Effects of protein-coding variants on blood metabolite measurements and clinical biomarkers in the UK Biobank

Authors

Abhishek Nag, Ryan S. Dhindsa, Lawrence Middleton, ..., Dirk S. Paul, Andrew R. Harper, Slavé Petrovski

Correspondence

slav.petrovski@astrazeneca.com

Blood metabolite measurements can help diagnose and monitor human disease. We analyzed 412,393 UK Biobank exomes to assess rare variant contribution to 355 metabolic blood measurements, revealing significant associations for 205 genes. We provide several examples of how this resource can help elucidate disease mechanisms.

Nag et al., 2023, The American Journal of Human Genetics *110*, 487–498
March 2, 2023
https://doi.org/10.1016/j.ajhg.2023.02.002

CellPress

Effects of protein-coding variants on blood metabolite measurements and clinical biomarkers in the UK Biobank

Abhishek Nag,[1] Ryan S. Dhindsa,[2,3,4] Lawrence Middleton,[1] Xiao Jiang,[1] Dimitrios Vitsios,[1] Eleanor Wigmore,[1] Erik L. Allman,[1] Anna Reznichenko,[5] Keren Carss,[1] Katherine R. Smith,[1] Quanli Wang,[2] Benjamin Challis,[5] Dirk S. Paul,[1] Andrew R. Harper,[1,6] and Slavé Petrovski[1,7,*]

Summary

Genome-wide association studies (GWASs) have established the contribution of common and low-frequency variants to metabolic blood measurements in the UK Biobank (UKB). To complement existing GWAS findings, we assessed the contribution of rare protein-coding variants in relation to 355 metabolic blood measurements—including 325 predominantly lipid-related nuclear magnetic resonance (NMR)-derived blood metabolite measurements (Nightingale Health Plc) and 30 clinical blood biomarkers—using 412,393 exome sequences from four genetically diverse ancestries in the UKB. Gene-level collapsing analyses were conducted to evaluate a diverse range of rare-variant architectures for the metabolic blood measurements. Altogether, we identified significant associations ($p < 1 \times 10^{-8}$) for 205 distinct genes that involved 1,968 significant relationships for the Nightingale blood metabolite measurements and 331 for the clinical blood biomarkers. These include associations for rare non-synonymous variants in *PLIN1* and *CREB3L3* with lipid metabolite measurements and *SYT7* with creatinine, among others, which may not only provide insights into novel biology but also deepen our understanding of established disease mechanisms. Of the study-wide significant clinical biomarker associations, 40% were not previously detected on analyzing coding variants in a GWAS in the same cohort, reinforcing the importance of studying rare variation to fully understand the genetic architecture of metabolic blood measurements.

Introduction

Metabolic blood measurements represent intermediates or end products of biochemical pathways that can be used to diagnose and monitor human disease. Applying metabolic blood measurements as intermediate traits has helped researchers dissect the genetic basis of several complex human diseases.[3] Furthermore, genetic underpinnings of metabolic blood measurements can offer insights into human disease mechanisms and, in turn, provide potential therapeutic opportunities. Large-scale genome-wide association studies (GWASs) have identified many statistically significant genetic loci that regulate levels of metabolic blood measurements.[4–14] However, difficulty in interpreting the functional effects of non-coding variants that often underlie these loci have stymied the clinical impact of many of these associations.[15] In addition, certain important rare variant studies have also expanded our understanding of genetic determinants of the human metabolome, but the sample sizes for these studies were limited.[16–20]

The UK Biobank (UKB)[21] is a large population-based resource of ~500,000 participants with genetic data linked to a diverse set of phenotypic measurements. To date, analyses of human metabolic measures in this cohort have focused largely on common and low-frequency variation derived from genotype data.[1,2] These studies identified associations for a few thousand independent loci and have greatly enhanced our understanding of the genetic architecture of human metabolic measurements.

Associations for rare protein-coding variants have demonstrably greater translational potential given their often-larger effect sizes[22] and our ability to interpret their functional impact.[23] The availability of exome sequences in the UKB now allows for a more systematic exploration of protein-coding variants across the allele frequency spectrum (including ultra-rare to rare variants) to identify associations that may either underpin known GWAS loci or represent previously unreported loci regulating metabolic blood measurements. Here we present variant- and gene-level (collapsing) association analyses for 325 Nightingale assay-derived blood metabolite measurements (N = 99,283) and 30 clinical blood biomarkers (N = 393,351) across multiple genetic ancestries in the UKB.

[1]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK; [2]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Waltham, MA, USA; [3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [4]Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, TX 77030, USA; [5]Translational Science and Experimental Medicine, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden; [6]Early Respiratory and Immunology, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK; [7]Department of Medicine, University of Melbourne, Austin Health, Melbourne, VIC, Australia
*Correspondence: slav.petrovski@astrazeneca.com
https://doi.org/10.1016/j.ajhg.2023.02.002.

Material and methods

UK Biobank (UKB) resource

The UKB resource[21] is a prospective cohort study of ~500,000 individuals from across the United Kingdom, aged between 40 and 69 years. Whole-exome sequencing was performed in 454,988 participants. For the sequenced cohort, the average age at recruitment was 56.5 years and 54% were female. Participant data, obtained through questionnaires and assessment visits, include health records that are periodically updated by the UKB, self-report survey information, linkage to death and cancer registries, urine and blood biomarkers, imaging data, accelerometer data, and various other phenotypic endpoints.[21] All study participants provided informed consent. For this study, data from the UKB resource were accessed under the application number 26041. The protocols for UKB are overseen by The UK Biobank Ethics Advisory Committee (EAC); for more information, see https://www.ukbiobank.ac.uk/ethics/.

Metabolic blood measurements

Concentrations of 168 blood metabolites (107 "non-derived" metabolites and 61 composite metabolites) and 81 metabolite ratios, pertaining to lipoprotein lipids, fatty acids and their compositions, and various other low-molecular-weight metabolites, were measured in a subset of randomly selected 121,695 UKB participants by Nightingale Health using nuclear magnetic resonance (NMR) spectroscopy[24] (Table S1A). Metabolite measurements were normalized for known sources of technical variation using the ukbnmr R package (https://github.com/sritchie73/ukbnmr/). The package first adjusts the subset of 107 non-derived metabolites for technical variation using robust linear regression, and then derives the remaining metabolite measurements post-adjustment. We also derived an additional 76 metabolite ratios that were not present in the original Nightingale data but were suggested by the authors of the ukbnmr package as having potential biological significance.[24] Thus, in total, 325 metabolite measurements (107 non-derived metabolites, 61 composite metabolites, 81 metabolite ratios, and 76 ukbnmr-derived metabolite ratios) were analyzed in this study. Additionally, we also analyzed routine clinical blood biomarkers (n = 30) related to glucose and lipid metabolism, renal and liver function, that were measured in the majority of the 500,000 UKB participants (Table S1B). There were eight biochemical measurements—namely, albumin, apolipoprotein B, creatinine, glucose, HDL-cholesterol, LDL-cholesterol, triglycerides, and total cholesterol—that were measured as a part of both the Nightingale metabolites and the clinical biomarkers. Across the eight biochemical measurements, there was strong positive correlation between their values in the two datasets, with Pearson's correlation coefficients (r) ranging from 0.65 (albumin) to 0.93 (HDL-cholesterol and triglycerides) (Table S2). However, given that the observed differences in these biochemical measurements between the two datasets might reflect differences in sample processing, assay method, and data normalization steps, we analyzed all measurements. Rank-based inverse-normal transformation was applied to all metabolite measurements and clinical biomarkers prior to performing association analyses.

Additionally, for four clinical biomarkers—namely, LDL-cholesterol, total cholesterol, apolipoprotein B, and urate—we adjusted for the effect of commonly prescribed medications known to influence their levels. For LDL-cholesterol, total cholesterol, and apolipoprotein B, we adjusted for the effect of statins based on their "statin adjustment factors," previously estimated in the UKB as 0.684, 0.749, and 0.719, respectively.[2] Similar to the method described for statins,[2] we adjusted urate levels for the effect of allopurinol using an "allopurinol adjustment factor (0.810)," which was calculated based on the subset of participants that had repeat assessment for urate levels and had started taking allopurinol between the enrollment visit and the second visit. All medication-adjusted values were calculated only for the European ancestry participants given that the available sample sizes for the non-European ancestries in this study were insufficient to permit medication correction factors to be accurately calculated.

