymous or non-coding variants with significant associations ($P < 5\times10^{-9}$) from the imputation dataset (Supplementary Table 13a–c). We applied the PheWAS analysis for those variants with a *P* value threshold of $P < 1\times10^{-7}$.

For the resulting associations, we searched the NHGRI-EBI GWAS Catalog to determine whether they were already reported in previous studies[82]. Specifically, we identified the LD proxy ($r^2 > 0.9$) of the PheWAS target variants and manually inspected the reported associations for those variants. For associations with no supporting prior studies, we additionally queried Open Target Genetics and

manually assessed whether the associations were new[83]. In addition, we also searched the FinnGen study (Freeze R2, http://r2.finngen.fi/) and asked whether the PheWAS target variants and their LD proxies had similar associations. These PheWAS results and the reference to the prior association reports are summarized in Supplementary Table 13a–c.

For the CNV PheWAS, we queried summary statistics from previous CNV association tests for the 173 traits of interest[33]. Results for a burden of *HNF1B* CNVs are shown in Supplementary Fig. 6a, along with the corresponding meta-analyzed summary statistics for biomarker traits described in this work.

Correlation of genetic effects across relevant phenotypes. We used LD Score regression in genetic correlation mode to estimate genetic correlation effects between biomarkers and the 166 medically relevant phenotypes used in the PheWAS analysis. The exact arguments were as follows.

ldsc.py --rg <traits> --ref-ld-chr ldsc/1000G.EUR.QC/ --w-ld-chr ldsc/weights_hm3_no_hla/weights.

Causal inference. For the final results, all lead variants with $P < 5 \times 10^{-8}$ were kept for the MR analyses. All MR calculations were performed using TwoSampleMR, which was also used to perform trait munging[84].

We used the Rücker model-selection framework for causal inference as follows[49–51,85]. For each exposure–outcome pair, we started with a simple fixed-effects IVW MR analysis and computed the model's significance and the Q statistic for heterogeneity. If the significance of Q was <0.01, then we used it as evidence for heterogeneity and switched to a mixed-effects IVW model instead. We then computed an MR-Egger model and compared it to the IVW-selected model. Let Q_{ivw} be the Q statistic of the IVW model and Q_e be the Q statistic of the MR-Egger model. We computed the significance of the difference $Q_e - Q_{ivw}$ using a χ^2_1 distribution and switched to the MR-Egger model if the result was significant ($P < 0.01$). The significance of all selected models was adjusted using a BY FDR correction at 5%[86]. Network visualization of the results was performed using Cytoscape versions 3.7 and 3.8 (ref. [87]).

Polygenic prediction within and across populations. To construct PRSs for each of the traits, we applied the BASIL algorithm implemented in the R package snpnet. This method is capable of finding the exact solution for L_1-penalized multivariate regression (lasso) on an ultrahigh-dimensional large dataset through an iterative procedure built on top of the R package glmnet[69,70,88]. Because this method considers all of the genetic variants available in the input dataset and performs variable selection and multivariate regression fit simultaneously, it is suitable for polygenic risk prediction from a large-scale dataset.

We randomly split the white British individuals into training (70%, $n = 223{,}327$ with non-missing phenotypes for at least one biomarker trait), validation (10%, $n = 31{,}929$) and test (20%, $n = 63{,}818$) sets and used both training and validation sets to fit multivariate lasso regression models. The validation set was used to find the optimal penalization (sparsity) parameter with respect to the predictive performance (R^2). To maximize the performance of polygenic prediction, we combined the directly genotyped variants, the imputed HLA allelotypes and the CNV dataset with PLINK version 1.9 and used this as the input genotype dataset consists of 1,080,968 variants. For each biomarker phenotype, we applied the R package snpnet for the log-transformed and covariate-adjusted phenotypes and regression coefficients, BETAs[70].

