

Assessing the contribution of rare-to-common protein-coding variants to circulating metabolic biomarker levels via 412,394 UK Biobank exome sequences

Abhishek Nag¹, Lawrence Middleton¹, Ryan S. Dhindsa^{2,3}, Dimitrios Vitsios¹, Eleanor Wigmore¹, Erik L. Allman¹, Anna Reznichenko⁴, Keren Carss¹, Katherine R. Smith¹, Quanli Wang², Benjamin Challis⁴, Dirk S. Paul¹, Andrew R. Harper¹, Slavé Petrovski¹

¹Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK

²Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Waltham, USA

³Department of Molecular and Human Genetics, Baylor College of Medicine and Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, TX 77030, USA

⁴Translational Science and Experimental Medicine, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

Corresponding author:

Slavé Petrovski
Vice-President, Centre for Genomics Research,
Discovery Sciences, BioPharmaceuticals R&D
AstraZeneca
Cambridge

United Kingdom

Email: slav.petrovski@astrazeneca.com

1 **Abstract**

2
3 Genome-wide association studies have established the contribution of common and low
4 frequency variants to metabolic biomarkers in the UK Biobank (UKB); however, the role of
5 rare variants remains to be assessed systematically. We evaluated rare coding variants for
6 198 metabolic biomarkers, including metabolites assayed by Nightingale Health, using
7 exome sequencing in participants from four genetically diverse ancestries in the UKB
8 (N=412,394). Gene-level collapsing analysis – that evaluated a range of genetic
9 architectures – identified a total of 1,303 significant relationships between genes and
10 metabolic biomarkers ($p < 1 \times 10^{-8}$), encompassing 207 distinct genes. These include
11 associations between rare non-synonymous variants in *GIGYF1* and glucose and lipid
12 biomarkers, *SYT7* and creatinine, and others, which may provide insights into novel disease
13 biology. Comparing to a previous microarray-based genotyping study in the same cohort, we
14 observed that 40% of gene-biomarker relationships identified in the collapsing analysis were
15 novel. Finally, we applied Gene-SCOUT, a novel tool that utilises the gene-biomarker
16 association statistics from the collapsing analysis to identify genes having similar biomarker
17 fingerprints and thus expand our understanding of gene networks.

18 **Introduction**

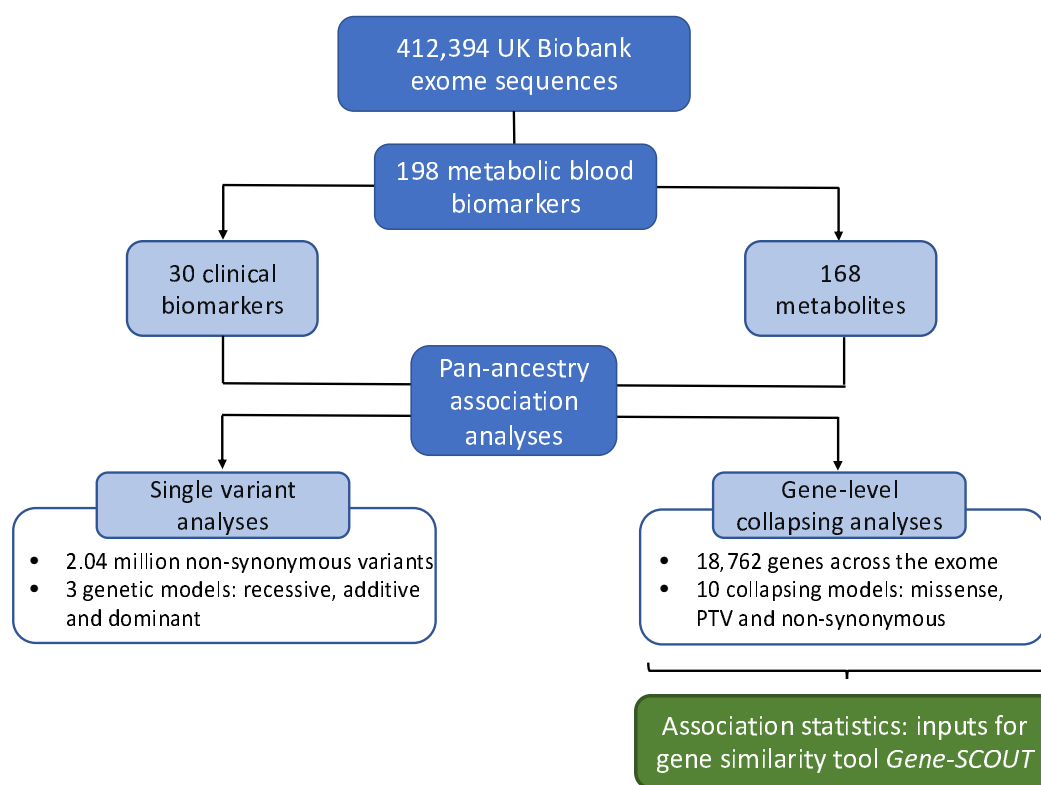
19 Metabolic blood biomarkers represent intermediate or end products of biochemical pathways
20 that can be used to diagnose and monitor human disease. The application of metabolic
21 biomarkers as intermediate traits to dissect the genetic basis of complex human diseases is
22 well-established. Investigating the genetic underpinnings of blood biomarkers can offer novel
23 insights into human disease mechanisms and, in turn, provide potential therapeutic targets.
24 Large-scale genome-wide association studies (GWAS) have so far identified hundreds of
25 genetic loci that regulate blood biomarker and metabolite levels¹⁻¹¹; however, difficulty in
26 mapping these loci to causal genes and interpreting functional effects of non-coding variants
27 have stymied the clinical impact for many of these associations¹².