Whole-exome sequencing and bioinformatics pipeline

Whole-exome sequences for 454,988 UKB participants were generated at the Regeneron Genetics Center as part of a pre-competitive data generation collaboration between AbbVie, Alnylam Pharmaceuticals, AstraZeneca, Biogen, Bristol-Myers Squibb, Pfizer, Regeneron, and Takeda.[25] The exome-sequencing procedure and the relevant quality control (QC) steps have been detailed previously in Szustakowski et al.[25] and Wang et al.[26] The FASTQ sequences that were made available were first aligned, following which single-nucleotide variants (SNVs) and small indels were called using Illumina's DRAGEN Bio-IT Platform Germline Pipeline v.3.0.7 (https://emea.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html) on the Amazon Web Services cloud compute platform available at AstraZeneca's Center for Genomics Research. SNPEff v.4.3[27] was used to annotate the "most damaging effect" predicted for each protein-coding variant. In addition, we used certain other bioinformatic tools such as missense tolerance ratio (MTR) scores[28] to identify regions of protein-coding genes under constraint for missense variants, and REVEL[29] to prioritize coding variants based on their predicted deleteriousness. Further details on how these tools were applied to the UKB exome-sequencing dataset have been previously described.[26]

Selection of UKB samples for association analyses

Prior to performing the association analyses, we excluded samples from the available UKB exome-sequencing dataset (N = 454,796) based on the following QC measures[26] (Figure S1).

DNA contamination: VerifyBAMID freemix (measure of DNA contamination) > 4%
Coverage depth: ≥10× for <94.5% of the consensus coding sequence (CCDS release 22)
Relatedness: second-degree relatives or closer (equivalent to kinship coefficient > 0.0884), as estimated using the –kinship function in KING v.2.2.2[30]

Additionally, to account for differing genetic ancestry when performing the analyses, we assigned samples to one of the four major ancestral groups (minimum 1,000 participants): European (N = 394,694), South Asian (N = 8,078), East Asian (N = 2,209), and African (N = 7,412) (Table S3). This was done by excluding participants: (1) with predicted probability of genetic ancestry <0.99 (for European ancestry) or <0.95 (for the remaining ancestries), as estimated using PEDDY v.0.4.2; or (2) with one or more of their top four principal component (PC) values lying outside four standard deviations from their corresponding genetic ancestry PC distribution.

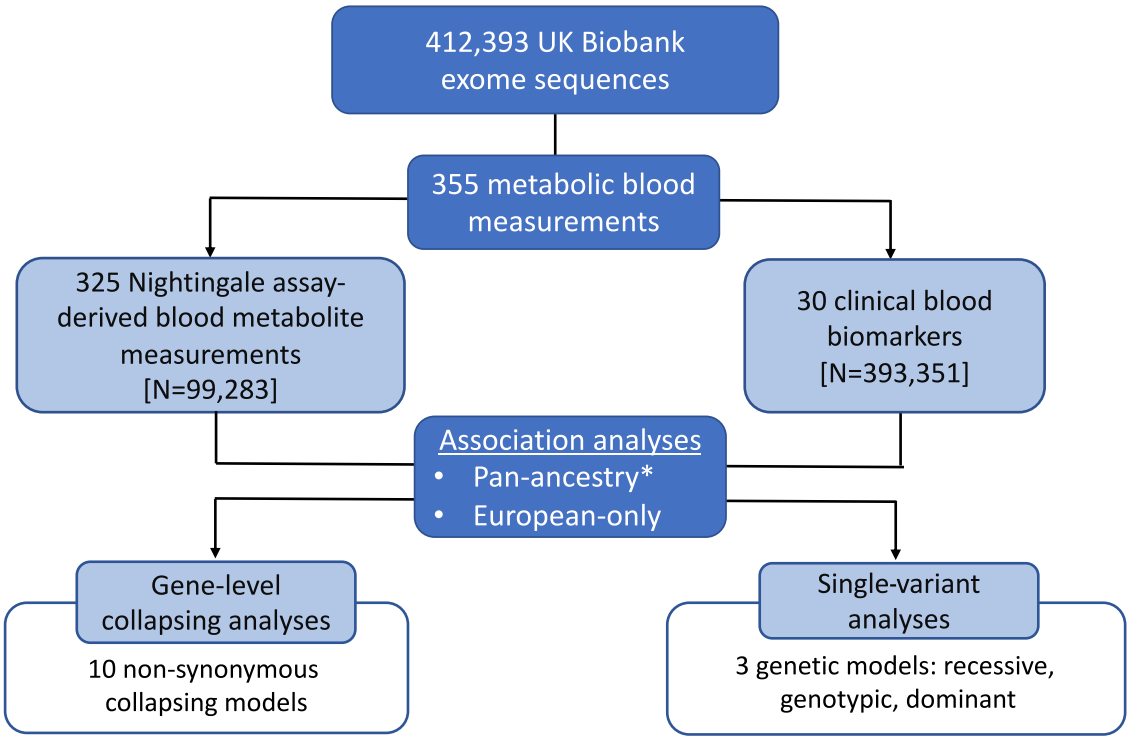

Figure 1. A schematic of the genetic association analyses for the metabolic blood measurements using UK Biobank exome sequences
The UK Biobank exome sequences were used to conduct single-variant and gene-level (collapsing) association analyses for 325 Nightingale assay-derived blood metabolite measurements and 30 clinical blood biomarkers. (*The pan-ancestry analysis included the four major [n > 1,000] ancestral groups observed in the UK Biobank: Europeans [n = 394,694], Africans [n = 7,412], South Asians [n = 8,078], and East Asians [n = 2,209]).

Pan-ancestry association analysis for metabolic blood measurements

A number of stringent variant-level QC steps, detailed previously,[26] were applied to select variant calls with highest confidence for association testing. Briefly, the variant-level QC criteria included coverage depth, genotype and mapping quality scores, DRAGEN variant status, read position rank sum score (RPRS), mapping quality rank sum score (MQRS), alternate allele read proportion for heterozygous calls, proportion of samples failing any of these QC criteria, and gnomAD-related filters.

Association testing between the metabolic blood measurements and the variants in the exome-sequencing dataset was conducted using two complementary analytical approaches (Figure 1): single variant exome-wide association study (ExWAS) and gene-level collapsing analysis.

We conducted the association analyses for all four ancestries combined ("pan-ancestry" analysis), and separately in the European ancestry participants as this comprised the single largest ancestral group in this resource. Further details about these statistical analyses and corresponding models are described in the following sections.

Single-variant exome-wide association study (ExWAS)

In the single-variant analysis (hereafter referred to as ExWAS), variants that passed the QC steps were filtered further to include those that were present in a minimum of six individuals (equivalent to MAF > 0.0008% in the European ancestry subset). We additionally excluded variants that had one of the following annotations as their most damaging effect as per SNPEff: *3_prime_UTR*, *5_prime_UTR*, *initiator_codon_variant*, *non_coding_transcript_exon_variant*, and *synonymous_variant*. The remaining non-synonymous coding variants (N = 2,043,019 in the European ancestry subset) were used to perform the ExWAS.

The ExWAS was conducted by fitting a linear regression model adjusted for age, sex, and BMI (for the Nightingale metabolite measurements only), using the tool PEACOK (https://github.com/astrazeneca-cgr-publications/PEACOK/) that was developed as a modification of the R package PHESANT.[31] For the pan-ancestry analysis, we additionally included the categorical ancestral group and top five ancestry principal components as covariates. For each of the 355 metabolic measurements, three different genetic models were evaluated in the ExWAS: (1) genotypic (AA vs. AB vs. BB), (2) dominant (AA + AB vs. BB), and (3) recessive (AA vs. AB + BB), where A and B denote the reference and alternative alleles, respectively. A significance cut-off of $p < 1 \times 10^{-8}$ was adopted for the ExWAS, as defined in our previous phenome-wide association study (PheWAS).[26,32].

Gene-level collapsing analysis

In order to boost power to detect associations for rare variants (including private mutations) having the same direction of effect, we adopted a collapsing framework to test the aggregate effect of rare functional variants in a gene. Overall, 10 different collapsing models (9 dominant and 1 recessive) were implemented per gene to evaluate a range of genetic architectures. Additionally, a synonymous collapsing model was used solely for the purpose of establishing an empirical negative control.[26]

As outlined in Table S4, the criteria for qualifying variants (QVs)[33] for the collapsing models were based on the following parameters: type of variant (missense, non-synonymous, or

PTV), minor allele frequency, *in silico* deleteriousness predictors (REVEL and MTR), and type of genetic model (dominant or recessive). The following variant annotations were used to define PTVs: *exon_loss_variant*, *frameshift_variant*, *start_lost*, *stop_gained*, *stop_lost*, *splice_acceptor_variant*, *splice_donor_variant*, *gene_fusion*, *bidirectional_gene_fusion*, *rare_amino_acid_variant*, and *transcript_ablation*. Hemizygous genotypes for the X chromosome also qualified for the recessive model.