Using the β values from multivariate lasso regression, we computed the PRS for each individual with the PLINK2 subcommand --score[89]. To evaluate the performance of the models, we computed R^2 values for log-transformed phenotypes for individuals in the held-out white British test set ($n = 63{,}818$), as well as for self-identified non-British white ($n = 23{,}595$), African ($n = 6{,}021$), South Asian ($n = 7{,}341$) and East Asian ($n = 1{,}082$) populations. To assess the incremental predictive performance compared to the covariates, we evaluated the R^2 values for the risk score computed from the covariate (defined as the difference between the log-transformed phenotype value and log-transformed and covariate-adjusted phenotype values), as well as the combined risk score (the sum of the covariate score and genotype PRS, Supplementary Table 17a). Polygenic score accuracy was generally independent of residualization strategy (Supplementary Table 17b).

For the evaluation of multi-PRS models, we also trained snpnet PRS models for disease outcomes using the R package snpnet in the same fashion as for the biomarker phenotypes, except that we used the binomial family for logistic regression and AUC as the criterion to select the sparsity parameter.

Evaluation of snpnet PRS models with the MESA cohort is described in the Supplementary Note.

Single-trait biomarker PRS-PheWAS. We started by enumerating all our 166 high-confidence traits that were replicated between ICD codes and self-reported, cancer, family history and manually curated traits[31,75], as described in the PheWAS analysis above. For each of the 35 biomarkers, we used R's fisher.test implementation of Fisher's exact test between the 40th and 60th percentiles and the top and bottom 1% and 1–10% of PRS in the union of the unrelated non-British white individuals and the held-out test set of unrelated white British individuals. We then corrected for multiple hypotheses using a Bonferroni-adjusted Q value less than 5% within each biomarker and reported the enrichment as the OR estimate from Fisher's exact test.

Models for multi-PRS prediction of disease outcomes. To perform out-of-sample validation, we trained L_1-regularized logistic regression models with glmnet using only the 35 biomarker PRSs and the snpnet PRS for the trait of interest as predictors[69]. Results were evaluated using the area under the ROC curve in the union of the held-out test set of self-identified white British individuals and all unrelated, self-identified non-British white individuals for which the corresponding phenotypes were available (as used in the cross-population testing; see above). We also performed the lasso regression additionally including age, sex, genotyping array and the top ten global principal components of the genotyping matrix as covariates for each outcome (referred to as 'age/sex/PCs') and additional information provided in the Supplementary Note.

Finally, we derived versions of the multi-PRS model with these covariates and also relevant pre-existing polygenic scores for gallstones[90], type 2 diabetes[91,92] and heart attack[93,94] in the model (Supplementary Table 20). We refer to models trained only on covariates and trait polygenic scores as 'baseline' models and those which additionally include the 35 biomarker PRSs as 'multi-PRS' models throughout the manuscript.

Evaluation of multi-PRS prediction in an external cohort. The FinnGen Data Freeze 3 comprised 135,300 Finnish participants, with phenotypes derived from ICD (eight, ninth and tenth revisions) diagnosis codes obtained from national registries, including the national Finnish hospital discharge and cause-of-death registries as a part of the FinnGen project (Supplementary Table 24).

FinnGen samples were genotyped with Illumina and Affymetrix arrays (Thermo Fisher Scientific). Genotype imputation was carried out by using the population-specific SISu version 3 imputation reference panel with Beagle 4.1 (version 08Jun17.d8b, https://faculty.washington.edu/browning/beagle/b4_1.html) as described in the following protocol: https://doi.org/10.17504/protocols.io.nmndc5e. Post-imputation quality control involved excluding variants with INFO score <0.6.

We estimated a full weighting matrix for each SNP from the corresponding coefficients of the regression model and then applied the per SNP-weighted model to individuals in the FinnGen study. To assess the risk for first disease events, HRs and 95% CIs per s.d. increment were estimated with Cox proportional hazards models after evaluation of the proportionality assumption. For the comparison, on type 2 diabetes and myocardial infarction, with models including the existing polygenic scores, scores were standardized within the population before being applied using the weights of standardized PRSs (Supplementary Tables 22 and 23) to capture the differences in SNP sets used across the scores. An R script to perform these analyses, which takes the raw PRS for each outcome of interest as input, is available in the code repository. With age as the time scale, the survival models were stratified by sex and adjusted for batch, and the first ten principal components of ancestry were calculated within Finns.

