Article

# Genetic analysis of circulating metabolic traits in 619,372 individuals

https://doi.org/10.1038/s41586-026-10532-5

Received: 5 May 2025

Accepted: 13 April 2026

Published online: 20 May 2026

Open access

Check for updates

Ralf Tambets[1], Mihkel Jesse[1], Jaanika Kronberg[2,3], Adriaan van der Graaf[4], Erik Abner[2], Urmo Võsa[2], Ida Rahu[1], Nele Taba[2], Anastassia Kolde[2,3], Dzvenymyra Yarish[1], Sariyya Abdullayeva[1], Anastasiia Alekseienko[2], Andres Veidenberg[2], Estonian Biobank Research Team*, Krista Fischer[2,3], Zoltán Kutalik[4,5], Tõnu Esko[2,6], Kaur Alasoo[1,6] ✉ & Priit Palta[2,6] ✉

Interpreting the association of genetic variants with complex traits can be improved by gaining a greater understanding of the molecular consequences of these variants. Although genome-wide association studies (GWAS) for complex diseases routinely profile over one million individuals[1–5], studies of molecular traits have lagged behind. Here we performed a GWAS meta-analysis for 249 circulating metabolic traits in the Estonian Biobank and the UK Biobank in up to 619,372 individuals. We identified 88,127 common and low-frequency locus–trait associations from 8,398 loci that converged on shared genes and pathways. Using statistical fine mapping, systematic phenome-wide colocalization and *cis*-Mendelian randomization, we explored putative causal links between metabolic traits and disease outcomes. We predict that although plasma branched-chain amino acids (BCAAs) have been associated with type 2 diabetes in observational studies[6,7], lowering BCAA levels by targeting the BCAA catabolism pathway is unlikely to reduce type 2 diabetes risk. Leveraging our large sample size and high-quality genotype imputation, we found that 19.4% of the confidently fine-mapped variants had minor allele frequencies between 0.1 and 1%, and these variants were twofold enriched for predicted missense and splice-altering variants. Our results highlight the value of integrating low-frequency variants into genetic association studies.

Recent large-scale GWAS of metabolic traits have continued to uncover novel associations and biological insights[8–14]. However, for more than half of the metabolic traits that are captured by nuclear magnetic resonance (NMR) spectroscopy, the proportion of heritability explained by genome-wide significant variants remains below 50% (ref. 12), indicating that much larger sample sizes are needed to identify the remaining genetic effects. Furthermore, most existing association studies using the Nightingale Health NMR platform have been limited to common variants[8–10,12] and exome sequencing[13,15], leaving the full genome-wide spectrum of low-frequency genetic variation unexplored. Finally, larger sample sizes and increased statistical power also bring new challenges for interpreting genetic associations, particularly when genetic variants have pleiotropic effects on several correlated metabolic traits[8–10]. In particular, there is a growing concern that naive use of these associations in the Mendelian randomization[16] framework can lead to spurious and misleading findings[17,18].

## Association testing and meta-analysis

We performed GWAS for 249 metabolic traits (Supplementary Table 1) in the Estonian Biobank (EstBB; $n$ = 185,352) and 6 genetic ancestry groups from the UK Biobank (UKBB; $n$ = 434,020) (Extended Data Fig. 1). The UKBB genetic ancestry groups were defined previously by the Pan-UKBB project[19] and are listed in Table 1. Relying on the population-specific genotype imputation panel for the EstBB[20] and the Genomics England[21] and TopMed[22] imputation panels for the UKBB allowed us to test 10–96 million variants across genetic ancestry groups (up to nine times more than previous studies using the same NMR platform[8,12,13]). On the basis of minor allele frequency (MAF), we stratified these variants into three bins: common variants (MAF > 1%), low-frequency variants (MAF between 0.1% and 1%) and rare variants (MAF < 0.1%). The number of significant locus–trait pairs ranged from 37 (UKBB_AMR) to 62,543 (UKBB_EUR), and the number of independent lead variants ($r^2$ < 0.8) ranged from 24 to 6,014, with most associations detected in the UKBB_EUR and EstBB subsets (Table 1). We observed high genetic correlation for matched metabolic traits between the EstBB ($n$ = 185,352) and UKBB_EUR ($n$ = 413,897) subsets (median genetic correlation (rg) = 0.91, mean rg = 0.89), indicating that genetic effects are largely shared between the two biobanks (Supplementary Table 2).

In the meta-analysis of EstBB and UKBB_EUR (meta_EUR; $n$ = 599,249), we identified 86,886 locus–trait pairs, corresponding to 8,260 independent lead variants ($r^2$ < 0.8). This represented an approximately

[1]Institute of Computer Science, University of Tartu, Tartu, Estonia. [2]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia. [3]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia. [4]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [5]University Center for Primary Care and Public Health, Unisanté, University of Lausanne, Lausanne, Switzerland. [6]These authors jointly supervised this work: Tõnu Esko, Kaur Alasoo, Priit Palta. *A list of authors appears at the end of the paper. ✉e-mail: kaur.alasoo@ut.ee; priit.palta@ut.ee

**Table 1 | Number of significant locus–metabolic trait pairs and unique lead variants ($r^2$ < 0.8) detected in each genetic ancestry group and the two meta-analyses**

| Biobank and genetic ancestry group | Locus–trait pairs | Unique lead variants | Sample size |
|---|---|---|---|
| EstBB | 26,736 (1,012) | 2,313 (90) | 185,352 |
| UKBB_EUR (European) | 62,543 (1,498) | 6,014 (142) | 413,897 |
| UKBB_CSA (Central/South Asian) | 1,070 | 113 | 8,652 |
| UKBB_AFR (African) | 916 | 143 | 6,439 |
| UKBB_EAS (East Asian) | 358 | 41 | 2,604 |
| UKBB_MID (Middle Eastern) | 92 | 44 | 1,500 |
| UKBB_AMR (Admixed American) | 37 | 24 | 928 |
| Meta_EUR | 86,886 (2,124) | 8,260 (156) | 599,249 |
| Meta_ALL | 88,127 (2,089) | 8,398 (156) | 619,372 |

The significance threshold was set to $P < 5\times10^{-8}$ for common and low-frequency variants (MAF > 0.1%) and to $P < 6.25\times10^{-10}$ for rare variants (MAF < 0.1%). The numbers of rare variant associations are shown in parentheses. A version of this table further accounting for the 249 metabolic traits tested is presented in Extended Data Table 1.

tenfold increase compared with Karjalainen et al.[8] ($n$ = 136,016; 8,578 locus–trait pairs) and a 63% increase compared with a parallel study by Zoodsma et al.[13] on the overlapping set of UKBB samples ($n$ = 450,016; 52,662 locus–trait pairs). The estimated heritability of individual metabolic traits ranged from 2.8% for acetoacetate to 19.5% for HDL_size (median 10.2%), and we observed a clear linear relationship between heritability and the number of loci associated with each metabolic trait (Supplementary Fig. 1 and Supplementary Table 3). On average, 93% of the lead variant associations detected by Karjalainen et al.[8] were replicated in our meta_EUR analysis with a highly concordant direction of effect (Supplementary Fig. 2). We also detected many novel associations for all tested metabolic traits. The fraction of novel associations ranged from 27% for 3-hydroxybutyrate to 85% for lactate (Fig. 1a). Altogether, we identified 7,790 novel independent lead variants ($r^2$ < 0.8) not previously reported by Karjalainen et al.[8], including 163 lead variants on the X chromosome. To identify additional conditionally distinct association signals, we performed statistical fine mapping around the 84,762 meta_EUR locus–trait pairs that had MAF > 0.1% in the UKBB_EUR subset. We restricted the fine-mapping analysis to the summary statistics from the UKBB_EUR subset, as this allowed us to use in-sample linkage disequilibrium (LD) calculated from the overlapping set of 413,897 individuals and thus avoid a major potential source of false positives[23]. We identified 116,467 independent credible sets, 31,392 (27%) of which were fine-mapped to 3,000 distinct putative causal variants (posterior inclusion probability (PIP) > 0.8). These included 271 putative missense variants predicted by Ensembl VEP and 172 putative splice-altering variants predicted by either SpliceAI[24] or AlphaGenome[25] (Supplementary Table 4). Notably, 28 missense variants were also predicted to affect splicing. As an example, a fine-mapped missense variant (19-48806519-G-C, PIP = 1) in the *BCAT2* gene was associated with BCAA levels and was also predicted to disrupt splicing by SpliceAI[24] (score = 0.41) and AlphaGenome[25] (score = 0.143)[26].

In addition to the EUR genetic ancestry group, we also performed GWAS in five smaller genetic ancestry groups from the UKBB (AFR, AMR, CSA, EAS and MID) (Table 1). Including these summary statistics in our meta-analysis (meta_ALL) increased the number of independent lead variants from 8,260 to 8,398 (Table 1), 43 of which were not tested in the EstBB and UKBB_EUR cohorts owing to low allele frequency (allele count <20). Focusing separately on each genetic ancestry group, we found that between 9 and 48% of the lead variants had MAF < 0.1% in the UKBB_EUR analysis, with the highest proportion observed in the UKBB_AFR subset (68 out of 143 lead variants) (Extended Data Table 2). This highlights the need to substantially increase the sample sizes for under-represented genetic ancestry groups to enable the discovery of ancestry-specific associations[27].

## Colocalization across molecular layers

To demonstrate how genetic associations with metabolic traits can help to interpret disease associations, we used gpu-coloc[26] to colocalize all 86,886 signals from our meta_EUR analysis with GWAS summary statistics for up to 7,228 traits across three biobanks (FinnGen r12[28], Pan-UKBB[19], Million Veterans Program (MVP)[29] and FinnGen+MVP + UKBB meta-analysis (https://public-mvp-ukbb.finngen.fi/)) as well as gene expression, splicing and protein QTLs from the expression quantitative trait loci (eQTL) Catalogue[30], the INTERVAL eQTL study[31] and the UKBB Pharma Proteomics Project (UKB-PPP)[32]. We also included GWAS meta-analysis summary statistics for coronary artery disease (CAD; Aragam (2022))[5] and for type 2 diabetes (T2D; Suzuki (2024))[1] (Fig. 1b). Using a stringent colocalization posterior probability (CLPP) threshold ((PP.H4) > 0.9), we detected a total of 932,864 colocalization events which involved all 249 studied metabolic traits at least once. We detected at least one colocalization for 53.4% of the metabolic trait signals. Similarly, 20.2% of the CAD loci and 16.3% of the T2D loci colocalized with at least one metabolic trait. For biobanks, the percentage of colocalizing loci ranged from 4.9% (Pan-UKBB) to 11% (MVP). The lowest rate of colocalization was observed for eQTLs and protein quantitative trait loci (pQTLs), where 1.5–4.2% of the loci colocalized with at least one metabolic trait. Finally, approximately 2.5% of these colocalizations involved low-frequency (MAF < 1%) metabolic trait loci.

To better understand the patterns of shared colocalizations across complex traits, metabolic traits and molecular QTLs, we converted the 932,864 colocalization events into a colocalization graph, where independent association signals were defined as nodes and colocalization events (PP.H4 > 0.9) between these nodes were defined as edges. This graph contained 2,946 connected components, which we defined as colocalization clusters (see example in Extended Data Fig. 2). The sizes of the clusters ranged from 1 edge to 62,504 edges and the number of metabolic traits in each cluster ranged from 1 to 228. The two largest sources of colocalizing associations were the Pan-UKBB (involved in 1,643 clusters) and the eQTL Catalogue (involved in 1,576 clusters), which is likely to reflect the very large number of associations present in those datasets (Fig. 1b). As expected, we re-discovered several known colocalizations involving CAD, T2D and inflammatory conditions (Supplementary Note 1).