For a given collapsing model, the effect of QVs in each studied gene (n = 18,762) was calculated as the difference in the mean of a metabolic blood measurement between individuals harboring and not harboring the QVs, using a linear regression model in PEACOK.[31] Covariates used in the pan-ancestry linear regression model were identical to that described for the ExWAS, described above.

A significance cut-off of $p < 1 \times 10^{-8}$ was set for the collapsing analysis based on the observed p value distribution for the synonymous model and an n-of-1 permutation of the phenotypes, as has been described previously.[26] Consistent with the threshold identified in our previous work, we observed that only 2 associations (of total ~6.5 million) for the synonymous model and no association (of total ~69.6 million) for the n-of-1 permutation achieved a p value less than 1×10^{-8} in the pan-ancestry analysis of the 355 metabolic blood measurements (Figure S2).

Medication-adjusted collapsing analysis

For a subset of the metabolic blood measurements (304 Nightingale metabolite measurements and 4 clinical biomarkers), additional collapsing analyses were also conducted in the European ancestry participants after adjusting for medication usage, using the following approach:

Nightingale metabolite measurements: "statin use" (binary variable) was included as an additional covariate in the analysis. This was done for 304 (of the total 325) lipid-related metabolite measurements.

Clinical biomarkers: medication-adjusted values were precalculated based on "adjustment factors" that had been estimated in the UKB. Apolipoprotein B, LDL-cholesterol, and total cholesterol values were adjusted using their respective "statin adjustment factor," and urate values were adjusted using "allopurinol adjustment factor."

Association analysis of clinical phenotypes documented in the UKB (PheWAS)

We harmonized and union mapped the clinical phenotypes available in the UKB, as previously described.[26] Phenome-wide collapsing analysis (PheWAS) for 15,719 clinical phenotypes was performed for the 11 collapsing models, as described in our previously published study.[26] Similar to the analysis of the metabolic blood measurements, we conducted both a pan-ancestry and a European-only PheWAS for the clinical phenotypes. We queried the results of the PheWAS for genes of interest that emerged from the analysis of the metabolic blood measurements. All results are available in the public PheWAS portal at https://azphewas.com.

Additionally, we also performed an association analysis between each of the 355 metabolic blood measurements (325 Nightingale blood metabolite measurements and 30 clinical biomarkers) and the 15,719 clinical phenotypes using a linear regression model to capture any association patterns between the metabolic measurements and the ICD-10 chapter mappings of the clinical phenotypes.

Comparing results from collapsing analysis to microarray-based genome-wide association study

We explored the hypothesis that applying a collapsing framework—that tests the aggregate effect of rare functional variants in a gene identified using exome sequencing—could extend our knowledge of gene-biomarker relationships beyond those identified via microarray-based studies. In order to do that, we compared our findings with a study that conducted single-variant association analysis for clinical biomarkers in the UKB using microarray data[2] (hereafter referred to as Sinnott-Armstrong et al. GWAS).

Of the 30 clinical blood biomarkers that we studied, 28 were also analyzed in the Sinnott-Armstrong et al. GWAS. To be consistent with what was done in the Sinnott-Armstrong et al. GWAS, we used the statin-adjusted values for LDL-cholesterol, total cholesterol, and apolipoprotein B, and the medication-unadjusted values for the remaining biomarkers. An additional seven biomarkers, that comprised of four directly measured urinary biomarkers and an additional three derived measurements, were also analyzed in the Sinnott-Armstrong et al. GWAS. For the purpose of comparing findings, we additionally performed gene-level collapsing analysis for the four directly measured urinary biomarkers in the UKB (i.e., sodium in urine, potassium in urine, microalbumin in urine, and creatinine [enzymatic] in urine). Thereafter, for the set of 32 biomarkers common to both studies (28 blood and 4 urinary biomarkers), we compared the gene-biomarker relationships that achieved significance ($p < 1 \times 10^{-8}$) in the collapsing analysis with the gene-biomarker relationships corresponding to the significant coding variant associations reported in the Sinnott-Armstrong et al. GWAS. We considered a comparatively relaxed significance threshold of $p = 1 \times 10^{-7}$ for the Sinnott-Armstrong et al. GWAS results in order to be conservative when attributing a gene-biomarker relationship as being specific to the collapsing analysis.

We also hypothesized that the various variant-level "purifying" filters implemented for selecting QVs in the collapsing analysis can enable a more direct estimate for the effect of gene aberrations (e.g., protein-truncating variants) on biomarker levels. To investigate this hypothesis, we compared the effect sizes for gene-biomarker relationships that achieved significance in both the gene-level collapsing analysis and the Sinnott-Armstrong et al. GWAS. For each such gene-biomarker relationship, we selected (1) the *model* with the highest absolute beta in the collapsing analysis and (2) the individual *variant* with the highest absolute beta as reported in the Sinnott-Armstrong et al. GWAS. For the latter, we adopted the absolute beta estimated in the genotypic model in our ExWAS (for the corresponding gene-biomarker relationship) as a substitute, to account for possible differences in trait transformation, association model, or covariates between our study and the Sinnott-Armstrong et al. GWAS. Critically, the absolute betas we estimated in the ExWAS were highly correlated with those estimated by the Sinnott-Armstrong et al. GWAS (Spearman's rho = 0.99) (Figure S3). We then compared the absolute beta of the collapsing model (step 1) with that of the individual variant (step 2). This approach provides a means to compare the effect size of aberrations in genes on biomarker levels estimated from

individual coding variants captured by microarrays with that estimated from an aggregate of rare coding variants identified by exome sequencing.

Drug targets enrichment analysis for genes associated with the metabolic blood measurements

A drug targets enrichment analysis was performed for the genes that were significantly associated with one of the Nightingale metabolite measurements or clinical biomarkers in the collapsing analysis. This was based on five publicly available lists: a custom list (n = 387; https://raw.githubusercontent.com/ericminikel/drug_target_lof/master/data/drugbank/drug_gene_match.tsv) that was originally derived from DrugBank[34] and another four lists that were derived from Informa Pharmaprojects database (https://pharmaintelligence.informa.com/products-and-services/clinical-planning/pharmaprojects).

For each gene that was significantly associated ($p < 1 \times 10^{-8}$) with one of the metabolic blood measurements in our collapsing analysis, only the most significant association was considered. The enrichment analysis was based on the same p value threshold as that used to define significance in this study, i.e., $p < 1 \times 10^{-8}$. The relationship between drug target status and significant genes identified in our analysis was assessed using Fisher exact test for each of the five gene lists. Specifically, for each of the gene lists, we created a contingency table that compared the intersection of the significant genes from the collapsing analysis with the gene list and the intersection of all the genes that we tested in our collapsing analysis (N = 18,762) with the gene list. We also performed a similar enrichment analysis for associations derived from the GWAS Catalog (https://www.ebi.ac.uk/gwas/) and the OMIM database (https://www.omim.org) to serve as reference. For the GWAS Catalog associations, we included the most significant associations per gene.

Results

In this study, we focused our analyses on 355 metabolic blood measurements available in the UKB: 325 Nightingale assay-derived metabolite measurements and an additional 30 clinical biomarkers. We classified the Nightingale metabolite measurements and the clinical biomarkers into the following 10 different biological classes or physiological systems based on known biology: lipids, amino acids, liver, glucose-related/glycolysis, renal, hormones, ketone bodies, bone and joint, inflammatory, and immune (Tables S1A and S1B). The Nightingale metabolite measurements that are available for 121,695 UKB participants mainly include lipoprotein lipids, fatty acids and their compositions, and various other low-molecular-weight metabolites[24] (Table S1A). The clinical biomarkers, which have been measured in nearly all the UKB participants, are more diverse and relate to glucose and lipid metabolism, renal and liver function, and others (Table S1B). The 355 metabolic blood measurements that we studied correlate with each other, particularly those belonging to the same biological class (Figure 2A). We, therefore, considered various Pearson's correlation coefficient (r^2) thresholds of 0.8, 0.5, and 0.2 to determine independence, which resulted in 113, 54, and 28 independent metabolic blood measurements in the dataset, respectively. Additionally, we evaluated the biological relevance of the metabolic blood measurements by testing their association with 15,719 clinical traits documented in the UKB[26] (material and methods) (Figure 2B).