Statistics. For computational and statistical analyses, we used Jupyter Notebook with Python (3.6 and 2.7) and R kernels (http://jupyter.org/), R (versions 3.5.2 and 3.4.0), R studio (3.5.2), R tidyverse package version 1.3.0 and Stata version 15. Software and packages used for specific analyses are listed in the corresponding subsections above. The P values were computed from two-sided tests, unless otherwise specified.

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

Data availability

Summary-level data generated in this work are available at the NIH's instance of figshare (the meta-analyzed GWAS summary statistics (https://doi.org/10.35092/yhjc.12355382), the fine-mapped associations (https://doi.org/10.35092/yhjc.12344351), the snpnet PRS coefficients (https://doi.org/10.35092/yhjc.12298838) and the multi-PRS weights (https://doi.org/10.35092/yhjc.12355424), please see the Supplementary Note for details)[77]. Other data are displayed in the Global Biobank Engine (https://biobankengine.stanford.edu).

Code availability

Analysis scripts and notebooks are available on GitHub at https://github.com/rivas-lab/biomarkers/.

References

78. Levey, A. S. et al. A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150**, 604–612 (2009).
79. Wainberg, M. et al. Homogeneity in the association of body mass index with type 2 diabetes across the UK Biobank: a Mendelian randomization study. *PLoS Med.* **16**, e1002982 (2019).
80. Eastwood, S. V. et al. Algorithms for the capture and adjudication of prevalent and incident diabetes in UK Biobank. *PLoS ONE* **11**, e0162388 (2016).

81. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
82. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
83. Carvalho-Silva, D. et al. Open targets platform: new developments and updates two years on. *Nucleic Acids Res.* **47**, D1056–D1065 (2019).
84. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
85. Slob, E. A. W., Groenen, P. J. F., Thurik, A. R. & Rietveld, C. A. A note on the use of Egger regression in Mendelian randomization studies. *Int. J. Epidemiol.* **46**, 2094–2097 (2017).
86. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
87. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
88. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
89. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
90. Joshi, A. D. et al. Four susceptibility loci for gallstone disease identified in a meta-analysis of genome-wide association studies. *Gastroenterology* **151**, 351–363 (2016).
91. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
92. Läll, K., Mägi, R., Morris, A., Metspalu, A. & Fischer, K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet. Med.* **19**, 322–329 (2017).
93. Abraham, G. et al. Genomic prediction of coronary heart disease. *Eur. Heart J.* **37**, 3267–3278 (2016).
94. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