## Plasma lactate and platelet activation

To prioritize novel disease colocalizations while limiting the effect of horizontal pleiotropy that could complicate interpretation, we focused on the 291 colocalization clusters that colocalized with at least one disease end-point from the FinnGen + UKBB + MVP meta-analysis and involved 5 or fewer metabolic traits. This analysis revealed three high-confidence colocalizations between pulmonary embolism and plasma lactate at the *GP6* (rs1654425, 19-55027612-T-C, cluster 25; Extended Data Fig. 2), *GRK5* (rs10886430, 10-119250744-A-G, cluster 104) and *ZFPM2* (rs6993770, 8-105569300-A-T, cluster 696) loci (Fig. 1c and Supplementary Table 5). At all three loci, increased plasma lactate was associated with an increased risk of pulmonary embolism (FinnGen + UKBB + MVP meta-analysis) and an increased risk of deep vein thrombosis (MVP). For the *GP6* and *GRK5* loci, we also detected colocalizations with the corresponding *cis*-pQTLs in the UKB-PPP plasma proteomics dataset, and for *GRK5* we also detected a colocalizing eQTL signal in platelets[33] as well as in whole blood[31]. Although we did not detect a colocalizing *cis*-QTL effect at the *ZFPM2* locus, our lead variant (rs6993770) has previously been localized to a megakaryocyte-specific enhancer for other platelet traits[34]. Finally, although all three loci also

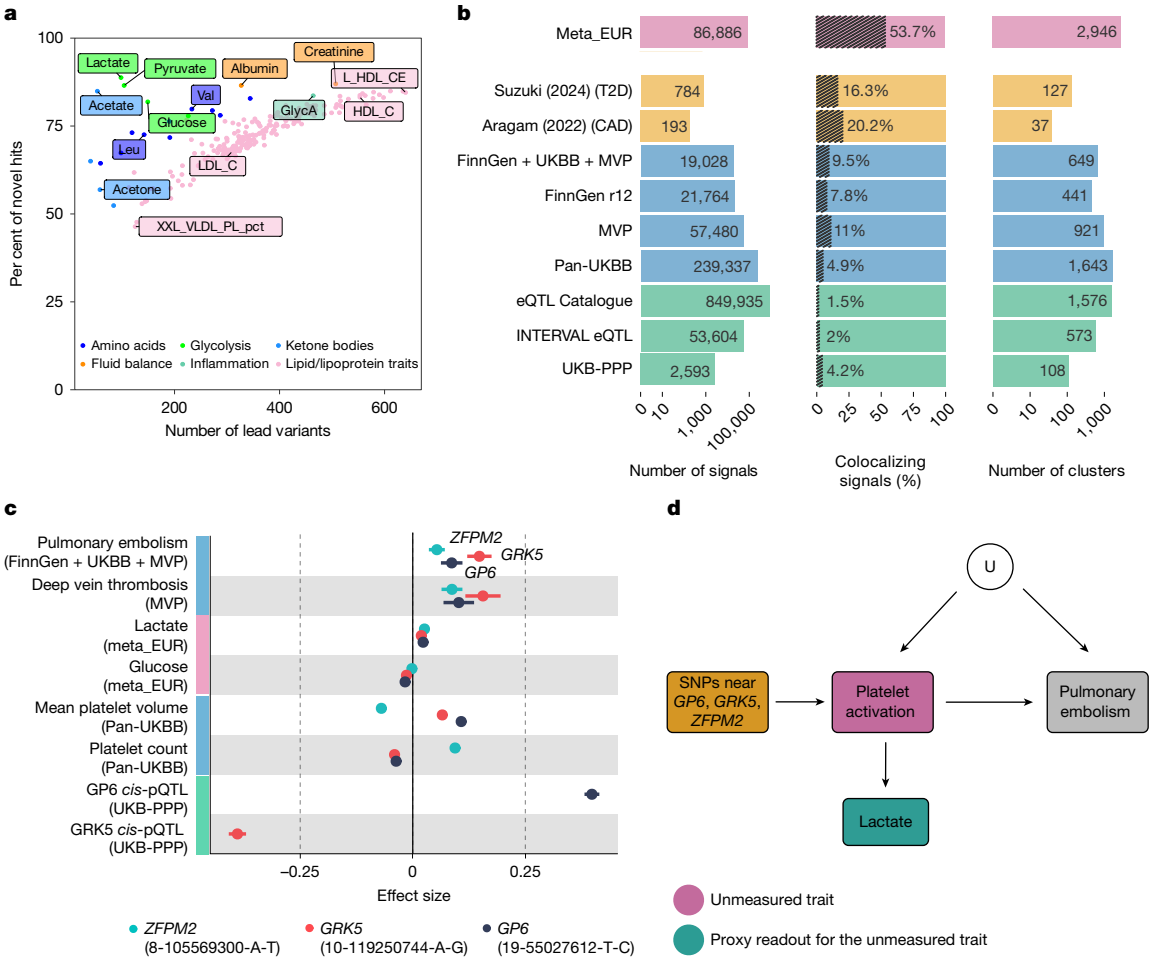



**Fig. 1 | Known and novel genetic associations with metabolic traits.** **a**, Number of genome-wide significant loci ($P < 5 \times 10^{-8}$) detected for each metabolic trait ($n = 249$) and the proportion of those associations that were not detected by Karjalainen et al.[8]. **b**, Overview of the datasets included in the colocalization analysis. The first column shows the number of signals included for colocalization from each dataset. The second column shows the proportion of these signals that colocalize with at least one metabolic trait. The third column indicates the number of colocalization clusters that involve at least one colocalization from each dataset. **c**, Forest plot of colocalizing genetic associations at the *ZFPM2*, *GRK5* and *GP6* loci between pulmonary embolism, deep vein thrombosis, lactate, glucose, mean platelet volume, platelet count and *cis*-pQTL signals for GP6 and GRK5 proteins. The points show the standardized GWAS effect size (beta) and error bars show the 95% confidence intervals. **d**, Proposed causal model linking genetic variants at the *GP6*, *GRK5* and *ZFPM2* loci via platelet activation to increased pulmonary embolism risk. Lactate is likely to act as a proxy readout of platelet activation without having a direct causal effect on pulmonary embolism risk. U, unmeasured confounders.

colocalized with both platelet count and mean platelet volume, the effect size directions were not consistent. At the *GP6* and *GRK5* loci, increased mean platelet volume and decreased platelet count were associated with increased risk for pulmonary embolism, whereas at the *ZFPM2* locus, both effects were the opposite (Fig. 1c). To further characterize these three loci, we focused on fine-mapped credible sets from the *GP6*, *GRK5* and *ZFPM2* loci and performed colocalization against all harmonized GWAS credible sets from the Open Targets Platform[35] using the CLPP method (Methods). We found that the *GP6* credible set (6 variants, maximum PIP = 0.48) colocalized (CLPP = 0.15) with a GWAS hit for platelet reactivity to collagen-related peptide[36], first reported in ref. 37. Similarly, the fine-mapped variant at the *GRK5* locus (10-119250744-A-G, PIP = 0.99) colocalized (CLPP = 0.99) with a GWAS signal for thrombin-induced platelet activation in two independent studies[36,38]. This is consistent with experimental data in mice indicating that *GRK5* is a negative regulator of platelet activation[39]. The *ZFPM2* fine-mapped enhancer variant (8-105569300-A-T, PIP = 0.52) had pleiotropic effects on several blood cell type composition and cytokine traits[34]. Notably, at all three loci, the alleles associated with increased lactate and increased pulmonary embolism risk were also associated with a decrease in plasma glucose (Fig. 1c). As activated platelets generate energy by converting glucose to lactate[40,41], this suggests that at these three loci, plasma lactate might serve as a biomarker for platelet activation[11] (Fig. 1d). This mirrors observational studies linking higher plasma lactate levels to increased mortality in patients with pulmonary embolism[42–44]. However, our analysis does not imply that plasma lactate itself has a direct causal effect on pulmonary embolism risk. In fact, when we focused on all genome-wide variants associated with plasma lactate (including those not colocalizing with pulmonary embolism), the association was inconclusive at best (Extended Data Fig. 3).

## Low-frequency and rare variants

Whereas previous NMR GWAS studies have primarily focused on common variation (MAF > 1%)[8,12,13], we tested all variants with minor allele count greater than 20. To further explore these low-frequency associations, we first focused on our fine-mapping results. We identified 10,016 (8.6%) credible sets where the MAF of the variant with the highest PIP was between 0.1% and 1%. These credible sets belonged to 786 probably independent signal clusters and contained 583 distinct confidently fine-mapped variants (PIP > 0.8). Notably, 135 out of 583 (23.1%) of the

low-frequency fine-mapped variants were predicted to be missense or splice variants (Supplementary Table 4). By contrast, only 11.5% of the common (MAF > 1%) fine-mapped variants were predicted to alter coding sequence or splicing, highlighting the increased interpretability of many low-frequency associations. On the basis of this analysis, we prioritized several low-frequency missense variants in known metabolic genes, such as variants in *PCSK9*, *ABCA1* and *ANGPTL4* associated with lipid traits, a missense variant (12-102840493-G-A) in *PAH* associated with phenylalanine and tyrosine, a missense variant (X-24503382-A-G) in *PDK3* associated with pyruvate, and four independent missense variants in the *HAL* gene associated with histidine (Supplementary Table 4).

However, to ensure accurate LD information, our fine mapping was restricted to the UKBB_EUR subset of samples (69% of all samples) and only included variants with MAF > 0.1%. To find associations that might have been missed by fine mapping, we alternatively focused on all lead variants from the meta_EUR meta-analysis with MAF < 1% ($n$ = 480 variants, 5.9% of all leads) (Supplementary Fig. 3). These corresponded to 324 low-frequency variants (MAF between 0.1% and 1%) and 156 rare variants with MAF < 0.1%. Reassuringly, 96 out of 324 low-frequency variants (30%) belonged to at least one fine-mapped credible set, and 68 out of 324 variants (21%) were predicted to be either missense or splice variants (Supplementary Table 6), suggesting that many of these low-frequency lead variants represent causal variants. The fraction of predicted missense and splice variants was slightly lower among the rare variants (23 out of 156 (15%)) (Supplementary Table 6), perhaps reflecting a slightly increased rate of false positives in this allele frequency bin or complex haplotype effects[45].

## Convergence of common and rare variants

As an example of insights gained from rare missense and splice variants, we observed convergence of common and rare variants on the BCAA catabolism pathway. The first two steps of BCAA catabolism are transamination of valine, leucine and isoleucine catalysed by branched-chain aminotransferase (encoded by *BCAT1* and *BCAT2* genes) followed by oxidative decarboxylation catalysed by the branched-chain α-keto acid dehydrogenase (BCKDH) complex[46] (Fig. 2a). The BCKDH complex is made up of three proteins: the E1 subunit encoded by *BCKDHA* and *BCKDHB*, the E2 subunit encoded by *DBT*, and the E3 subunit encoded by *DLD*[46] (Fig. 2b). The activity of the BCKDH complex is further regulated by BCKDK, which inhibits its activity, and protein phosphatase 2Cm (encoded by *PPM1K*), which reactivates it (Fig. 2b).

Common variant associations for three of the six genes (*BCAT2*, *DBT* and *PPM1K*) have been reported in previous GWAS studies for BCAAs. We additionally detected a rare (MAF = 0.012%, $P = 2.6 \times 10^{-13}$) missense variant, 19-41414070-A-G (rs771686663), in *BCKDHA* (Fig. 2c) and a rare (MAF = 0.047%, $P = 6.5 \times 10^{-29}$) splice region variant, 16-31111297-T-A (rs118042732), in *BCKDK* (Fig. 2d). In both cases, the predicted variant effects were directionally consistent with the sign of the GWAS associations (Fig. 2e). The *BCKDHA* missense variant was predicted by AlphaMissense[47] to be deleterious and was associated with increased BCAA levels. By contrast, the rs118042732 splice region variant in *BCKDK* was predicted by both SpliceAI (score = 0.30) and AlphaGenome (score = 0.34) to lead to splice acceptor loss and was associated with decreased BCAA levels (consistent with BCKDK being a negative regulator of the BCKDH complex) (Fig. 2b). Reassuringly, both *BCKDK* ($P < 1 \times 10^{-30}$) and *BCKDHA* ($P < 1 \times 10^{-13}$) were also found to be associated with BCAAs in a parallel effort that performed rare variant burden testing and exome-wide association testing using overlapping UKBB NMR samples[13], confirming that our rare variant imputation is reliable.

Finally, we detected a novel common variant (7-107837919-T-A, MAF 52%) association at the *DLD* locus (beta = −0.01, $P = 9.8 \times 10^{-13}$). Thus, we identified GWAS hits for all six key enzymes involved in the catabolism of BCAAs. This illustrates that very large sample sizes are needed to saturate the discovery of key regulators of biological processes owing to either very small effects of some common variants on the target genes (*DLD*) or very low allele frequency of the genetic variants that affect those genes (*BCKDHA* and *BCKDK*). Of note, whereas the GWAS and burden testing analysis by Zoodsma et al.[13] identified a largely divergent set of genes in this pathway (*BCAT2*, *BCKDK* and *BCKDHA* from burden testing and *BCAT2*, *DLD*, *DBT* and *PPM1K* from GWAS), we detected all six genes from a single analysis. This is consistent with recent reports that the differences between GWAS and burden testing results can be largely explained by differential statistical power[48].