Pan-ancestry collapsing analysis for 355 metabolic blood measurements

We performed a gene-level collapsing analysis to test the aggregate effect of rare functional variants in each gene in relation to the Nightingale blood metabolite measurements and the clinical blood biomarkers (Figure 1). We employed 10 previously established non-synonymous collapsing models to capture a diverse range of rare-variant genetic architectures (material and methods).[26] In a pan-ancestry analysis, which included participants from four major ancestral groups in the UKB (Europeans, Africans, South Asians, and East Asians) (Table S3), we identified 1,968 distinct relationships between genes and the Nightingale metabolite measurements and 331 between genes and clinical biomarkers ($p < 1 \times 10^{-8}$) (Tables S5A, S5B, 3A, and 3B). The protein-truncating variant (PTV)-exclusive models (with differing allele frequency cut-offs: 0.1% and 5%) accounted for 67% of the significant gene-metabolite and gene-biomarker relationships. These included many biologically plausible associations between genes and their direct protein products—e.g., *ALB* and albumin (beta = −2.48, 95% CI: [−2.65, −2.30], $p = 5.1 \times 10^{-161}$), *CST3* and cystatin C (beta = −3.36, 95% CI: [−3.49, −3.23], $p < 1 \times 10^{-300}$), and *SHBG* and sex hormone binding globulin (beta = −1.35, 95% CI: [−1.42, −1.28], $p < 1 \times 10^{-300}$)—as well as several other important associations—e.g., *PLIN1* and HDL-cholesterol (beta = 0.28, 95% CI: [0.21, 0.35], $p = 1.4 \times 10^{-15}$), *CREB3L3* and triglycerides (beta = 0.32, 95% CI: [0.22, 0.43], $p = 2.0 \times 10^{-9}$), and others (Tables S5A and S5B).

Exome-wide, 205 distinct genes were significantly associated with at least one metabolite measurement or clinical biomarker. Among the genes that showed the highest number of significant associations were well-established positive controls, including *APOC3*, *APOB*, *ANGPTL3*, *LDLR*, and *PCSK9*.[4,10,35] There were strong correlations among the metabolic blood measurements we analyzed, like within the lipids class (Figure 2A); however, the additional granularity that can be gleaned from the associations observed in this study, including ones that were not reported in previous rare variant studies of the human metabolome,[16–20] may provide insights into the underlying biology for some genes. For example, PTVs in the transcription factor-encoding *CREB3L3* were specifically associated with increased levels of "extremely large VLDL" fractions among the measured Nightingale blood metabolites (Table S5A). This finding supports the notion that the link between *CREB3L3* and fatty liver disease[36] may in fact be a manifestation of metabolic syndrome in general,[37] since the overproduction of large VLDL particles is a known hallmark of metabolic syndrome.[38]

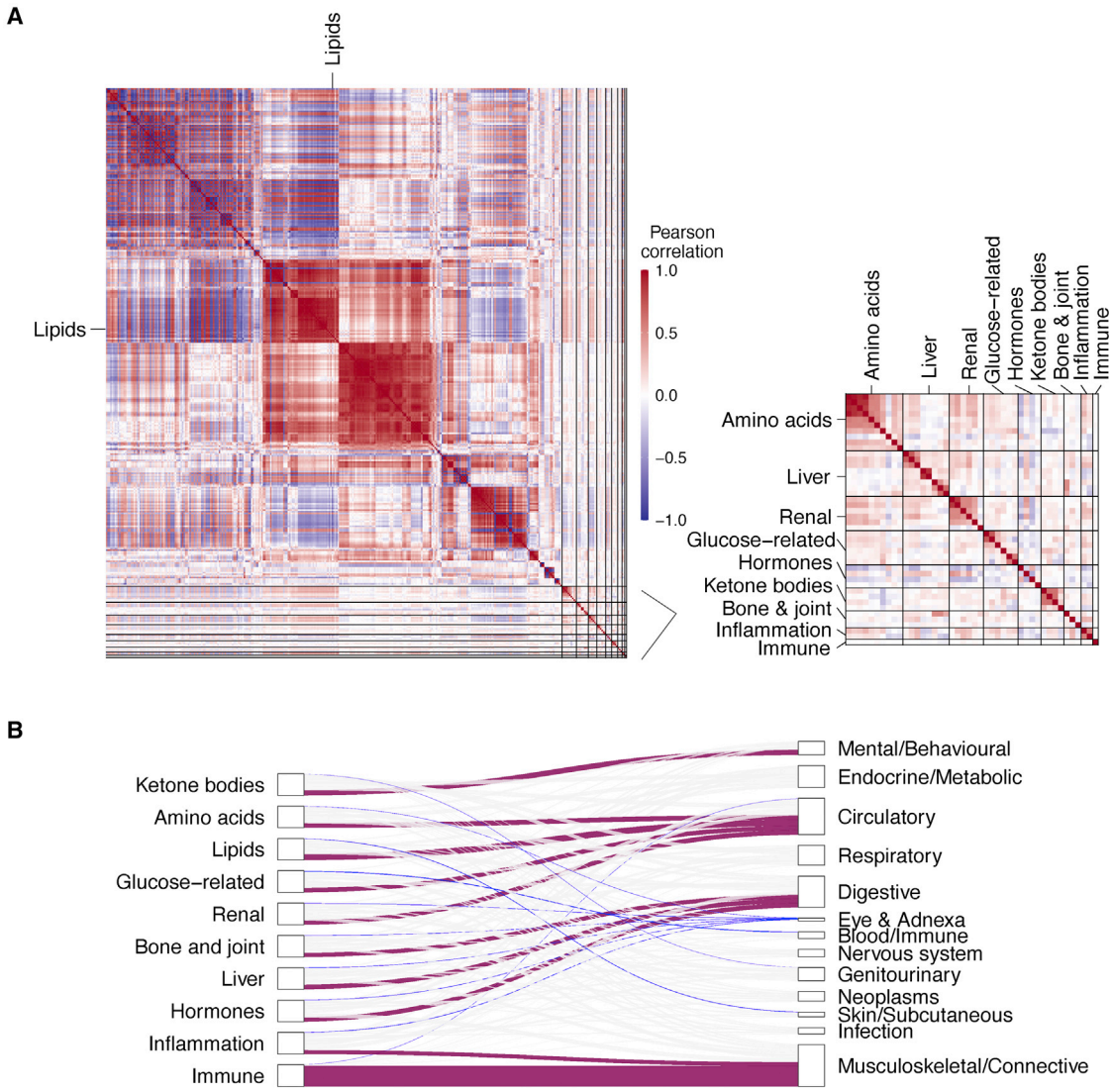

Figure 2. Characteristics of metabolic blood measurements analyzed in this study

The 355 metabolic blood measurements (325 Nightingale assay-derived metabolite measurements and 30 clinical biomarkers) analyzed in this study were grouped into the following 10 biological classes or physiological systems: lipids, amino acids, liver, glucose-related/glycolysis, renal, hormones, ketone bodies, bone and joint, inflammatory, and immune (Tables S1A and S1B).

(A) The correlation structure between any two metabolic blood measurements is shown, grouped by the biological classes.

(B) Association between the metabolic blood measurements and the 15,719 clinical traits documented in the UKB was tested.[26] For each class of metabolic biomarker measurements, the ICD-10 chapter for which the highest (in purple) and the least (in blue) number of significant associations ($p < 1 \times 10^{-8}$) with individual clinical traits was observed has been highlighted.

Another notable association was between PTVs in the *PLIN1*—loss of which causes a rare autosomal-dominant condition of partial lipodystrophy and severe dyslipidemia[39]—and decreased triglycerides (beta = −0.23, 95% CI: [−0.30, −0.16], $p = 1.6 \times 10^{-10}$) and increased HDL-cholesterol levels (beta = 0.28, 95% CI: [0.21, 0.35], $p = 1.4 \times 10^{-15}$). This finding of a favorable lipid profile among individuals harboring *PLIN1* PTVs in the UKB was also recently reported in another study.[40] Using the Nightingale metabolite data, we additionally found that the *PLIN1* PTVs showed strongest association signals with certain lipid fractions (triglycerides, phospholipids, and cholesteryl esters) contained in HDL particles (Table S5A), which may provide further insights on the biological role of *PLIN1*. Consistent with the metabolite and biomarker findings, individuals harboring *PLIN1* PTVs had a lower risk of developing atherosclerotic heart disease in the UKB phenome-wide association study or PheWAS (material and methods) (OR = 0.56 [95% CI: 0.38, 0.82], $p = 1.6 \times 10^{-3}$), although this association is not study-wide significant.