Acknowledgements

This research was conducted using the UK Biobank Resource under application number 24983, 'Generating effective therapeutic hypotheses from genomic and hospital linkage data' (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf). Based on the information provided in protocol 44532, the Stanford IRB has determined that the research does not involve human subjects as defined in 45 CFR 46.102(f) or 21 CFR 50.3(g). All participants in the UK Biobank study provided written informed consent (more information is available at https://www.ukbiobank.ac.uk/2018/02/gdpr/). The FinnGen project is approved by the Finnish Institute for Health and Welfare (THL), approval number THL/2031/6.02.00/2017, amendments THL/1101/5.05.00/2017, THL/341/6.02.00/2018,THL/2222/6.02.00/2018, THL/283/6.02.00/2019), Digital and population data service agency VRK43431/2017-3, VRK/6909/2018-3, the Social Insurance Institution (KELA) KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019 and Statistics Finland TK-53-1041-17. The Biobank Access Decisions for FinnGen samples and data include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7 Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, Auria Biobank AB17-5154, Biobank Borealisof Northern Finland_2017_1013, Biobank of Eastern Finland 1186/2018, Finnish Clinical Biobank Tampere MH0004, Central Finland Biobank 1-2017, and Terveystalo Biobank STB 2018001. The following biobanks are acknowledged for collecting the FinnGen project samples: Auria Biobank (https://www.auria.fi/biopankki), THL Biobank (https://thl.fi/fi/web/thl-biopankki), Helsinki Biobank (https://www.terveyskyla.fi/helsinginbiopankki), Biobank Borealis of Northern Finland (https://www.oulu.fi/university/node/38474), Finnish Clinical Biobank Tampere (https://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (https://ita-suomenbiopankki.fi), Central Finland Biobank (https://www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (https://www.veripalvelu.fi/verenluovutus/biopankkitoiminta) and Terveystalo Biobank (https://www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). All Finnish Biobanks are members of the BBMRI.fi infrastructure (www.bbmri.fi). Statin adjustment analyses were further conducted via UK Biobank application 7089 using a protocol approved by the Partners HealthCare Institutional Review Board. We thank all the participants in the UK Biobank and FinnGen studies. We thank A. Paterson and members of the Rivas, Pritchard and Bejerano labs for their feedback. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (NIH) and by the NCI, the National Human Genome Research Institute (NHGRI), NHLBI, NIDA, NIMH and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 19 October 2020. This work was supported by the NHGRI of the NIH under awards R01HG010140 (M.A.R.), R01EB001988-21 (T.H.) and R01HG008140 (J.K.P.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Some of the computing for this project was performed on the Sherlock cluster at Stanford University. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. N.S.-A. is supported by the Department of Defense through a National Defense Science and Engineering grant and by a Stanford Graduate Fellowship. Y.T. is supported by a Funai Overseas Scholarship from the Funai Foundation for Information Technology and the Stanford University School of Medicine. N.M. is supported by the Academy of Finland (no. 331671). H.M.O. is supported by the Academy of Finland (no. 309643). F.R. is supported by a National Heart, Lung, and Blood Institute grant (1K01HL144607). The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and by 12 industry partners (AbbVie Inc, AstraZeneca UK Ltd., Biogen MA Inc., Celgene Corporation, Celgene International II Sàrl, Genentech Inc, Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc. and Novartis AG). M.A.R. is in part supported by the NHGRI of the NIH under award R01HG010140 (M.A.R.) and an NIH Center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01 HG009080). The land upon which some of this work was performed is the ancestral and unceded land of the Muwekma Ohlone, and we pay our respects to their elders past and present.

Author contributions

M.A.R., Y.T. and N.S.-A. conceived and designed the study. N.S.-A., Y.T., D.A., N.M., C.B., M.A., G.R.V., J.P.P., J.Q., A.S. and M.A.R. carried out the statistical and computational analyses with advice from M.W., H.M.O., F.R., T.L.A., V.A., R.T., T.H., S.R., J.K.P. and M.J.D. T.K., A.S.H. and T.L.A. organized reagents. N.S.-A., Y.T., D.A., M.A., G.R.V. and M.A.R. carried out quality control of the data. M.A.R., Y.T. and N.S.-A. supervised computational and statistical aspects of the study. The manuscript was written by N.S.-A., Y.T., D.A., M.A., G.R.V., V.A. and M.A.R. and revised by all the co-authors. All co-authors approved of the final version of the manuscript.

Competing interests

The Board of Trustees of the Leland Stanford Junior University filed a US Provisional Application 'Methods for diagnosis of polygenic diseases and phenotypes from genetic variation' (serial no. 62/852,738) describing this work. J.K.P., M.A.R., N.S.-A. and Y.T. are designated as inventors of the patent. M.A.R. is on the SAB of 54gene and the computational advisory board for Goldfinch Bio and has advised BioMarin, Third Rock Ventures, MazeTx and Related Sciences. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-00757-z.

Correspondence and requests for materials should be addressed to N.S.-A., Y.T. or M.A.R.

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

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Nasa Sinnott-Armstrong, Yosuke Tanigawa, Manuel A. Rivas

Last updated by author(s): Nov 16, 2020

Reporting Summary

Nature Research 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 Research 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	We used array-genotype data and phenotype data from the UK Biobank [Ref:10] and FinnGen (http://finngen.fi/en). The primary data used to generate the analyses presented here are available in the UK Biobank access management system (https://amsportal.ukbiobank.ac.uk/) for application 24983,"Generating effective therapeutic hypotheses from genomic and hospital linkage data" (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf). No software was used for data collection.