## Extent of horizontal pleiotropy across metabolic traits

To understand the shared genetic control of various classes of metabolic traits, we explored genetic correlations between all 249 metabolic traits. Although the median genetic correlation across all traits was low (rg = 0.16), there were high genetic correlations between various lipoprotein traits (median rg = 0.52) as well as between other closely regulated metabolites, such as BCAAs (rg = 0.97) (Extended Data Fig. 4 and Supplementary Table 7). To characterize the molecular mechanisms underlying genetic correlations, we focused on the lead variants that were shared ($r^2 > 0.8$) between metabolic traits. Among the 249 metabolic traits, most lead variants were significantly associated ($P < 5 \times 10^{-8}$) with multiple metabolites (mean = 10, median = 2). Most prominently, a common missense variant (MAF = 40%) in *GCKR* (2-27508073-T-C, *GCKR*:p.Leu446Pro) was significantly associated ($P < 5 \times 10^{-8}$) with 229 (out of 249) metabolites (Fig. 3a). However, in many other cases, pleiotropy was restricted to the same class of metabolites, such as the 5-75360714-T-C (rs12916) variant at the *HMGCR* locus, which is associated with multiple lipid traits (Supplementary Fig. 4).

To characterize the effect of pleiotropic genetic effects on interpreting disease associations, we performed genome-wide Mendelian randomization using all 249 metabolic traits as exposures and either CAD[5] or T2D[1] as outcomes (see Methods). For CAD, 211 of the 249 tested metabolic traits (85%) yielded significant Mendelian randomization estimates (false discovery rate (FDR) < 5%; Fig. 3b), whereas for T2D (Fig. 3c), the number of significant associations was 157 (63% of tested traits) (Supplementary Table 8). Reassuringly, we recapitulated known causal effects between genetically regulated low-density lipoprotein (LDL) cholesterol and CAD (beta = 0.43, $P$ value = $6.02 \times 10^{-42}$) and between glucose and T2D (beta = 0.67, $P$ value = $6.3 \times 10^{-12}$). We also detected a known negative association between genetically lower LDL cholesterol and T2D[49] (beta = −0.11, $P$ value = $2.37 \times 10^{-4}$). Finally, we detected genome-wide significant Mendelian randomization estimates between BCAA levels and both CAD and T2D (Fig. 3b,c). However, these genome-wide Mendelian randomization estimates, beyond the known causal effects of LDL and glucose, can be tricky to interpret owing to extensive genetic correlation between the metabolic traits (Extended Data Fig. 4 and Supplementary Table 7), widespread horizontal pleiotropy, and large heterogeneity between the effect estimates from individual variants (Supplementary Table 8).

## Evaluating drug targets

To limit the effect of horizontal pleiotropy, we interrogated a subset of the genome-wide Mendelian randomization associations using a more conservative *cis*-Mendelian randomization approach[50,51]. Instead of capturing average genome-wide effects of circulating metabolic traits, *cis*-Mendelian randomization uses genetic variation in the *cis* region of the target gene to estimate the effect of perturbing gene function on disease risk[51]. If these genes have a direct biological effect on metabolic traits, then we can use the variant effect on those traits as a proxy readout for the (unmeasured) effect of these variants on gene function[11]. First, we focused on three genes with direct effect on regulating plasma LDL cholesterol levels: *LDLR*, *HMGCR* and *PCSK9* (Fig. 4a). In all three cases, we observed robust causal effects of perturbing these

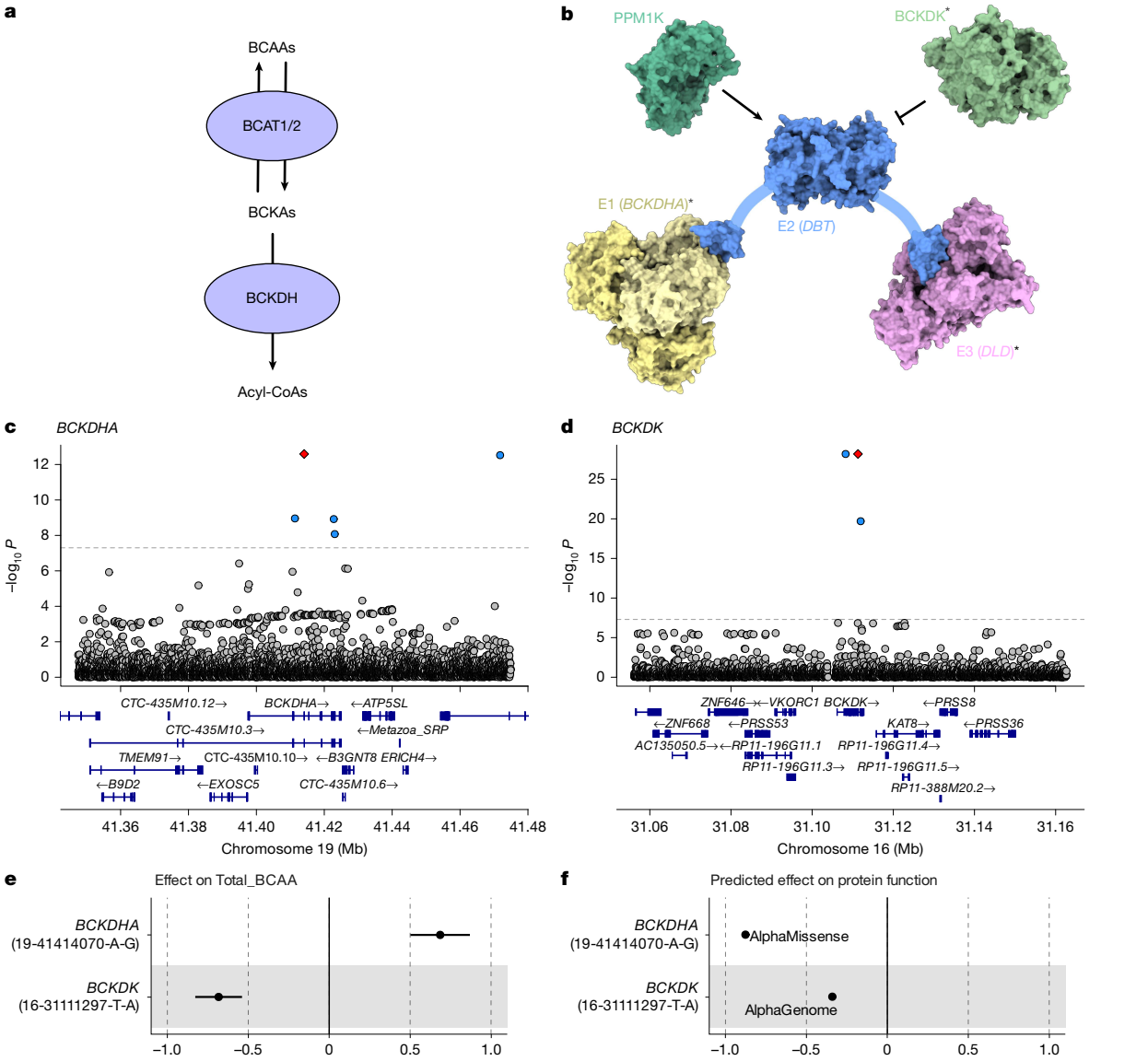



**Fig. 2 | Convergence of common and low-frequency associations at the BCAA catabolism pathway. a**, BCAAs are converted to branched-chain keto acids (BCKAs) by BCAA aminotransferase (encoded by *BCAT1* and *BCAT2*). This process is reversible. BCKAs can be further catabolised by the branched-chain α-keto acid dehydrogenase (BCKDH) complex to acyl-CoAs. **b**, The BCKDH complex is made up of subunits E1 (encoded by *BCKDHA* and *BCKDHB*), E2 (encoded by *DBT*) and E3 (encoded by *DLD*). The activity of the BCKDH complex is controlled by a kinase (BCKDK) that inhibits its function and a phosphatase (PPM1K) that reactivates it[46]. Asterisks indicate newly identified associations. **c**, GWAS association signal for total BCAA level (Total_BCAA) in the *BCKDHA* gene region. The *BCKDHA* missense variant 19-41414070-A-G is highlighted in red. **d**, GWAS association signal for Total_BCAA in the *BCKDK* gene region. The predicted *BCKDK* splice loss variant 16-31111297-T-A is highlighted in red. **e**, The effect of the *BCKDHA* missense variant and the *BCKDK* splice loss variant on Total_BCAA (beta ± 95% confidence interval). **f**, Predicted variant effect on protein function from AlphaMissense and AlphaGenome models (arbitrary units).

genes on CAD risk, as previously reported[9,49]. We then estimated the causal effect of lowering LDL cholesterol through these mechanisms on T2D. At the *HMGCR* locus, we detected a negative association between genetically regulated LDL cholesterol and T2D, which is consistent with previous Mendelian randomization studies as well as large clinical trials demonstrating that statin use is associated with increased T2D risk[49]. Notably, although the effect of genetically regulated LDL cholesterol on CAD risk was even higher at the *LDLR* and *PCSK9* loci, the effect on T2D was strongly attenuated relative to *HMGCR*[49]. This is consistent with clinical trials of *PCSK9* inhibitors that did not detect increased risk of T2D as a side effect[52]. However, our estimates could still be biased by LD between linked causal variants (Supplementary Note 2).

Reassured by the ability of *cis*-Mendelian randomization to rediscover known associations with lipid-lowering drug targets, we next followed up the significant genome-wide Mendelian randomization estimates between BCAAs and both CAD and T2D (Fig. 3b,c). Although the association between genetically regulated BCAAs and T2D has been reported before[6], genome-wide Mendelian randomization necessarily averages effects across multiple distinct mechanisms, only some of which might influence T2D risk. Furthermore, recent studies have suggested that the genome-wide Mendelian randomization signal between BCAAs and T2D first observed by Lotta et al.[6] might be confounded by horizontal pleiotropy and reverse causality[13,53,54]. As discussed above, a prominent mechanism that regulates plasma BCAA levels is BCAA catabolism, which is controlled by BCAT2 and the BCKDH complex (Fig. 2a). To clarify the contradictory results obtained from genome-wide Mendelian randomization analysis and motivated by the recent discovery of a clinical candidate BCKDK kinase inhibitor[55],

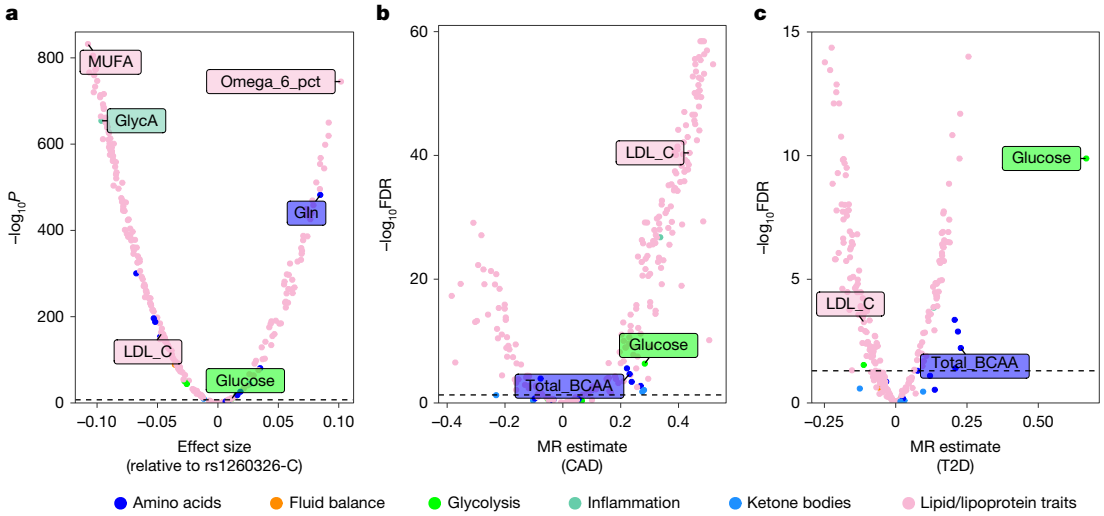



**Fig. 3 | Extent of pleiotropic associations across metabolic traits.** **a**, Pleiotropic effects of the *GCKR* missense variant 2-27508073-T-C (rs1260326) on 249 metabolites. The dotted line represents the genome-wide significance threshold ($5 \times 10^{-8}$). **b**, Genome-wide Mendelian randomization estimates using all 249 metabolic traits as exposures and CAD as outcome. The *y* axis shows negative $\log_{10}$-transformed FDR-adjusted *P* values, with a horizontal line drawn at $P = 0.05$. **c**, Genome-wide Mendelian randomization estimates using all 249 metabolic traits as exposures and T2D as outcome. The *y* axis shows negative $\log_{10}$-transformed FDR-adjusted *P* values, with a horizontal line drawn at $P = 0.05$.

we sought to assess whether lowering plasma BCAA levels via specific inhibition of the BCKDK kinase could reduce T2D and CAD risk.