We next systematically evaluated the clinical phenotypes associated with the 49 genes that were significantly associated with at least one blood metabolite measurement (material and methods). Twelve of these genes (24%) were associated with at least one clinical phenotype at a p value threshold of $p < 1 \times 10^{-6}$ (Table S6). As expected, given the nature of the measured metabolites, most of these

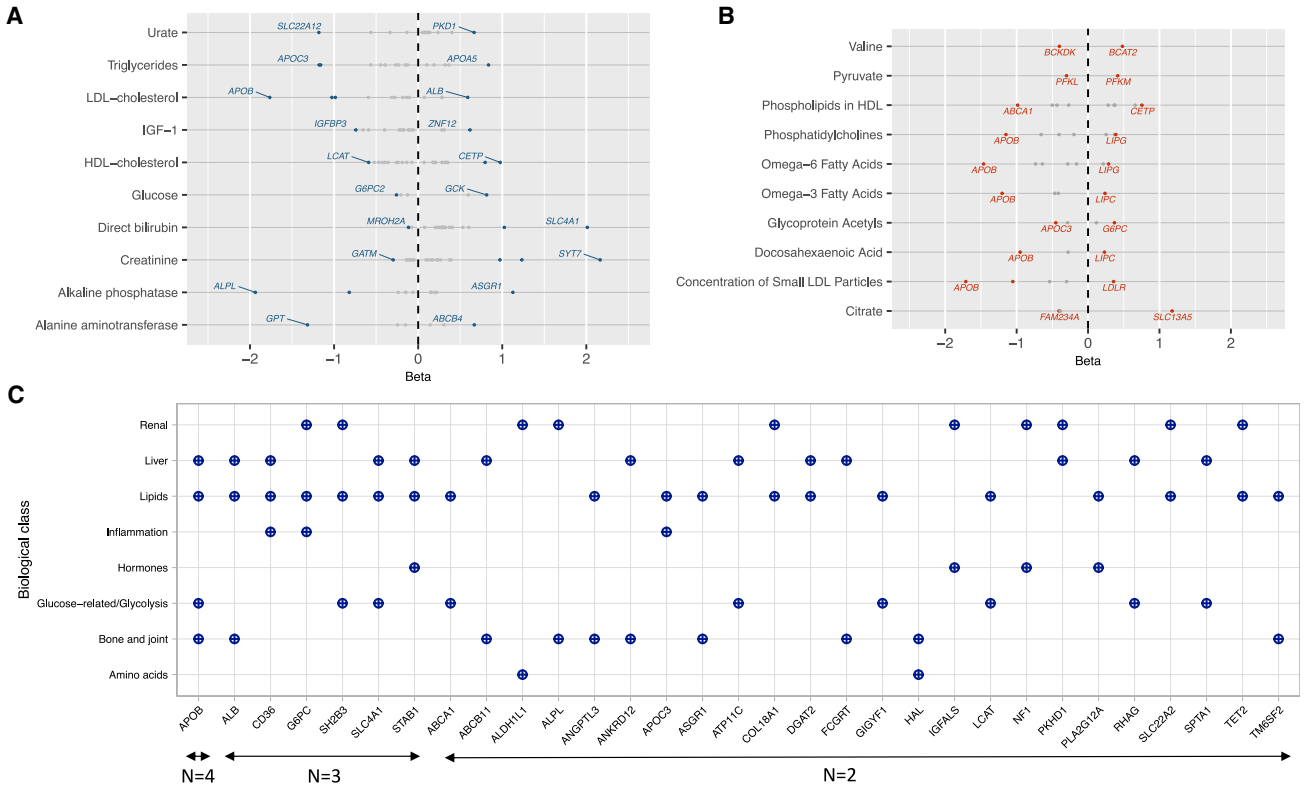

Figure 3. Significant relationships between genes and metabolic blood measurements identified in the pan-ancestry gene-level collapsing analysis

(A and B) Among the significant gene relationships (Fisher's exact test $p < 1 \times 10^{-8}$) identified in the gene-level collapsing analysis (indicated using dots) for select clinical biomarkers (3A) and Nightingale metabolite measurements (3B), the genes with the highest effect sizes have been shown.

(C) The plot shows the 31 genes that were significantly associated (Fisher's exact test $p < 1 \times 10^{-8}$) with metabolic blood measurements that spanned multiple biological classes or physiological systems in the gene-level collapsing analysis. For each such gene, the associated biological classes have been indicated.

associations were with cardiovascular, renal, and metabolic diseases, such as *PKD1* with polycystic kidney disease and *LDLR* with hypercholesterolemia.

We next conducted separate gene-level collapsing analyses in each individual ancestry groups (European, South Asian, East Asian, and African) to detect ancestry-specific associations (Tables S7A–S7C). In the European-only analysis—the largest ancestral group in our study—we observed a decrease in the number of significant associations ($p < 1 \times 10^{-8}$) compared to the pan-ancestry analysis: 1,936 significant relationships between genes and the Nightingale metabolite measurements (versus 1,968) and 323 between genes and clinical biomarkers (versus 331). This finding was consistent with the fact that the European ancestry participants (n = 394,694) represented a subset of the pan-ancestry UKB sequence panel (N = 412,393). There were, however, 34 (2%) gene-metabolite and 14 (4%) gene-biomarker study-wide significant associations in the European-only analysis that were not study-wide significant in the pan-ancestry analysis (Figure S4). This included previously reported associations between *DRD5* and urate,[41,42] *SLC7A9* and cystatin C,[42] and others, reflecting stronger effects in Europeans compared to the other ancestries. Conversely, for the gene-metabolite and gene-biomarker relationships that were study-wide significant in the pan-ancestry analysis but not in the Europeans-only analysis (66/1,968 [3%] and 22/331 [6%], respectively) (Figure S4), no significant difference was observed in effect sizes between the two analyses (Mann-Whitney p = 0.57), suggesting that this was largely driven by greater statistical power in the pan-ancestry analysis rather than ancestry-specific effects.

One such association detected exclusively in the pan-ancestry analysis was between rare non-synonymous variants in the membrane-trafficking gene *SYT7* and blood creatinine levels (model = "rec," number of individuals harboring QVs = 5, beta = 2.17 [95% CI: 1.46, 2.87], $p = 1.6 \times 10^{-9}$). Since 3 of the 5 individuals harboring recessive *SYT7* QVs were observed among the South Asian and African ancestry participants, the pan-ancestry analysis empowered detection of this association, which was not study-wide significant in the European ancestry subset (beta = 1.17 [95% CI: 0.06, 2.28], p = 0.04). We further sought to assess whether recessive QVs in *SYT7* were associated with any clinical phenotypes in the pan-ancestry UKB PheWAS (material and methods). Consistent with the biomarker findings, we observed a strong increased risk of glomerular disease among individuals harboring

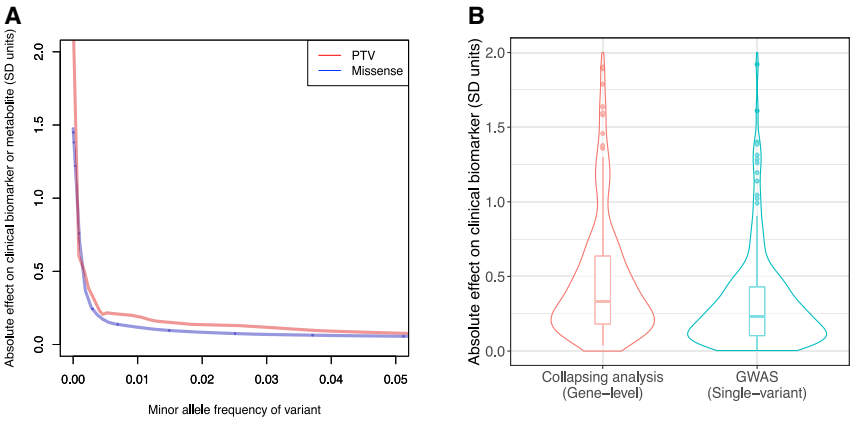

Figure 4. Analysis of coding variants captured using microarrays and exome sequencing in relation to metabolic blood measurements in the European ancestry participants

(A) Absolute effect sizes for missense variants and PTVs significantly associated ($p < 1 \times 10^{-8}$) with metabolic blood measurements in the single variant analysis (ExWAS) as a function of their minor allele frequency. In cases where a missense variant or a PTV was significantly associated with more than one measurement, the association with the highest absolute effect size was plotted.