Data analysis	See Method section for the full details of the data analysis. We provide all the analysis scripts on GitHub (https://github.com/rivas-lab/biomarkers. All resources not described at these links are available upon request to the corresponding authors. We used the following software for the data analysis. - PLINK v1.90b and v2.00a (https://www.cog-genomics.org/plink/2.0/) - Jupyter notebook with Python (3.6 and 2.7) and R kernels (http://jupyter.org/) - R (version 3.5.2 and 3.4.0) and R studio (3.5.2) - Stata version 15 - Ensembl Variant Effect Predictor (VEP, April 2017 version) with Loftee plugin (https://github.com/konradjk/loftee, version v0.3-beta) - R snpnet package version 0.3 (https://github.com/rivas-lab/snpnet/releases) - R bma package version 3.18.12 - R UpSetR package version 1.4.0 - R tidyverse package version 1.3.0 - METAL meta-analysis wrapper version 2011-03-25 - FINEMAP version 1.3 and 1.4 - LD score regression version 1.0.1 - GCTA 1.26.0 - Cytoscape version 3.7 and 3.8

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Supplementary Data 1-4 are available at NIH's instance of figshare (https://doi.org/10.35092/yhjc.12355382, https://doi.org/10.35092/yhjc.12344351, https://doi.org/10.35092/yhjc.12298838, and https://doi.org/10.35092/yhjc.12355424, please see Supplementary Note for details). Other data is displayed in the Global Biobank Engine (https://biobankengine.stanford.edu).

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	We used the UK Biobank population cohort and analyzed in total of 363,228 individuals in 5 subpopulations in the study: self-identified White British (n = 337,151 individuals), African (6,498), East Asian (1,772), South Asian (7,962), and self-identified non-British White (24,909). We also used FinnGen (n=135,500) cohort. We took the maximum number of individuals who does not meet the data exclusion criteria (see the details below).
Data exclusions	To minimize the variabilities due to population structure in our dataset, we restricted our association analyses to include 363,228 individuals based on the following five criteria reported by the UK Biobank in the file "ukb_sqc_v2.txt": 1. used to compute principal components ("used_in_pca_calculation" column) 2. not marked as outliers for heterozygosity and missing rates ("het_missing_outliers" column) 3. do not show putative sex chromosome aneuploidy ("putative_sex_chromo- some_aneuploidy" column) 4. have at most 10 putative third-degree relatives ("excess_relatives" column). Other individuals who does not meet the criteria above was excluded from the analysis. Please refer to "Genotype data preparation" subsection in Methods section for more details.
Replication	Our findings are replicated with self-identified non-British White, MESA, and FinnGen cohorts when applicable (see main text). Notably, multi-PRS analysis is successfully replicated in FinnGen.
Randomization	We randomly split the training, validation, and test set split was performed without using the phenotype information. Please see "Polygenic prediction within and across populations" subsection in the Method section.
Blinding	We randomly split the training, validation, and test set split was performed without using the phenotype information. Blinding was not applicable to other aspects of our study because we allocated individual groupings only in the polygenic prediction analysis.

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 used the UK Biobank population cohort and analyzed 5 subpopulations in the study: (n = 337,151 individuals, see below), African (6,498), East Asian (1,772), South Asian (7,962), and self-identified non-British White (24,909). We also used FinnGen 3 (n=135,500) cohort.
Recruitment	This research has been conducted using the UK Biobank Resource under Application Number 24983, "Generating effective therapeutic hypotheses from genomic and hospital linkage data" (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf) and FinnGen (http://finngen.fi/en).
Ethics oversight	Based on the information provided in Protocol 44532 the Stanford IRB has determined that the research does not involve human subjects as defined in 45 CFR 46.102(f) or 21 CFR 50.3(g). All participants of UK Biobank provided written informed consent (more information is available at https://www.ukbiobank.ac.uk/2018/02/gdpr/). For FinnGen, all patients and control subjects provided informed consent, a biobank research consent, based on the Finnish Biobank Act. Recruitment protocols followed the biobank protocols approved by Valvira, the National Supervisory Authority for Welfare and Health. The Ethical Review Board of the Hospital District of Helsinki and Uusimaa approved the FinnGen study protocol Nr HUS/990/2017. All DNA samples and data in this study were pseudonymized.

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