The most direct way to assess this would be to perform *cis*-Mendelian randomization with genetic variants in the *cis* region of *BCKDK* as genetic instruments, plasma BCAA levels as a proxy exposure and T2D as outcome. However, the *BCKDK* region lacks strong common variant associations, and the splice donor variant that we detected (Fig. 2d) is too rare to have sufficient power for *cis*-Mendelian randomization. Instead, we focused on *cis* variation near *DBT* and *PPM1K*, two other members of the BCKDH complex (Fig. 2b) that have robust common variant associations. We also included *cis* variation near *BCAT2*, an enzyme that is directly upstream of the BCKDH complex in the BCAA catabolism pathway (Fig. 2a). In all three gene regions, the results were broadly consistent with a null effect of BCKDK inhibition on T2D risk (Fig. 4b). None of these loci had genome-wide significant hits for T2D (Supplementary Figs. 5 and 6). Although some Mendelian randomization method and outcome GWAS combinations (Methods) did yield non-null causal effect estimates, these were not consistent across the three *cis* regions (Supplementary Tables 9–11). Thus, current genetic evidence does not support the idea that *BCKDK* inhibition would have a large beneficial effect on reducing T2D risk.

## Limitations

Our study has several limitations. First, 97% of the samples included in the analysis were of predominantly European genetic ancestries. This skew limited our ability to detect genome-wide significant signals in other genetic ancestry groups and may influence the generalizability of our findings across genetic ancestry groups. As a result, the number of genome-wide significant signals increased by only 1.4% (Table 1) when samples from other UKBB genetic ancestry groups (AFR, AMR, CSA, EAS and MID) were included in the meta-analysis. Secondly, owing to substantial methodological challenges in meta-analysis fine mapping[23], we performed statistical fine mapping only on the UKBB_EUR subset of the samples. Thus, we probably missed many weaker secondary signals at the genome-wide significant loci. Finally, our analysis was necessarily limited to the 249 (mostly lipoprotein) metabolic traits detectable by the Nightingale Health NMR platform. Previous mass spectrometry-based metabolic trait GWAS studies have profiled a much broader spectrum of metabolites (up to 1,919 traits[56]), but at the cost of significantly lower throughput (at most 19,994 individuals[57]).

Although biobank-scale datasets provide unprecedented power for genetic discovery, they also introduce complexities in interpreting genetic associations due to pervasive pleiotropy. Our results reinforce previous reports of extensive pleiotropy across metabolic trait GWASs[8–10]. Some of this pleiotropy is readily interpretable, such as co-regulation between various lipid traits[9] or opposing effects between substrates and products of enzymatic reactions[10]. However, given our large sample size, we also detected more cryptic pleiotropic effects, such as the *GCKR* missense variant that was associated with 231 out of 249 tested metabolic traits (Fig. 3a). Such extensive overlaps exemplify that as cohorts grow larger, the detection of pleiotropic signals becomes more pronounced, making it more difficult to disentangle direct and indirect genetic effects. As a result, we caution against interpreting genome-wide Mendelian randomization results as evidence of direct causal effects of tested metabolic traits on the outcomes of interest[58]. *Cis*-Mendelian randomization analyses are potentially less susceptible to these pleiotropic effects, but still require careful consideration of LD and detailed understanding of metabolic pathways to ensure that Mendelian randomization assumptions are met[11,50,51].

## Conclusion

We have created a comprehensive resource of both common and low-frequency genetic variants associated with 249 metabolic traits in up to 619,372 individuals across multiple ancestry groups. We have demonstrated the utility of the resource for GWAS interpretation via statistical fine mapping, systematic phenome-wide colocalization, low-frequency variant prioritization and *cis*-Mendelian randomization analyses. To ensure that our results can be used as widely as possible, we have publicly released all summary statistics via the GWAS Catalog[59] and we have also made the association and colocalization results easy to query via two online browsers: https://nmrmeta.gi.ut.ee/ and https://elixir.ut.ee/eqtl/nmr-coloc. Although we were still underpowered to detect many novel associations with rare variants (156 unique lead variants with MAF < 0.1%), our analysis clearly highlights the value of including low-frequency (MAF between 0.1% and 1%) variants in GWAS discovery and interpretation. Whereas 8.6% of the credible sets had low-frequency lead variants, this proportion increased to 19.4% for

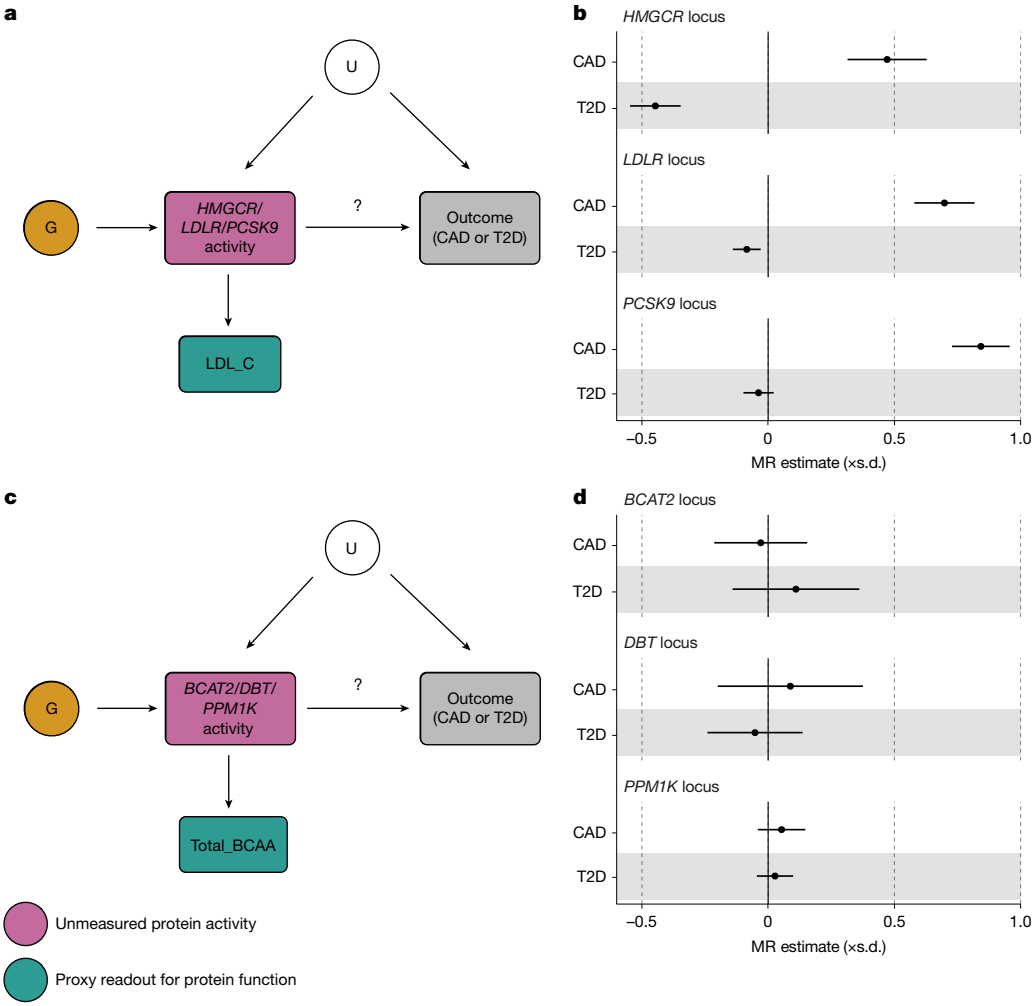



**Fig. 4 | Drug target evaluation with *cis*-Mendelian randomization.** **a**, *Cis*-Mendelian randomization using genetic variants G from the *cis* regions of *HMGCR*, *LDLR* and *PCSK9* genes is used to estimate the causal effect of inhibiting the corresponding gene function on risk for T2D[1] and CAD[5]. LDL cholesterol (LDL_C) is used as a proxy readout for variant effects on *HMGCR*, *LDLR* and *PCSK9* function. **b**, Mendelian randomization estimates from **a**, with 95% confidence intervals. **c**, *Cis*-Mendelian randomization using genetic variants G from the *cis* regions of *BCAT2*, *DBT* and *PPM1K* genes is used to estimate the causal effect of inhibiting the corresponding protein function on risk for T2D[1] and CAD[5]. Total_BCAA is used as a proxy readout for variant effects on *BCAT2*, *DBT* and *PPM1K* function. **d**, Mendelian randomization estimates from **c**, with 95% confidence intervals. U, unmeasured confounders.

confidently fine-mapped (PIP > 0.8) variants, suggesting that owing to less extensive LD, low-frequency signals are easier to fine map than common signals. Notably, these low-frequency fine-mapped variants were also twice as likely to be predicted as missense or splice-altering than common fine-mapped variants (23.1% versus 11.5%), thus providing a clear hypothesis about the causal gene and a likely mechanism of action. While traditional multi-cohort GWAS meta-analyses have been limited to common variants[1,5], low-frequency variants are now routinely included in large biobanks such as FinnGen[28], MVP[29] and Pan-UKBB[19]. As a result, 2.5% of our colocalizations involved metabolic trait loci with MAF < 1%. Furthermore, for many common complex diseases, cross-biobank meta-analyses (such as the FinnGen + MVP + UKBB meta-analysis (https://public-mvp-ukbb.finngen.fi/)) are now achieving comparable statistical power to traditional multi-cohort studies, thus making it possible to include low-frequency variants in colocalization and *cis*-Mendelian randomization analyses to reveal disease mechanisms and prioritize drug targets. However, taking full advantage of these low-frequency meta-analysis associations requires accurate fine mapping of conditionally distinct signals, which is currently an active area of research with several novel methods proposed[23,60,61].

Our results demonstrate that even for well-studied metabolic traits, increasing GWAS sample size can still yield novel discoveries and biological insights. However, as the discovery of associations increases, it is increasingly important to also invest in scalable tools and computational methods to support the interpretation and prioritization of these associations.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10532-5.


1. Suzuki, K. et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* **627**, 347–357 (2024).
2. Zhou, W. et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human disease. *Cell Genom.* **2**, 100192 (2022).
3. Yengo, L. et al. A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
4. COVID-19 Host Genetics Initiative A second update on mapping the human genetic architecture of COVID-19. *Nature* **621**, E7–E26 (2023).
5. Aragam, K. G. et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat. Genet.* **54**, 1803–1815 (2022).