(B) For the gene-clinical biomarker relationships that were significantly associated in both our gene-level collapsing analysis and the coding variant analysis in the Sinnott-Armstrong et al. GWAS (n = 215), the effect size estimates from the two analyses were compared. For each such gene-clinical biomarker relationship, this was done by selecting the collapsing model from the collapsing analysis and the individual variant from the GWAS with the highest absolute effect sizes. The effect sizes estimated in the collapsing analysis were significantly higher than that in the GWAS (Mann-Whitney $p = 8.0 \times 10^{-6}$).

recessive *SYT7* QVs (OR = 92.1 [95% CI: 12.1, 713.2], $p = 2.6 \times 10^{-5}$); however, this clinical association is not study-wide significant.

Medication adjustment of Nightingale metabolites and clinical biomarker levels can improve signal detection

To account for the effect of commonly prescribed medications on metabolic blood measurements, we additionally performed collapsing analysis for the lipid-related Nightingale metabolites (n = 304) and four clinical biomarkers (LDL-cholesterol, total cholesterol, apolipoprotein B, and urate) in the European ancestry subset (material and methods). We observed an improvement in signal detection, with 202 (10%) gene-metabolite relationships and 31 (37%) gene-biomarker relationships that were significant only in the medication-adjusted analysis (Tables S7A and S7B). This includes the association between putatively damaging missense variants and PTVs ("flexdmg" model) in the target gene of statins, i.e., *HMGCR* and LDL-cholesterol (medication-adjusted: beta = −0.19 [95% CI: −0.24, −0.13], $p = 1.7 \times 10^{-11}$; medication-unadjusted: beta = −0.15 [95% CI: −0.21, −0.10], $p = 6.1 \times 10^{-8}$), which validates the significance of correcting for medication effect to untangle true effects of natural aberration in genes. Finally, for the gene-metabolite and gene-biomarker relationships that were significant in either the medication-adjusted or the -unadjusted analyses (n = 2,193), a comparison of the absolute effect sizes showed significantly higher estimates in the medication-adjusted analysis (Mann-Whitney p = 0.03) (Figure S5).

Genes associated with metabolic blood measurements across multiple biological classes

We observed 31 genes that were significantly associated with metabolic blood measurements across two or more biological classes or physiologic systems that we defined (e.g., renal, liver, lipids, etc.; Figure 3C). One such gene, *GIGYF1*, encodes a tyrosine kinase receptor signaling protein. Among the European ancestry participants, rare PTVs in this gene were significantly associated with HbA1c (beta = 0.73 [95% CI: 0.57, 0.88], $p = 4.5 \times 10^{-20}$), glucose (beta = 0.59 [95% CI: 0.42, 0.76], $p = 7.9 \times 10^{-12}$), and lipid-related biomarkers, including total cholesterol$_{adj}$ (beta = −0.60 [95% CI: −0.76, −0.44], $p = 2.8 \times 10^{-13}$), LDL-cholesterol$_{adj}$ (beta = −0.54 [95% CI: −0.70, −0.38], $p = 1.1 \times 10^{-10}$), and apolipoprotein B$_{adj}$ (beta = −0.50 [95% CI: −0.66, −0.33], $p = 2.4 \times 10^{-9}$)]. In the UKB PheWAS (material and methods), rare PTVs in *GIGYF1* were significantly associated with an increased risk of type 2 diabetes, consistent with the findings for glucose-related biomarkers (OR = 4.0 [95% CI: 2.7, 5.8], $p = 1.0 \times 10^{-10}$), as well as with an increased risk of hypothyroidism (OR = 4.2 [95% CI: 2.7, 6.6], $p = 7.1 \times 10^{-9}$) (https://azphewas.com). Since hypothyroidism is known to raise LDL-cholesterol levels, we subsequently tested the *GIGYF1*-LDL-cholesterol association adjusting for hypothyroidism diagnosis. The association signal between *GIGYF1* PTVs and statin-adjusted LDL-cholesterol levels remained significant following the adjustment (beta = −0.55 [95% CI: −0.71, −0.38]; $p = 6.2 \times 10^{-11}$). By leveraging information from more than 400,000 UKB exomes, our study provides a comprehensive picture regarding *GIGYF1*'s biomarker fingerprint and associated clinical traits, consistent with previous common[10] and rare variant associations[43,44] reported at this locus.

Assessing allelic series via gene-level collapsing analysis

In addition to the gene-level collapsing analysis, we conducted a variant-level analysis (ExWAS) between non-synonymous coding variants and the 355 metabolic blood measurements (Figure 1). There was an inverse relationship between the minor allele frequency (MAF) of the associated variants and their absolute effect size on the metabolic blood measurements (Figure 4A). Among variants that were significantly associated ($p < 1 \times 10^{-8}$) with metabolic blood measurements in the ExWAS was

the splice variant in *HSD17B13* (GenBank: NM_178135.5: c.812+2dup) that is known to protect against chronic liver disease.[45] In addition to the expected associations with aspartate aminotransferase (AST) and alanine aminotransferase (ALT), this variant was associated with several other Nightingale metabolite measurements relevant to cholesterol and triglyceride metabolism (Table S8A). These associations are consistent with the putative role of *HSD17B13* in lipogenesis.[46] Importantly, the metabolite signature that we observe with natural inhibition of HSD17B13 (Table S9) could offer an additional, and potentially more specific biomarker profile, for monitoring treatment response with therapeutic HSD17B13 inhibitors in the future.

We compared the ExWAS signals to the gene-level collapsing results and found that 16% (316/1,968) of gene-metabolite relationships and 31% (101/331) of gene-biomarker relationships that were study-wide significant in the pan-ancestry collapsing analysis did not achieve significance in the corresponding variant-level analysis (Tables S8A and S8B). Next, we also compared significant findings from the Europeans-only collapsing analysis (Table S7B and S10) with a previous microarray-based GWAS in the same population[2] for the set of 32 clinical biomarkers common to the two studies (material and methods). Approximately 40% (142/357) of the significant gene-biomarker relationships from the collapsing analysis were not detected in that GWAS (Table S11), including associations for well-known drug target genes such as *HMGCR* (with LDL-cholesterol) and *PPARG* (with HDL-cholesterol). Moreover, for the 215 gene-biomarker relationships that were detected in both the collapsing analysis and the GWAS, it was observed that effect sizes estimated in the collapsing analysis—based on an aggregate effect of rare non-synonymous variants—were significantly higher (Mann-Whitney $p = 8.0 \times 10^{-6}$) (Table S12 and Figure 4B). Collectively, these results demonstrate that applying a rare variant gene-based collapsing analysis to large-scale exome-sequencing data increases power to capture associations that are driven by an allelic series, and thus expand our understanding of the genetic architecture of traits.

SLC4A1, for which 25 (89%) of total 28 PTVs in our dataset were detected in just a single individual (Figure S6), was one such gene that showed multiple signals in the gene-level collapsing analysis but none in the ExWAS. Specifically, *SLC4A1* PTVs were significantly associated with a strong reduction in HbA1c (beta = −2.21 [95% CI: −2.62, −1.79], $p = 1.4 \times 10^{-25}$) and LDL-cholesterol$_{adj}$ (beta = −1.28 [95% CI: −1.63, −0.94], $p = 3.7 \times 10^{-13}$), while also showing strong increases in total bilirubin (beta = 1.67 [95% CI: 1.34, 2.00], $p = 1.1 \times 10^{-22}$) and direct bilirubin (beta = 2.01 [95% CI: 1.65, 2.37], $p = 1.8 \times 10^{-28}$). Rare variants in this gene have been previously implicated in several hematological phenotypes, including hereditary spherocytosis, cryohydrocytosis, and renal tubular acidosis with hemolytic anemia. In the UKB PheWAS of clinical phenotypes (material and methods), *SLC4A1* PTVs were indeed statistically associated with hereditary spherocytosis and hereditary hemolytic anemia, consistent with their well-established prior associations (https://azphewas.com). Importantly, changes in RBC morphology as well as hemolytic anemia are known to cause a falsely low HbA1c reading. We, therefore, tested the *SLC4A1* gene-level biomarker associations adjusting for relevant red blood cell indices (red cell distribution width and mean corpuscular hemoglobin concentration) and clinical diagnoses (hereditary spherocytosis and hereditary hemolytic anemia). The *SLC4A1* associations remained significant even in the adjusted analyses (Table S13). Although this could suggest that PTVs in this gene influence multiple metabolic pathways, it is more likely that these results reflect under-reporting of hereditary spherocytosis and hereditary hemolytic anemia in the UKB. This demonstrates the importance of anchoring gene-biomarker associations with relevant clinical endpoints to account for potential pathophysiological confounders of biomarkers readouts.

Discussion

We used 454,796 UKB exome sequences to explore the contribution of protein-coding variation for 325 predominantly lipid-related Nightingale assay-derived blood metabolite measurements (n = 99,283) and 30 clinical blood biomarkers (n = 393,351) using both variant- and gene-level analyses. By exploring the full allelic frequency spectrum of protein-coding variation in relation to a range of metabolic blood measurements, our study expands the understanding of the genetic architecture of these measures in the UKB. Previous studies have explored the genetic basis of metabolic traits in the UKB, but they were either based on microarray data, focused on specific clinical biomarkers of interest,[44] or used an earlier tranche of the exome sequence data.[43] We observed that 40% of the study-wide significant associations for the clinical biomarkers identified in this study were not previously reported on analyzing coding variants in a GWAS in the same cohort,[2] which reinforces the importance of studying rare variation to fully understand the genetic architecture of metabolic blood measurements. We also illustrated the significance of adjusting metabolite and clinical biomarker values for commonly prescribed medications to improve signal detection, with the association between non-synonymous variants in *HMGCR* and LDL-cholesterol a good exemplar of this.

In this study, we analyzed the first tranche of the Nightingale assay-derived blood metabolite measurements that were released to the UKB research community only in 2021; thus, unlike the standard clinical biomarkers, the Nightingale blood metabolite measurements have not yet been extensively studied in this resource. Likewise, the role of rare coding variants in regulating the human

metabolome, including lipid-related metabolites, has previously been explored in certain exome array and sequencing-based studies in other cohorts.[16–20] These rare-variant studies have expanded our understanding of the human metabolome; however, the sample sizes for these studies included fewer than 10,000 individuals, thus limiting the statistical power to detect associations, especially for rarer variants. Here, in studying metabolite associations in 99,283 individuals, we detected 1,968 gene-metabolite associations. The associations we observed for *CREB3L3* and *PLIN1* variants with specific lipid metabolite measurements, for instance, elucidate gene-metabolite relationships from our study that were not reported in any of the previous rare variant-based metabolomics studies.[16–20]

There are several strengths of our study that might have implications for identifying or validating drug targets. First, by virtue of focusing on coding variants, the observed associations more often provide a direct causal link between variants in a gene and a metabolic trait.[35,45,47,48] Moreover, collapsing analyses draw their power from aggregating the signals of multiple rare variants (allelic series), which tend to be less impacted by local linkage disequilibrium structure, and are thus more likely to be enriched for causal genetic variants. Second, associations involving putative functional variants can also often provide clues to the desired therapeutic modulation, e.g., upregulation or downregulation of the target gene product, to mimic the protective effect for a given disease. For instance, we observed a total of 181 associations for rare (MAF $< 0.1\%$) PTVs with the 30 clinical biomarkers in the European ancestry participants (Table S14). This is greater than 3-fold enrichment compared to the 53 conditionally independent PTV associations reported in the microarray-based analysis of these clinical biomarkers in the same UKB population.[2] In an enrichment analysis that was performed based on drug targets derived from DrugBank and Informa Pharmaprojects database (material and methods), we observed a significant enrichment of 3.1×10^{-7}, 2.3×10^{-10}, 5.3×10^{-13}, and 2.2×10^{-8} for approved, phase I, phase II, and phase III drug targets, respectively, among the genes that were significantly associated with one of the Nightingale metabolite measurements or clinical biomarkers in our collapsing analysis (Figure S7). This emphasizes the importance of genetic perturbations—identified using large-scale exome sequencing—that associate with clinically relevant biomarkers as potential therapeutic targets for human diseases.

We used metabolic blood biomarkers as intermediate traits to study the genetic basis of complex human diseases because they can both bolster gene discovery and provide important insight into disease biology. Consistent with this, many of the genes that were significantly associated with one of the Nightingale metabolite measurements were also associated with relevant clinical phenotypes (Table S6). We also observe associations between certain biomarkers and variants in genes that encode them (e.g., *ALB* with albumin and *CST3* with cystatin C); although such associations serve as excellent positive controls that demonstrate the robustness of our analysis framework, they may not offer insights into disease pathophysiology.

This work collectively demonstrates the value of using a large collection of exome sequences linked to diverse metabolic blood measurements to better understand human diseases. Our study will not only be an important resource to identify therapeutic targets but also provide insights into the metabolic biomarker profile for human genetic perturbations that might be of interest in preclinical development.

Data and code availability

The association statistics for the analyses conducted in this study are available in the public PheWAS portal (https://azphewas.com). The association statistics from previously published GWASs of Nightingale metabolites[1] and clinical biomarkers[2] in the UKB are available through the Open Targets platform (https://genetics.opentargets.org/). Association tests described in this study were performed using a custom framework, PEACOK (PEACOK 1.0.7), that is available via GitHub (https://github.com/astrazeneca-cgr-publications/PEACOK/).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.02.002.

Acknowledgments

We thank the participants and investigators in the UKB study who made this work possible (Resource Application Number 26041); the UKB Exome Sequencing Consortium (UKB-ESC) members AbbVie, Alnylam Pharmaceuticals, AstraZeneca, Biogen, Bristol-Myers Squibb, Pfizer, Regeneron, and Takeda for funding the generation of the exome sequence data; the Regeneron Genetics Center for completing the sequencing and initial quality control of the exome sequencing data; and the AstraZeneca Center for Genomics Research Analytics and Informatics team for processing and analysis of sequencing data.

Author contributions

S.P. designed the study. A.N., R.S.D., L.M., X.J., D.V., E.W., Q.W., and S.P. performed the analyses and statistical interpretation. A.N., R.S.D., A.R.H., D.S.P., and S.P. drafted the manuscript. All authors contributed to the review and critical revision of the manuscript.

Declaration of interests

A.N., R.S.D., L.M., X.J., D.V., E.W., E.L.A., A.R., K.C., K.R.S., Q.W., B.C., D.S.P., A.R.H., and S.P. are current employees and/or stockholders of AstraZeneca.

Received: November 21, 2022
Accepted: January 30, 2023
Published: February 20, 2023