6. Lotta, L. A. et al. Genetic predisposition to an impaired metabolism of the branched-chain amino acids and risk of type 2 diabetes: a Mendelian randomisation analysis. *PLoS Med.* **13**, e1002179 (2016).
7. Vanweert, F., Schrauwen, P. & Phielix, E. Role of branched-chain amino acid metabolism in the pathogenesis of obesity and type 2 diabetes-related metabolic disturbances BCAA metabolism in type 2 diabetes. *Nutr. Diabetes* **12**, 35 (2022).
8. Karjalainen, M. K. et al. Genome-wide characterization of circulating metabolic biomarkers. *Nature* **628**, 130–138 (2024).
9. Richardson, T. G. et al. Characterising metabolomic signatures of lipid-modifying therapies through drug target mendelian randomisation. *PLoS Biol.* **20**, e3001547 (2022).
10. Smith, C. J. et al. Integrative analysis of metabolite GWAS illuminates the molecular basis of pleiotropy and genetic correlation. *eLife* **11**, e79348 (2022).
11. Rahu, I., Tambets, R., Fauman, E. B. & Alasoo, K. Mendelian randomization with proxy exposures: challenges and opportunities. *Genetics* **231**, iyaf210 (2025).
12. van der Meer, D. et al. Pleiotropic and sex-specific genetic mechanisms of circulating metabolic markers. *Nat. Commun.* **16**, 4961 (2025).
13. Zoodsma, M. et al. A genetic map of human metabolism across the allele frequency spectrum. *Nat. Genet.* **57**, 2445–2455 (2025).
14. Graham, S. E. et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
15. Nag, A. et al. Effects of protein-coding variants on blood metabolite measurements and clinical biomarkers in the UK Biobank. *Am. J. Hum. Genet.* **110**, 487–498 (2023).
16. Sanderson, E. et al. Mendelian randomization. *Nat. Rev. Methods Primers* **2**, 1–21 (2022).
17. Stender, S., Gellert-Kristensen, H. & Smith, G. D. Reclaiming mendelian randomization from the deluge of papers and misleading findings. *Lipids Health Dis.* **23**, 286 (2024).
18. Burgess, S., Woolf, B., Mason, A. M., Ala-Korpela, M. & Gill, D. Addressing the credibility crisis in Mendelian randomization. *BMC Med.* **22**, 374 (2024).
19. Karczewski, K. J. et al. Pan-UK Biobank genome-wide association analyses enhance discovery and resolution of ancestry-enriched effects. *Nat. Genet.* **57**, 2408–2417 (2025).
20. Mitt, M. et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* **25**, 869–876 (2017).
21. Shi, S. et al. A Genomics England haplotype reference panel and imputation of UK Biobank. *Nat. Genet.* **56**, 1800–1803 (2024).
22. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
23. Kanai, M. et al. Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. *Cell Genom.* **2**, 100210 (2022).
24. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
25. Avsec, Ž et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**, 1206–1218 (2026).
26. Jesse, M., Riet, A.-E. & Alasoo, K. Ultra-fast genetic colocalisation across millions of traits. Preprint at *bioRxiv* https://doi.org/10.1101/2025.08.25.672103 (2025).
27. Takeuchi, Y. et al. Genetic architecture of circulating metabolic biomarkers across ancestral populations. Preprint at *medRxiv* https://doi.org/10.64898/2025.12.03.25341540 (2025).
28. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
29. Verma, A. et al. Diversity and scale: genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024).
30. Kerimov, N. et al. eQTL Catalogue 2023: new datasets, X chromosome QTLs, and improved detection and visualisation of transcript-level QTLs. *PLoS Genet.* **19**, e1010932 (2023).
31. Tokolyi, A. et al. The contribution of genetic determinants of blood gene expression and splicing to molecular phenotypes and health outcomes. *Nat. Genet.* **57**, 616–625 (2025).
32. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
33. Momozawa, Y. et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, 2427 (2018).
34. Akbari, P. et al. A genome-wide association study of blood cell morphology identifies cellular proteins implicated in disease aetiology. *Nat. Commun.* **14**, 5023 (2023).
35. Buniello, A. et al. Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Res.* **53**, D1467–D1475 (2025).
36. Downes, K. et al. G protein-coupled receptor kinase 5 regulates thrombin signaling in platelets via PAR-1. *Blood Adv.* **6**, 2319–2330 (2022).
37. Jones, C. I. et al. Mapping the platelet profile for functional genomic studies and demonstration of the effect size of the GP6 locus. *J. Thromb. Haemost.* **5**, 1756–1765 (2007).
38. Rodriguez, B. A. T. et al. A platelet function modulator of thrombin activation is causally linked to cardiovascular disease and affects PAR4 receptor signaling. *Am. J. Hum. Genet.* **107**, 211–221 (2020).
39. Li, C. et al. G protein-coupled receptor kinase 5 regulates thrombin signaling in platelets. *Res. Pract. Thromb. Haemost.* **8**, 102556 (2024).
40. Heijnen, H. F., Oorschot, V., Sixma, J. J., Slot, J. W. & James, D. E. Thrombin stimulates glucose transport in human platelets via the translocation of the glucose transporter GLUT-3 from alpha-granules to the cell surface. *J. Cell Biol.* **138**, 323–330 (1997).
41. Detwiler, T. C. & Zivkovic, R. V. Control of energy metabolism in platelets. A comparison of aerobic and anaerobic metabolism in washed rat platelets. *Biochim. Biophys. Acta* **197**, 117–126 (1970).
42. Vanni, S. et al. Prognostic value of plasma lactate levels among patients with acute pulmonary embolism: the thrombo-embolism lactate outcome study. *Ann. Emerg. Med.* **61**, 330–338 (2013).
43. Vanni, S. et al. High plasma lactate levels are associated with increased risk of in-hospital mortality in patients with pulmonary embolism: high plasma lactate levels and PE. *Acad. Emerg. Med.* **18**, 830–835 (2011).
44. Leidi, A. et al. Risk stratification in patients with acute pulmonary embolism: current evidence and perspectives. *J. Clin. Med.* **11**, 2533 (2022).
45. Hawkes, G. et al. Whole-genome sequencing analysis identifies rare, large-effect noncoding variants and regulatory regions associated with circulating protein levels. *Nat. Genet.* **57**, 626–634 (2025).
46. Mann, G., Mora, S., Madu, G. & Adegoke, O. A. J. Branched-chain amino acids: catabolism in skeletal muscle and implications for muscle and whole-body metabolism. *Front. Physiol.* **12**, 702826 (2021).
47. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
48. Spence, J. P. et al. Specificity, length and luck drive gene rankings in association studies. *Nature* **649**, 918–925 (2026).
49. Lotta, L. A. et al. Association between low-density lipoprotein cholesterol-lowering genetic variants and risk of type 2 diabetes: a meta-analysis. *JAMA* **316**, 1383–1391 (2016).
50. Burgess, S. et al. Guidelines for performing Mendelian randomization investigations: update for summer 2023. *Wellcome Open Res.* **4**, 186 (2019).
51. Gill, D. et al. Common pitfalls in drug target Mendelian randomization and how to avoid them. *BMC Med.* **22**, 473 (2024).
52. Carugo, S., Sirtori, C. R., Corsini, A., Tokgozoglu, L. & Ruscica, M. PCSK9 inhibition and risk of diabetes: should we worry?. *Curr. Atheroscler. Rep.* **24**, 995–1004 (2022).
53. Mahendran, Y. et al. Genetic evidence of a causal effect of insulin resistance on branched-chain amino acid levels. *Diabetologia* **60**, 873–878 (2017).
54. Wang, Q., Holmes, M. V., Davey Smith, G. & Ala-Korpela, M. Genetic support for a causal role of insulin resistance on circulating branched-chain amino acids and inflammation. *Diabetes Care* **40**, 1779–1786 (2017).
55. Filipski, K. J. et al. Discovery of first branched-chain ketoacid dehydrogenase kinase (BDK) inhibitor clinical candidate PF-07328948. *J. Med. Chem.* **68**, 2466–2482 (2025).
56. Schlosser, P. et al. Genetic studies of paired metabolomes reveal enzymatic and transport processes at the interface of plasma and urine. *Nat. Genet.* **55**, 995–1008 (2023).
57. Surendran, P. et al. Rare and common genetic determinants of metabolic individuality and their effects on human health. *Nat. Med.* **1**, 12 (2022).
58. Reed, Z. E. et al. Exploring pleiotropy in Mendelian randomisation analyses: what are genetic variants associated with 'cigarette smoking initiation' really capturing? *Genet. Epidemiol.* **49**, e22583 (2025).
59. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
60. Kartau, J. & Pirinen, M. FINEMAP-miss: fine-mapping genome-wide association studies with missing genotype information. *Bioinformatics* **41**, btaf616 (2025).
61. Yang, Z. et al. CARMA is a new Bayesian model for fine-mapping in genome-wide association meta-analyses. *Nat. Genet.* **55**, 1057–1065 (2023).




**Estonian Biobank Research Team**

**Mari Nelis[2], Georgi Hudjasov[2], Mait Metspalu[2], Reedik Mägi[2], Andres Metspalu[2] & Lili Milani[2]**

# Methods

### Estonian Biobank

The EstBB is a volunteer-based biobank at the Institute of Genomics, University of Tartu[62]. The current EstBB data freeze consists of 212,955 adult (age ≥ 18 years) participants, reflecting the age, sex and geographical distribution of the adult Estonian population, for whom biological samples as well a variety of health-related and demographic information have been collected. All biobank participants have signed a broad informed consent form and their blood sample collection was undertaken across the country between 2002 and 2021[62,63]. The activities of EstBB are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of EstBB. The Nightingale Health NMR platform was used to generate plasma metabolic trait profiles for all individual samples in the biobank. The assay covers 249 metabolic traits ranging from low molecular weight compounds to lipids and lipoproteins. Individual-level data analysis in EstBB was carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data according to release application 6-7/GI/8988 from the EstBB.

### UK Biobank

The UKBB is a longitudinal biomedical study of approximately half a million participants between 38–71 years of age from the UK[64]. Participant recruitment was conducted on a volunteer basis and took place between 2006 and 2010. Initial data were collected in 22 different assessment centres throughout Scotland, England and Wales. Data collection includes elaborate genotype, environmental and lifestyle data. Blood samples were drawn at baseline for all participants, with an average of 4 h since the last meal (that is, generally non-fasting). NMR metabolic traits (Nightingale Health, quantification library 2020) were measured from EDTA plasma samples (aliquot 3) during 2019–2024 from the entire cohort. Details on the NMR metabolomic measurements in UKBB have been described previously for the first tranche of ~120,000 samples[65]. The UKBB study was approved by the North West Multi-Centre Research Ethics Committee. This research was conducted using the UKBB Resource under application numbers 91233 and 30418.

### NMR data QC and normalization

NMR data generation in the EstBB and UKBB has been previously described[66]. During the quality control of the NMR metabolomics data, we detected a difference between distributions of several metabolic traits (notably Ala and His) driven primarily by spectrometer and batch effect. We removed this unwanted technical variation using the R package ukbnmr in both EstBB and UKBB data[67]. We excluded individuals with more than 5 missing metabolic trait measurements from the cohort, confirmed that none of the 249 metabolic traits had a significant number of missing measurements (8,000 for EstBB, 24,000 for UKBB), and applied inverse normal transformation to each metabolic trait to obtain the final dataset.

### Association testing and meta-analysis

Genotype imputation for the EstBB and UKBB cohorts is described in Supplementary Note 3. We conducted genome-wide association tests for each of the seven genetic ancestry groups separately using regenie v3.1.1[68], with sex, age, age squared and the top principal components (PCs) of the genotype data used as covariates (PC1–PC10 for EstBB, PC1–PC20 for UKBB). For step 1 (whole-genome model), we used genotype calls for UKBB and genotyping data for EstBB and included variants with a MAF of at least 1%, a minor allele count of at least 20, Hardy-Weinberg equilibrium exact test $P$ values of $10^{-15}$ or less, and maximum per-variant and per-sample missing genotype rates of 0.1. For step 2 (association testing using a linear regression model), we used imputed genotypes and selected variants with a minor allele count of at least 20 and an imputation INFO score of at least 0.6.

We performed two different inverse-variance weighted fixed-effect meta-analyses: meta_EUR on individuals of predominantly European genetic ancestry (EstBB cohort and EUR genetic ancestry group of UKBB), and meta_ALL which encompasses all seven genetic ancestry groups from UKBB and EstBB.

### Genetic correlations

We utilized LD score regression (LDSC)[69,70] to obtain pairwise genetic correlations for all 249 NMR metabolic traits. Correlations were calculated between biobanks for each metabolic trait and between all metabolic traits in three of the largest datasets (EstBB, UKBB_EUR and meta_EUR) using the European reference panel LD scores from 1000 Genomes, as provided by the authors of LDSC (https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2). Genetic correlations were also calculated between traits that were present in both Pan-UKBB lab measurements and NMR measurements in the UKBB_EUR cohort, as well as between two inflammation markers, CRP and GlycA. All Pan-UKBB summary statistics were lifted over to the GRCh38 genome version prior to analysis.

### Lead variant and locus definition

For the common and low-frequency variants (MAF > 0.1%, ~15 million variants), we used the standard genome-wide significance threshold of $P < 5 \times 10^{-8}$. However, our analysis also included up to 80 million rare variants (MAF < 0.1%) for which the standard $P$ value threshold was too lenient. To be conservative, we treated all rare variant tests as independent and used the Bonferroni correction to establish a more stringent significance threshold of $P < 0.05/80{,}000{,}000$ ($6.25 \times 10^{-10}$). We obtained the set of dataset–metabolic trait–variant triplets by iterating over variants that met these thresholds. The variant with the lowest $P$ value was designated as the lead variant within a 2 Mb locus. In each dataset, neighbouring loci were merged into one if their lead variants were in LD with an $r^2$ of at least 0.05. This was done to prevent very strong association signals (for example, $-\log_{10}P > 100$) from 'bleeding' outside the 2 Mb window. To better evaluate the independence of lead variants, we utilized PLINK v1.90b6.26 to calculate pairwise LD between all lead variants in a single genetic ancestry group, assigning them into shared cross-metabolic trait clusters if $r^2$ was at least 0.8. The variant with the smallest $P$ value was assigned as the lead variant for each cluster.

### Colocalization

We used gpu-coloc to perform large-scale genetic colocalization between metabolic traits and multiple large and publicly accessible repositories and biobanks. The prior probabilities for coloc were set to $p_1 = p_2 = 1 \times 10^{-4}$, $p_{12} = 5 \times 10^{-6}$ as recommended in the gpu-coloc manuscript[26]. For the eQTL Catalogue r7[30,71], FinnGen r12[28] and INTERVAL eQTL[31] datasets, we used fine-mapped SuSiE logarithms of Bayes factors (LBFs) available from these resources directly as input to gpu-coloc. The eQTL Catalogue r7 eQTL and sQTL (quantified by leafCutter) LBFs were downloaded from the eQTL Catalogue FTP Server (https://www.ebi.ac.uk/eqtl/). The FinnGen r12 LBFs were downloaded from the FinnGen website (https://www.finngen.fi/en/access_results). The INTERVAL eQTL LBFs were downloaded from Zenodo (https://doi.org/10.5281/zenodo.17956387).

For the other resources, we started with marginal summary statistics and converted these to approximate Bayes factors for colocalization using the algorithm from the approx.bf.estimates function of the coloc[72] R package (https://github.com/chr1swallace/coloc/blob/main/R/claudia.R#L96). To define loci for colocalization from each GWAS, we started with the variant with the smallest $P$ value and defined the region in a ±1 Mb window around that variant as the first locus. We then excluded all variants from the first locus from consideration and proceeded recursively to define additional loci until there were no additional variants with $P < 5 \times 10^{-8}$ remaining. The Pan-UKBB[19]

association summary statistics for 7,228 traits were downloaded from the https://pan.ukbb.broadinstitute.org/ website. We converted variant coordinates from GRCh37 genome build to GRCh38 with pyliftover[73] and corrected the reference and alternative alleles with pyfaidx[74]. The marginal summary statistics for the MVP[29] dataset were downloaded from the public dbGaP FTP server (https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs002453/phs002453.v1.p1/analyses/GIA/). We corrected the reference and alternative alleles using pyfaidx[74] and for the qualitative traits, we re-calculated the $-\log_{10} P$ value from odds ratios and credible intervals, as for some trait–variant pairs the original $P$ value provided by the authors was rounded to 0. The UKB-PPP[32] summary statistics were downloaded from Synapse (https://www.synapse.org/Synapse:syn51364943). For each of the 2,923 proteins, we only used the summary statistics from the *cis* region (±1 Mb) around the corresponding protein coding gene. The summary statistics from the FinnGen + MVP + UKBB meta-analysis were downloaded from https://public-mvp-ukbb.finngen.fi/. The summary statistics from the Suzuki (2024) study[1] were downloaded from http://www.diagram-consortium.org/downloads.html. The summary statistics from the Aragam (2022) study[5] were downloaded from the GWAS Catalog (accession GCST90132315).

### Statistical fine mapping

We used SuSiE v0.14.2[75,76] to identify conditionally distinct association signals around each meta_EUR lead variant that had MAF > 0.1% in the UKBB_EUR subset and excluded variants in the MHC region (chr. 6:28510120–33480577). We utilized the susie_rss method from the susieR R package with prior weights set to null and scaled prior variance set to 0.1. We use LDstore[77] to calculate in-sample LD matrices for the UKBB_EUR subset of samples on the UKBB's dnanexus platform. To reduce computational complexity, we divided the genomic regions containing variants of interest into 3 Mb wide windows with a 1 Mb overlap and calculated LD for each. This ensured that each lead variant and the variants up to 500 kb from the lead variant were always contained in at least one LD matrix, except for variants located near chromosome ends or within the excluded MHC region. We imported the LD matrices into R using the rbcor package (https://github.com/mkanai/rbcor). Fine-mapped variants were considered as putative splice-altering variants if at least one of their SpliceAI[24] or AlphaGenome[25] donor or acceptor scores across both strands was greater than 0.1. Top 3% of the variants had a SpliceAI score > 0.1 and top 3.8% of the variants had AlphaGenome score > 0.1. To further characterize individual loci, we also performed colocalization between our meta_EUR fine-mapped credible sets and all fine-mapped credible sets available from the Open Targets Platform[35]. We downloaded the credible set files from the Open Targets FTP server (https://ftp.ebi.ac.uk/pub/databases/opentargets/platform/25.09/output/credible_set/).

We used the CLPP method to test if the association signals represented by two credible sets colocalize[78]. We set the colocalization threshold to CLPP > 0.04 as recommended previously[26].

### Prioritization of functional variants

We first identified all independent lead variants in the meta_EUR analysis that had MAF < 1%. We then narrowed the set down by only including SNPs identified as missense or splice regions variants by the Ensembl Variant Effect Predictor (VEP)[79] or were predicted to alter splicing by SpliceAI[24] or AlphaGenome[25] (at least one of their donor or acceptor scores across both strands was greater than 0.1). This approach identified 91 variants that we were able to assign to putative effector genes (Supplementary Table 6).

### Genome-wide Mendelian randomization

Genome-wide Mendelian randomization studies seek to determine metabolic traits (exposures) that have a causal effect on any number of outcomes (often complex diseases). However, inferences from Mendelian randomization studies are valid only if certain assumptions are met[50,80]. A key assumption of Mendelian randomization is that the genetic variants are associated with the outcome only via the exposure of interest[58]. In practice, this assumption can be challenging to satisfy, because genetic variants can have pleiotropic effects on multiple metabolic traits[8–10]. We performed genome-wide Mendelian randomization between all 249 metabolic traits and two diseases, CAD[5] and T2D[1], resulting in a total of 498 analyses. For each metabolic trait, we identified instrumental variables using a greedy LD pruning approach applied to its lead variants with MAF > 1%. This involved: (1) assigning the lead with the lowest $P$ value in the initial set to the instrument set; (2) discarding that variant and all variants in LD with it ($r^2 < 0.01$) from the initial set; and (3) repeating steps A and B until no variants remained in the initial set. For the Mendelian randomization analysis itself, we used multiplicative random-effects inverse-variance weighted Mendelian randomization (IVW-MR) (implemented in MendelianRandomization R package[81]) as recommended in recent guidelines[50].

### *Cis*-Mendelian randomization

A promising alternative to genome-wide Mendelian randomization is *cis*-Mendelian randomization that focuses the analysis to a specific *cis* region around the target gene of interest[9,51]. While *cis*-Mendelian randomization is less susceptible to horizontal pleiotropy[11,50], it is limited by the number of independent association signals that can be identified at any gene region, thus requiring very well powered association studies. For the primary *cis*-Mendelian randomization analysis, we included only variants from the ±200 kb region around the gene body of the target gene that had MAF > 1%. For instrument selection, we used the LD information from the UKBB Genomics England imputation (Pan-UKBB EUR subset, $n = 413{,}897$) and used greedy pruning strategy to only retain variants with $P < 5 \times 10^{-8}$ and $r^2 < 0.01$. Although previous studies have used more relaxed $r^2$ thresholds for LD pruning[9], we found that our high statistical power required a more stringent filtering to avoid including many variants with low residual LD. We performed the primary *cis*-Mendelian randomization analysis using two pleiotropy-robust methods that we found to perform well in the *cis*-Mendelian randomization context in our previous benchmark study[82]: multiplicative random-effects IVW-MR implemented in the MendelianRandomization R package[81] and MRLocus[83]. For the IVW-MR method, we also specified 'weights = delta'. We also repeated the same analysis using MR-Egger[84].

To further assess the robustness of our Mendelian randomization results, we also tested two other *cis*-Mendelian randomization methods: MR-link-2[85] and MR-PCA[86]. The advantage of these methods is that instead of requiring instrument selection via LD pruning, they explicitly model the LD between all associated variants in the *cis* region. For MR-link-2 and MR-PCA, all association summary statistics were harmonized to the UK10K genotype ref. 87 and both methods were using the default parameters provided by the mr_link_2_standalone.py function: MAF > 0.005, regional definition ±250 kb, and instrument selection threshold $P < 5 \times 10^{-8}$ and variance explained of the LD matrix of 99%.

### Structural modelling

Models of the three subunits of the BCKDH complex in Fig. 2b were generated using AlphaFold 3 via the AlphaFold Server[88]. Structures of the regulatory proteins BCKDK and PPM1K were retrieved from the AlphaFold Protein Structure Database[89,90]. For clarity in visualization, disordered or unstructured regions at the N and C termini were manually removed. Molecular graphics were performed with UCSF ChimeraX[91].

### Reporting summary

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

## Data availability

Complete genetic ancestry group-specific and meta-analysis association summary statistics from this study can be downloaded from the GWAS Catalog[59] (accessions GCST90449363–GCST90451603, Supplementary Table 12). GWAS lead variants, fine-mapping credible sets, and colocalization results are available from Zenodo (https://zenodo.org/records/13937265, https://zenodo.org/records/18132538 (ref. 92) and https://zenodo.org/records/17945143 (ref. 93)). The meta_EUR meta-analysis results can also be viewed in our PheWeb browser at https://nmrmeta.gi.ut.ee/ and the colocalization results can be explored at https://elixir.ut.ee/eqtl/nmr-coloc. The individual-level UKBB data are available for approved researchers through the UKBB data-access protocol (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access). The individual-level data from Estonia Biobank can be accessed through a research application to the Institute of Genomics of the University of Tartu (https://genomics.ut.ee/en/content/estonian-biobank). Source data are provided with this paper.

## Code availability

Data analysis code is available from https://github.com/ralf-tambets/EstBB-UKBB-metaanalysis/.


62. Milani, L. et al. The Estonian Biobank's journey from biobanking to personalized medicine. *Nat. Commun.* **16**, 3270 (2025).
63. Leitsalu, L. et al. Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
64. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
65. Julkunen, H. et al. Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nat. Commun.* **14**, 604 (2023).
66. Nightingale Health Biobank Collaborative Group Metabolomic and genomic prediction of common diseases in 700,217 participants in three national biobanks. *Nat. Commun.* **15**, 10092 (2024).
67. Ritchie, S. C. et al. Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. *Sci. Data* **10**, 64 (2023).
68. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
69. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
70. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
71. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
72. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
73. pyliftover. *PyPI* https://pypi.org/project/pyliftover/ (2024).
74. Shirley, M. D., Ma, Z., Pedersen, B. S. & Wheelan, S. J. Efficient 'pythonic' access to FASTA files using pyfaidx. Preprint at *PeerJ* https://doi.org/10.7287/peerj.preprints.970v1 (2015).
75. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. B* **82**, 1273–1300 (2020).
76. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the 'sum of single effects' model. *PLoS Genet.* **18**, e1010299 (2022).
77. 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).
78. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
79. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
80. Skrivankova, V. W. et al. Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration. *BMJ* **375**, n2233 (2021).
81. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).
82. Tambets, R., Kolde, A., Kolberg, P., Love, M. I. & Alasoo, K. Extensive co-regulation of neighboring genes complicates the use of eQTLs in target gene prioritization. *HGG Adv.* **5**, 100348 (2024).
83. Zhu, A. et al. MRLocus: Identifying causal genes mediating a trait through Bayesian estimation of allelic heterogeneity. *PLoS Genet.* **17**, e1009455 (2021).
84. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
85. van der Graaf, A. et al. MR-link-2: pleiotropy robust *cis* Mendelian randomization validated in three independent reference datasets of causality. *Nat. Commun.* **16**, 6112 (2025).
86. Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B. B. & Hopewell, J. C. Mendelian randomization with fine-mapped genetic data: choosing from large numbers of correlated instrumental variables. *Genet. Epidemiol.* **41**, 714–725 (2017).
87. UK10K Consortium et al The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
88. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
89. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
90. Varadi, M. et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).
91. Meng, E. C. et al. UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
92. Tambets, R. & Alasoo, K. Fine-mapping results for the EstBB-UKBB NMR metabolic trait meta-analysis. *Zenodo* https://doi.org/10.5281/zenodo.18132538 (2026).
93. Alasoo, K. & Jesse, M. Colocalisation results for the Tambets et al NMR metabolic trait GWAS study. *Zenodo* https://doi.org/10.5281/zenodo.17945143 (2025).