Web resources

GenBank, https://www.ncbi.nlm.nih.gov/genbank/

References

1. Smith, C.J., Sinnott-Armstrong, N., Cichońska, A., Julkunen, H., Fauman, E.B., Würtz, P., and Pritchard, J.K. (2022). Integrative analysis of metabolite GWAS illuminates the molecular basis of pleiotropy and genetic correlation. Elife *11*, e79348.
2. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Ollila, H.M., Kiiskinen, T., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat. Genet. *53*, 185–194.
3. Suhre, K., Raffler, J., and Kastenmüller, G. (2016). Biochemical insights from population studies with genetics and metabolomics. Arch. Biochem. Biophys. *589*, 168–176.
4. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
5. Wuttke, M., Li, Y., Li, M., Sieber, K.B., Feitosa, M.F., Gorski, M., Tin, A., Wang, L., Chu, A.Y., Hoppmann, A., et al. (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat. Genet. *51*, 957–972.
6. Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.P., Kangas, A.J., Soininen, P., Würtz, P., Silander, K., et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat. Genet. *44*, 269–276.
7. Yet, I., Menni, C., Shin, S.-Y., Mangino, M., Soranzo, N., Adamski, J., Suhre, K., Spector, T.D., Kastenmüller, G., and Bell, J.T. (2016). Genetic influences on metabolite levels: a comparison across metabolomic platforms. PLoS One *11*, e0153672.
8. Suhre, K., Wallaschofski, H., Raffler, J., Friedrich, N., Haring, R., Michael, K., Wasner, C., Krebs, A., Kronenberg, F., Chang, D., et al. (2011). A genome-wide association study of metabolic traits in human urine. Nat. Genet. *43*, 565–569.
9. Shin, S.-Y., Fauman, E.B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., et al. (2014). An atlas of genetic influences on human blood metabolites. Nat. Genet. *46*, 543–550.
10. Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.v., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., et al. (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the million veteran program. Nat. Genet. *50*, 1514–1523.
11. Chambers, J.C., Zhang, W., Sehmi, J., Li, X., Wass, M.N., van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S.E., et al. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nat. Genet. *43*, 1131–1138.
12. Prins, B.P., Kuchenbaecker, K.B., Bao, Y., Smart, M., Zabaneh, D., Fatemifar, G., Luan, J., Wareham, N.J., Scott, R.A., Perry, J.R.B., et al. (2017). Genome-wide analysis of health-related biomarkers in the UK household longitudinal study reveals novel associations. Sci. Rep. *7*, 11008.
13. Wheeler, E., Leong, A., Liu, C.-T., Hivert, M.-F., Strawbridge, R.J., Podmore, C., Li, M., Yao, J., Sim, X., Hong, J., et al. (2017). 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.
14. Long, T., Hicks, M., Yu, H.-C., Biggs, W.H., Kirkness, E.F., Menni, C., Zierer, J., Small, K.S., Mangino, M., Messier, H., et al. (2017). Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nat. Genet. *49*, 568–578.
15. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The post-GWAS era: from association to function. Am. J. Hum. Genet. *102*, 717–730.
16. Yousri, N.A., Fakhro, K.A., Robay, A., Rodriguez-Flores, J.L., Mohney, R.P., Zeriri, H., Odeh, T., Kader, S.A., Aldous, E.K., Thareja, G., et al. (2018). Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. Nat. Commun. *9*, 333.
17. Yin, X., Chan, L.S., Bose, D., Jackson, A.U., VandeHaar, P., Locke, A.E., Fuchsberger, C., Stringham, H.M., Welch, R., Yu, K., et al. (2022). Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. Nat. Commun. *13*, 1644.
18. Rhee, E.P., Yang, Q., Yu, B., Liu, X., Cheng, S., Deik, A., Pierce, K.A., Bullock, K., Ho, J.E., Levy, D., et al. (2016). An exome array study of the plasma metabolome. Nat. Commun. *7*, 12360.
19. Davis, J.P., Huyghe, J.R., Locke, A.E., Jackson, A.U., Sim, X., Stringham, H.M., Teslovich, T.M., Welch, R.P., Fuchsberger, C., Narisu, N., et al. (2017). Common, low-frequency, and rare genetic variants associated with lipoprotein subclasses and triglyceride measures in Finnish men from the METSIM study. PLoS Genet. *13*, e1007079.
20. Bomba, L., Walter, K., Guo, Q., Surendran, P., Kundu, K., Nongmaithem, S., Karim, M.A., Stewart, I.D., Langenberg, C., Danesh, J., et al. (2022). Whole-exome sequencing identifies rare genetic variants associated with human plasma metabolites. Am. J. Hum. Genet. *109*, 1038–1054.
21. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
22. UK10K Consortium, Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
23. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.
24. Ritchie, S.C., Surendran, P., Karthikeyan, S., Lambert, S.A., Bolton, T., Pennells, L., Danesh, J., di Angelantonio, E., Butterworth, A.S., and Inouye, M. (2021). Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. Preprint at medRxiv.
25. Szustakowski, J.D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P.G., Sasson, A., Wong, E., Liu, D., Wade Davis, J., Haefliger, C., et al. (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat. Genet. *53*, 942–948.
26. Wang, Q., Dhindsa, R.S., Carss, K., Harper, A.R., Nag, A., Tachmazidou, I., Vitsios, D., Deevi, S.V.v., Mackay, A., Muthas, D., et al. (2021). Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature *597*, 527–532.

27. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80–92.
28. Traynelis, J., Silk, M., Wang, Q., Berkovic, S.F., Liu, L., Ascher, D.B., Balding, D.J., and Petrovski, S. (2017). Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res. *27*, 1715–1729.
29. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am. J. Hum. Genet. *99*, 877–885.
30. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
31. Millard, L.A.C., Davies, N.M., Gaunt, T.R., Davey Smith, G., and Tilling, K. (2018). Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int. J. Epidemiol. *47*, 29–35.
32. Fadista, J., Manning, A.K., Florez, J.C., and Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur. J. Hum. Genet. *24*, 1202–1205.
33. Petrovski, S., Todd, J.L., Durheim, M.T., Wang, Q., Chien, J.W., Kelly, F.L., Frankel, C., Mebane, C.M., Ren, Z., Bridgers, J., et al. (2017). An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. Am. J. Respir. Crit. Care Med. *196*, 82–93.
34. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082.
35. Cohen, J.C., Boerwinkle, E., Mosley, T.H., and Hobbs, H.H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N. Engl. J. Med. *354*, 1264–1272.
36. Nakagawa, Y., Oikawa, F., Mizuno, S., Ohno, H., Yagishita, Y., Satoh, A., Osaki, Y., Takei, K., Kikuchi, T., Han, S.-I., et al. (2016). Hyperlipidemia and hepatitis in liver-specific CREB3L3 knockout mice generated using a one-step CRISPR/Cas9 system. Sci. Rep. *6*, 27857.
37. McCann, M.A., Li, Y., Muñoz, M., Gil, V., Qiang, G., Cordoba-Chacon, J., Blüher, M., Duncan, S., and Liew, C.W. (2021). Adipose expression of CREB3L3 modulates body weight during obesity. Sci. Rep. *11*, 19400.
38. Adiels, M., Olofsson, S.-O., Taskinen, M.-R., and Borén, J. (2008). Overproduction of very low-density lipoproteins is the hallmark of the dyslipidemia in the metabolic syndrome. Arterioscler. Thromb. Vasc. Biol. *28*, 1225–1236.
39. Gandotra, S., le Dour, C., Bottomley, W., Cervera, P., Giral, P., Reznik, Y., Charpentier, G., Auclair, M., Delépine, M., Barroso, I., et al. (2011). Perilipin deficiency and autosomal dominant partial lipodystrophy. N. Engl. J. Med. *364*, 740–748.
40. Patel, K.A., Burman, S., Laver, T.W., Hattersley, A.T., Frayling, T.M., and Weedon, M.N. (2022). PLIN1 haploinsufficiency causes a favorable metabolic profile. J. Clin. Endocrinol. Metab. *107*, e2318–e2323.
41. Sinnott-Armstrong, N., Naqvi, S., Rivas, M., and Pritchard, J.K. (2021). GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. Elife *10*, e58615.
42. Barton, A.R., Sherman, M.A., Mukamel, R.E., and Loh, P.-R. (2021). Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. Nat. Genet. *53*, 1260–1269.
43. Jurgens, S.J., Choi, S.H., Morrill, V.N., Chaffin, M., Pirruccello, J.P., Halford, J.L., Weng, L.-C., Nauffal, V., Roselli, C., Hall, A.W., et al. (2022). Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. Nat. Genet. *54*, 240–250.
44. Deaton, A.M., Parker, M.M., Ward, L.D., Flynn-Carroll, A.O., BonDurant, L., Hinkle, G., Akbari, P., Lotta, L.A., et al.; Regeneron Genetics Center; and DiscovEHR Collaboration (2021). Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. Sci. Rep. *11*, 21565.
45. Abul-Husn, N.S., Cheng, X., Li, A.H., Xin, Y., Schurmann, C., Stevis, P., Liu, Y., Kozlitina, J., Stender, S., Wood, G.C., et al. (2018). A protein-truncating HSD17B13 variant and protection from chronic liver disease. N. Engl. J. Med. *378*, 1096–1106.
46. Luukkonen, P.K., Tukiainen, T., Juuti, A., Sammalkorpi, H., Haridas, P.A.N., Niemelä, O., Arola, J., Orho-Melander, M., Hakkarainen, A., Kovanen, P.T., et al. (2020). Hydroxysteroid 17-β dehydrogenase 13 variant increases phospholipids and protects against fibrosis in nonalcoholic fatty liver disease. JCI Insight *5*, e132158.
47. Akbari, P., Gilani, A., Sosina, O., Kosmicki, J.A., Khrimian, L., Fang, Y.-Y., Persaud, T., Garcia, V., Sun, D., Li, A., et al. (2021). Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity. Science *373*, eabf8683.
48. Nag, A., Dhindsa, R.S., Mitchell, J., Vasavda, C., Harper, A.R., Vitsios, D., Ahnmark, A., Bilican, B., Madeyski-Bengtson, K., Zarrouki, B., et al. (2022). Human genetics uncovers MAP3K15 as an obesity-independent therapeutic target for diabetes. Sci. Adv. *8*, eadd5430.