**Acknowledgements** We acknowledge the participants of the UK Biobank and Estonian Biobank for their contributions. The Estonian Genome Centre analyses were partially carried out in the High Performance Computing Center, University of Tartu. The research was conducted using the Estonian Center of Genomics/Roadmap II funded by the Estonian Research Council (project number TT17). We thank E. B. Fauman for feedback on an earlier version of the manuscript. This research has been conducted using the UK Biobank Resource under application numbers 91233 and 30418. Nightingale Health Plc is acknowledged for early access to the UK Biobank NMR metabolite data. UCSF ChimeraX was developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases. K.A., R.T. and M.J. were supported by the Estonian Research Council (grant no. PSG415 and MOB3ERC115). P.P., R.T., E.A, U.V., J.K. and T.E. were supported by the Estonian Research Council (grant no. PRG1291). I.R. was supported by the Estonian Research Council (grant no. PSG415). U.V. was supported by the Estonian Research Council (grant no. PSG1230). E.A. was supported by University of Tartu Development Fund bridging grant PLTGIARENG24. K.F. and A.K. were supported by the Estonian Research Council (grant no. PRG1197), Estonian Ministry of Education and Research Centres of Excellence (grant no. TK214) and European Union's Horizon Europe grant no. 101060011. N.T. was supported by the Estonian Research Council grant no. PRG1414. Z.K. was supported by the Department of Computational Biology of the University of Lausanne.


**Author contributions** R.T. performed the GWAS analysis and fine mapping on the EstBB and UKBB data. I.R. developed the initial GWAS workflow. N.T., A.K. and K.F. developed quality control criteria for the EstBB metabolite data. R.T. and K.A. designed the *cis*-Mendelian randomization and genome-wide Mendelian randomization analyses and interpreted the results. A.v.d.G. performed the MR-link-2 and MR-PCA analyses. M.J. performed genetic colocalization between molecular QTLs, metabolic traits and diseases. E.A. performed structural modelling of the BCKDH complex. D.Y. obtained SpliceAI and AlphaGenome predictions for all lead variants. A.V. set up the PheWeb browser. S.A. developed the colocalization browser. K.A., R.T., A.A. and P.P. prioritized association signals for follow-up analysis. T.E. established the NMR dataset within the EstBB. Estonian Biobank Research Team ensured the quality and provenance of EstBB data. P.P. led the conceptualization. P.P. and K.A. led the study design, funding acquisition and analysis planning. Z.K., T.E., K.A. and P.P. supervised the research. K.A., P.P., R.T, U.V., E.A. and J.K. wrote the manuscript with feedback from all authors.

**Competing interests** T.E. is the Head of the Supervisory Board of UniTartu Ventures OÜ, an investment and holding company of University of Tartu. A.v.d.G. has received consultancy fees from GERO. The other authors declare no competing interests.

**Additional information**

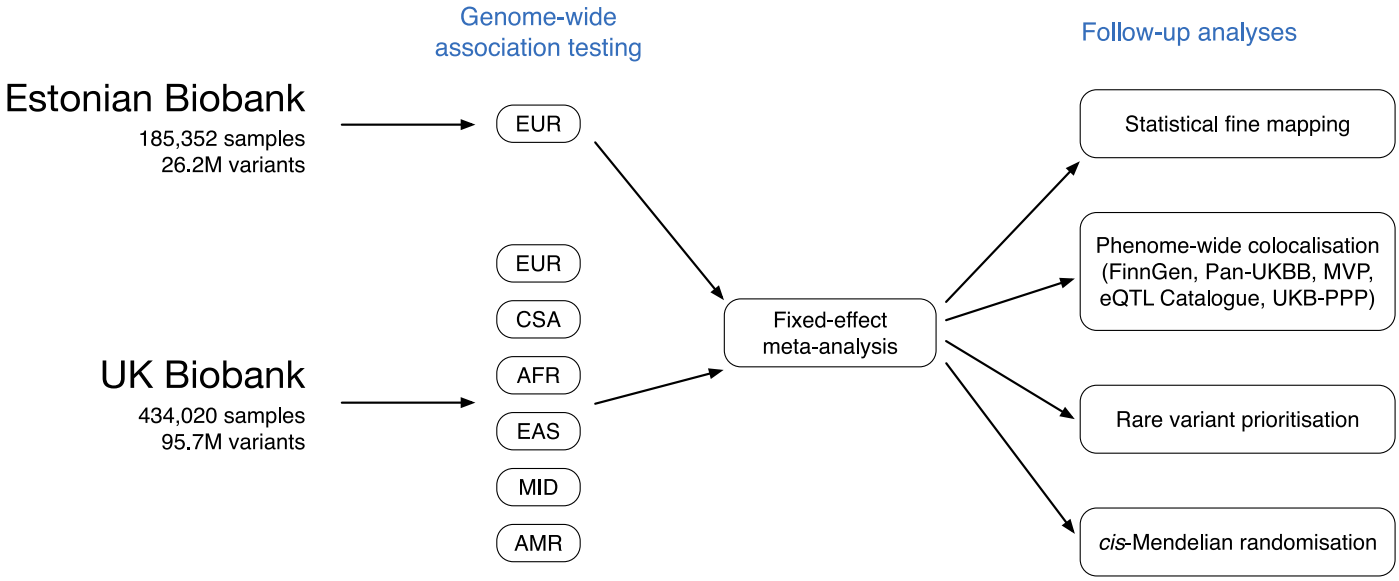

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10532-5.



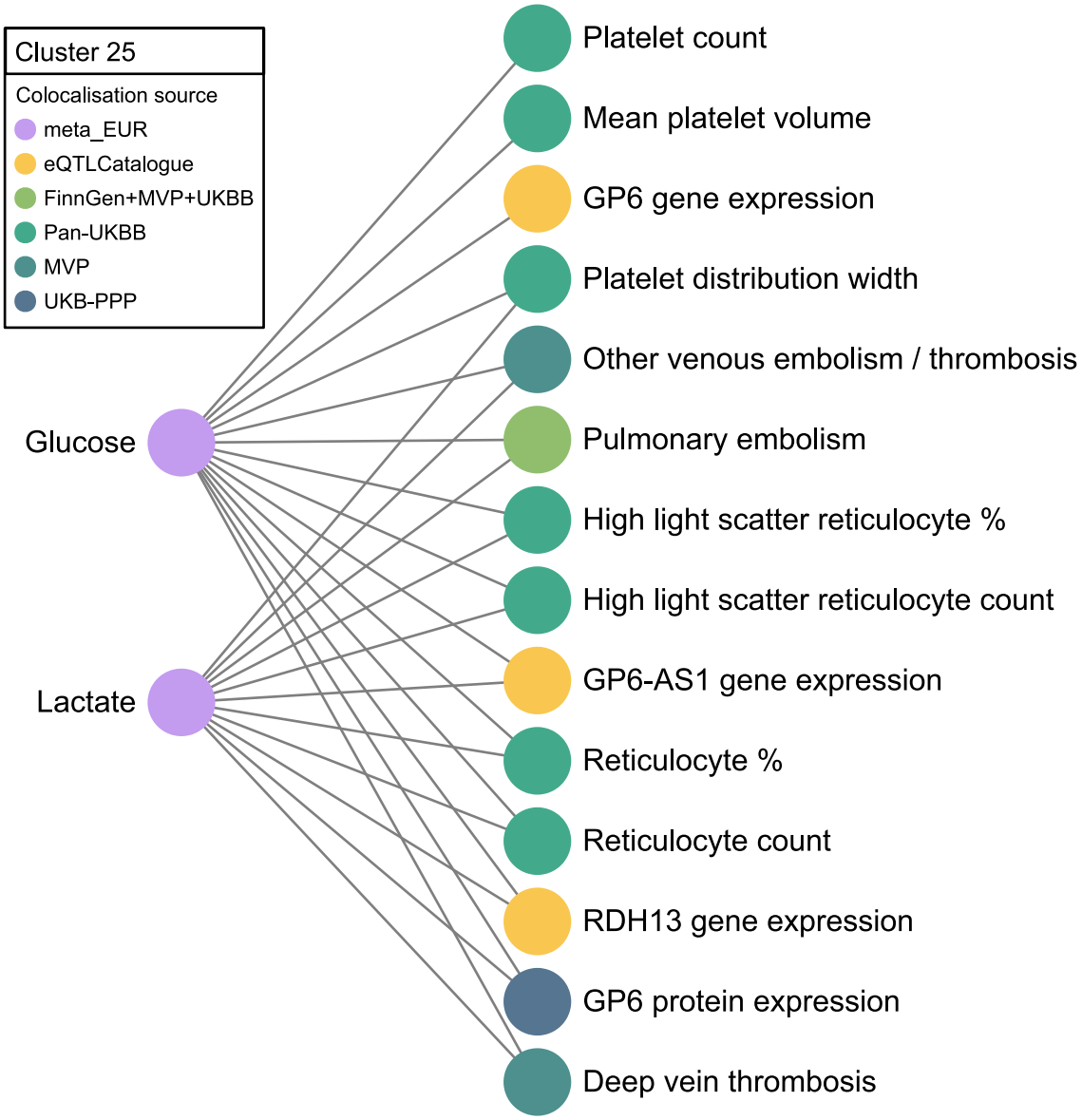
**Correspondence and requests for materials** should be addressed to Kaur Alasoo or Priit Palta.



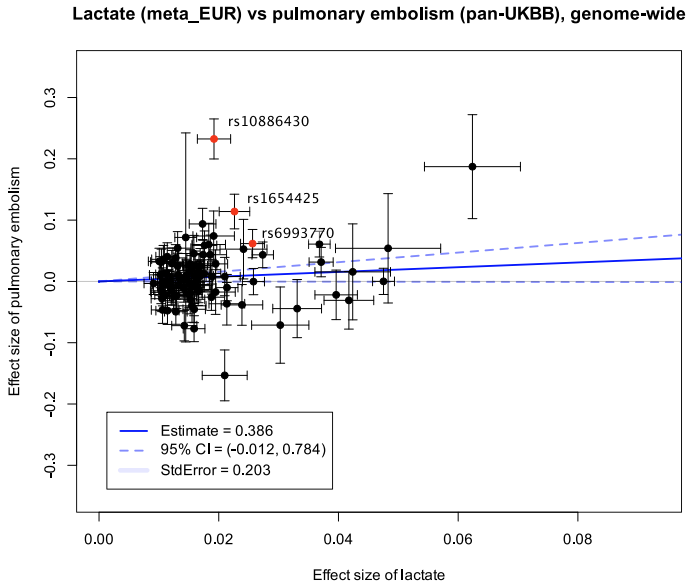
**Peer review information** *Nature* thanks Christopher Whelan and the other, anonymous, reviewers and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



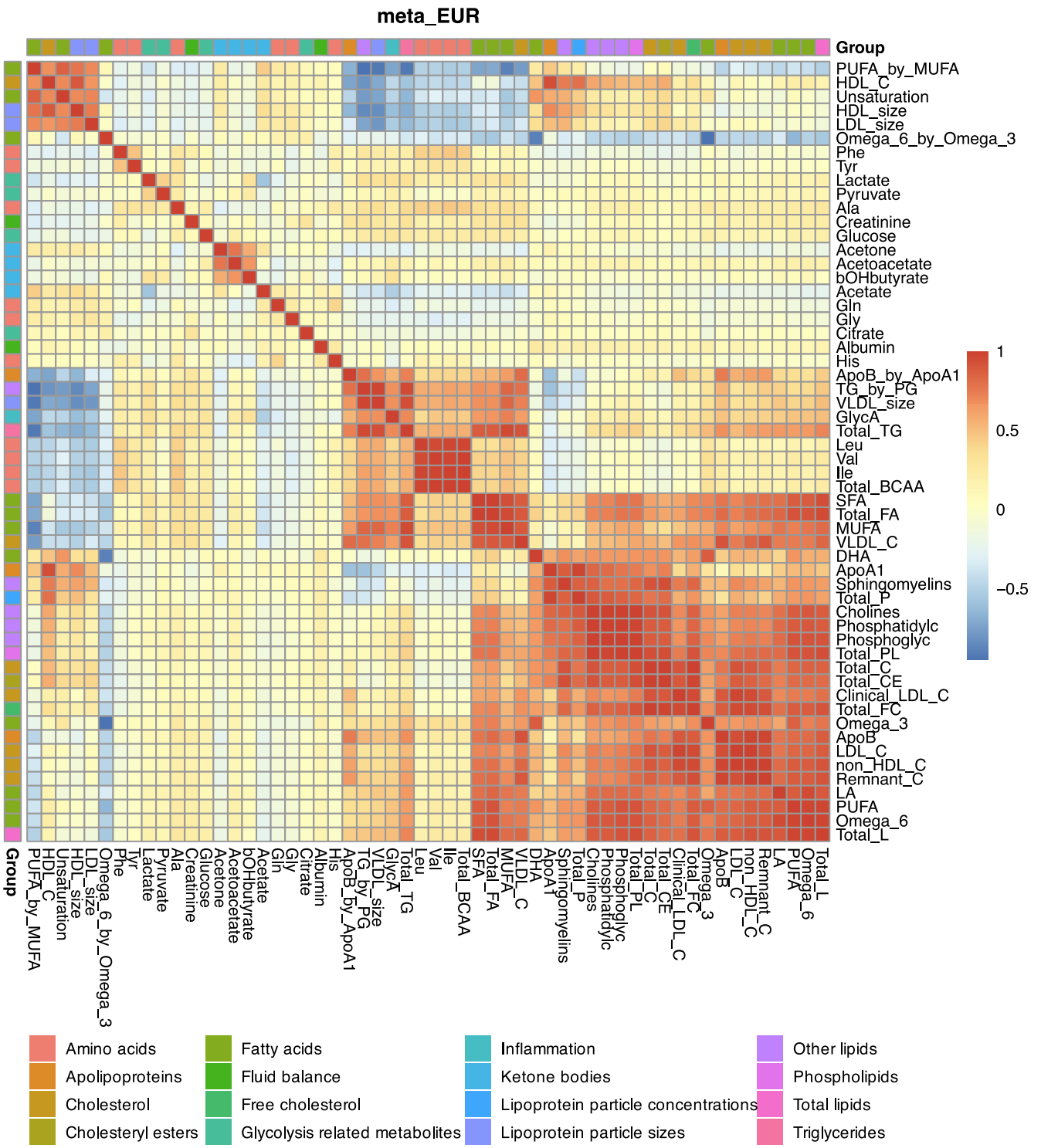

**Extended Data Fig. 1 | Outline of the study.** First, a separate GWAS was performed for each metabolic trait in the Estonian Biobank and six genetic ancestry groups of the UK Biobank: EUR (European), AFR (African), AMR (Admixed American), MID (Middle Eastern), EAS (East Asian), CSA (Central/South Asian) as defined by Pan-UKBB. These ancestry-specific GWAS results were then combined in a fixed-effect meta-analysis. Finally, several follow-up analyses were performed to demonstrate the value of the resource. MVP - Million Veterans Program; UKB-PPP - UK Biobank Pharma Proteomics Project.




**Extended Data Fig. 2 | Colocalization graph at the *GP6* locus.** Nodes correspond to association signals, and each edge indicates a colocalisation event (PP.H4 > 0.9) between the two signals. The graph has a bipartite structure, because we only tested for pairwise colocalizations between metabolic traits (left) and signals from other datasets (right). The genome-wide graph is sparse, making the identification of connected components (colocalisation clusters) straightforward.




**Extended Data Fig. 3 | Genome-wide random-effect inverse variance weighted MR using lactate (meta_EUR) as exposure and pulmonary embolism (FinnGen + MVP + UKBB) as outcome.** The points represent independent genome-wide significant ($p < 5 \times 10^{-8}$) GWAS hits for lactate in the meta_EUR meta-analysis. The lead variants at the *GP6* (rs1654425, 19-55027612-T-C), *GRK5* (rs10886430, 10-119250744-A-G) and *ZFPM2* (rs6993770, 8-105569300-A-T) loci have been highlighted. The points represent the standardized effect sizes of each genetic variant on the two traits with the error bars representing 95% confidence intervals.




**Extended Data Fig. 4 | Heatmap of pairwise genetic correlations between metabolic traits in the meta_EUR dataset.** The heatmap shows a representative subset of 56 metabolic traits from the main metabolic classes. The complete genetic correlation matrix for all 249 metabolic traits is presented in Supplementary Table 10.

**Extended Data Table 1 | Number of significant locus-metabolic trait pairs and unique lead variants**

| Biobank / genetic ancestry group | Locus-trait pairs | Unique lead variants | Sample size |
|---|---|---|---|
| EstBB | 17,364 | 1,164 | 185,352 |
| UKBB_EUR (European) | 39,666 | 3,154 | 413,897 |
| UKBB_CSA (Central/South Asian) | 686 | 39 | 8,652 |
| UKBB_AFR (African) | 531 | 38 | 6,439 |
| UKBB_EAS (East Asian) | 218 | 13 | 2,604 |
| UKBB_MID (Middle Eastern) | 16 | 8 | 1,500 |
| UKBB_AMR (Admixed American) | 5 | 3 | 928 |
| meta_EUR | 56,298 | 4,384 | 599,249 |
| meta_ALL | 57,103 | 4,417 | 619,372 |

Compared to Table 1, the significance threshold was set to $p < 2 \times 10^{-10}$ ($5 \times 10^{-8}/249$) to also account for the 249 tested metabolic traits.

## Extended Data Table 2 | Number of ancestry-specific lead variants detected in each UKBB non-EUR genetic ancestry group

| ancestry group | Unique lead variants | Variant missing in UKBB_EUR | Variant MAF < 0.1% in UKBB_EUR | Ancestry-specific lead variant fraction |
|---|---|---|---|---|
| UKBB_AFR | 143 | 37 | 31 | 47.6% |
| UKBB_AMR | 24 | 0 | 2 | 8.3% |
| UKBB_CSA | 113 | 14 | 14 | 24.8% |
| UKBB_EAS | 41 | 10 | 3 | 31.7% |
| UKBB_MID | 44 | 0 | 11 | 25% |

The counts are shown separately for lead variants either completely missing in the UKBB_EUR subset or having MAF < 0.1%.

nature portfolio

Corresponding author(s): Kaur Alasoo, Priit Palta

Last updated by author(s): Apr 1, 2026

# Reporting Summary

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

## Statistics

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

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*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<br>*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 | No software was used for data collection. |
| Data analysis | Genome-wide association tests were conducted using regenie v3.1.1. Custom inverse-variance weighted fixed-effect metaanalysis code was used, available at https://github.com/ralf-tambets/EstBB-UKBB-metaanalysis/. LD score regression (LDSC) v1.0.1 was employed to obtain pairwise genetic correlations for all 249 NMR metabolites. PLINK v1.90b6.26 was utilized to calculate pairwise LD between lead variants. susieR v0.14.276, LDStore v2.0 and rbcor (available at https://github.com/mkanai/rbcor) were used to perform statistical fine mapping. Colocalisation analyses were conducted using gpu-coloc, available at https://github.com/mjesse-github/gpu-coloc/. Missense and splice region variants were identified using Ensembl Variant Effect Predictor, SpliceAI and AlphaGenome APIs. "MendelianRandomization" R package v0.10.0 was used for genome-wide MR and cis-MR. In addition, MRLocus (v0.0.25), MR-link-2 (v1.0.0) and MR-PCA (v1.0.0) were used to assess the robustness of cis-MR results. Models of the three subunits of the BCKD complex in Figure 3b were generated using AlphaFold 3 via the AlphaFold Server. Molecular graphics were performed with UCSF ChimeraX v1.11. |

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

## Data

Policy information about availability of data

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

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

Complete genetic ancestry group-specific and meta-analysis association summary statistics from this study can be downloaded from the GWAS Catalog (accessions GCST90449363 - GCST90451603, Supplementary Table 12). GWAS lead variants, fine mapping credible sets, and colocalisation results are available from Zenodo (URLs https://zenodo.org/records/13937265, https://zenodo.org/records/18132538, and https://zenodo.org/records/17945143). The meta_EUR meta-analysis results can also be viewed in our PheWeb browser at https://nmrmeta.gi.ut.ee/ and the colocalisation results can be explored at https://elixir.ut.ee/eqtl/nmr-coloc. The individual-level UK Biobank data are available for approved researchers through the UK Biobank data-access protocol (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access). The individual-level data from Estonia Biobank can be accessed through a research application to the Institute of Genomics of the University of Tartu (https://genomics.ut.ee/en/content/estonian-biobank).

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | Sex was used as a covariate in genome-wide association tests and was assigned based on chromosome information in the genotyping data. Gender data was not taken into account. We did not perform sex-stratified GWAS as other GWAS studies of the same metabolic traits have have identified highly concordant genetic effects between sexes (e.g. Zoodsma et al, 2025). |
| Reporting on race, ethnicity, or other socially relevant groupings | We grouped UK Biobank participants into six genetic ancestry groups using information provided by the Pan-UKBB project. Both Estonian Biobank and UK Biobank consist predominantly of individuals of European descent. |
| Population characteristics | The Estonian Biobank is a volunteer-based biobank of approximately 200, 000 adult (18+) participants, reflecting the age, sex and geographical distribution of the adult Estonian population.<br>The UK Biobank is a longitudinal biomedical study of approximately half a million participants between 38-71 years old from the United Kingdom. |
| Recruitment | Recruitment of the Estonian Biobank participants was conducted on a volunteer basis. All biobank participants have signed a broad informed consent form and their blood sample collection was undertaken across the country between 2002 and 2021.<br><br>UK Biobank participant recruitment was conducted on a volunteer basis and took place between 2006 and 2010 in 22 different assessment centers throughout Scotland, England, and Wales. |
| Ethics oversight | Analysis of Individual level from the Estonian Biobank was carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data according to release application 6-7/GI/8988 from the Estonian Biobank.<br>The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee. This research was conducted using the UK Biobank Resource under application numbers 91233 and 30418. |

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

# Field-specific reporting

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

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

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

# Life sciences study design

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

| | |
|---|---|
| Sample size | Sample size refers to the number of participants in each genetic ancestry group. For autosomes, the sample sizes were 928 for UKBB_AMR; 1,500 for UKBB_MID; 2,604 for UKBB_EAS; 6,439 for UKBB_AFR; 8,652 for UKBB_CSA; 185,352 for EstBB; and 413,897 for UKBB_EUR. For X chromosome, the sample sizes were 925 for UKBB_AMR; 1,491 for UKBB_MID; 2,595 for UKBB_EAS; 6,411 for UKBB_AFR; 8,627 for UKBB_CSA; 185,352 for EstBB; and 412,523 for UKBB_EUR.<br><br>Previous GWAS studies of metabolic traits have demonstrated that these sampe sizes are sufficient to identify robust associations. |
| Data exclusions | Estonian Biobank individuals were excluded from the analysis if their genotyping call-rate was < 95%, if they were outliers of the absolute value of heterozygosity (> 3SD from the mean) or if sex defined based on heterozygosity of X chromosome did not match sex in phenotype |

data. To reduce the impact of population stratification on GWAS association analysis, principal component analysis of the genotype data was used to exclude a small number of participants that did not cluster with the other predominantly European ancestry individuals. Duplicate and monozygous twin detection was performed and one sample was removed out of the pair of duplicates.
The UK Biobank analysis was restricted to 434,020 individuals that were confidently assigned to one of six genetic ancestry groups by the Pan-UKBB project and had high quality NMR metabolic trait data available.
In both UK Biobank and Estonian Biobank, we excluded individuals with more than 5 missing metabolite measurements from the cohort.

Replication

We did not perform explicit replication of novel associations detected in the meta-analysis but we observed highly concordant genetic associations in the Estonian Biobank and UK Biobank (median genetic correlation 0.91). Also, 95% of the loci from a previous GWAS meta-analysis conducted on non-overlapping samples (Karjalainen et al, 2024) replicated in our study.

Randomization

Randomization was not relevant as no new primary data was collected in this study. We analyzed existing metabolic trait and genotype data data from the Estonian Biobank and the UK Biobank.

Blinding

The analysis conducted in this study did not involve group allocations were blinding would have been relevant and/or possible.

# 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 |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

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

## Plants

Seed stocks

*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes

*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication

*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*