28
29 The UK Biobank (UKB)¹³ is a large population-based resource of ~500,000 participants with
30 genetic data linked to a diverse set of phenotypic measurements. Genotype data from
31 microarrays and large population-based imputation panels have helped establish the
32 contribution of common and low frequency variants towards blood biomarkers in the UKB¹⁴.
33 The availability of exome sequences in the same population now allows for the exploration of
34 rare coding variants regulating metabolic blood biomarkers. Associations for rare coding
35 variants have demonstrably greater translational potential given their larger effect sizes¹⁵
36 and our ability to more directly interpret their functional impact¹⁶.

37
38 Using exome sequences from 412,394 unrelated participants across multiple genetic
39 ancestries in the UKB, we present findings of variant-level and gene-level (collapsing)
40 association tests for 198 metabolic blood biomarkers. We then introduce a novel tool, Gene-
41 SCOUT, that utilises this rich catalogue of gene-biomarker association statistics to identify
42 genes with similar biomarker fingerprints as a given (target) gene of interest and expand our
43 understanding of gene networks.

44 Results

45 In this study, we analysed 198 metabolic blood biomarkers, including 30 clinical blood
46 biomarkers related to glucose and lipid metabolism, renal and liver function (**Table S1A**),
47 and an additional 168 Nightingale assay blood metabolite measurements related to
48 lipoprotein lipids, fatty acids and their compositions, and various other low-molecular weight
49 metabolites¹⁷ (**Table S1B**). Most of the metabolic biomarkers pertain to lipid metabolism
50 (77%) and correlate highly with each other (**Figure 2a**). Many metabolic biomarkers also
51 demonstrate strong associations with clinical traits documented in the UKB (**Figure 2b**).

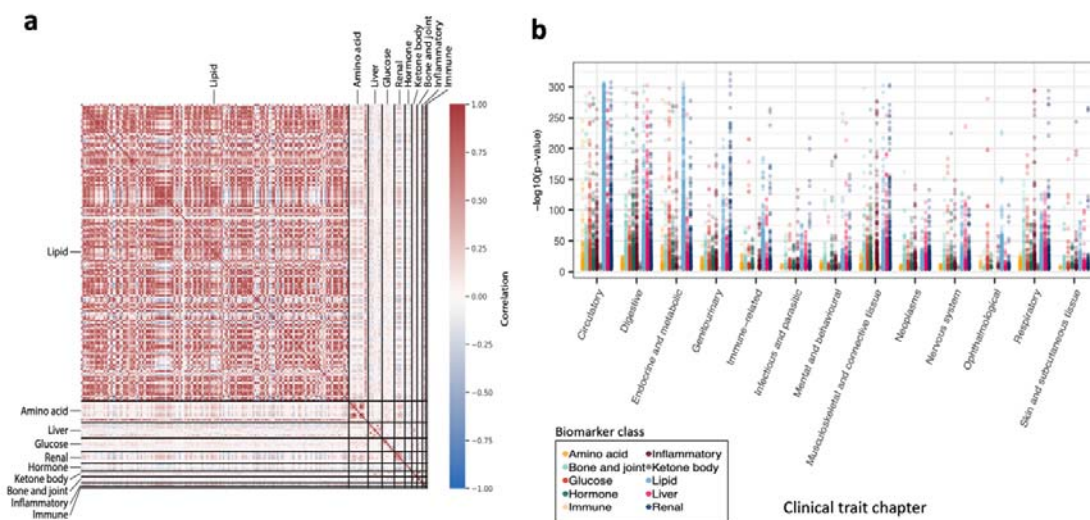


52 **Figure 1: A schematic of the association analyses that were conducted for the**
53 **metabolic blood biomarkers using the UK Biobank exome sequences**

54 The UK Biobank exome sequences were used to conduct single variant (under 3 genetic models) and
55 gene-level (under 10 collapsing models) association analyses for the clinical blood biomarkers (N=30)
56 and blood metabolite measurements (N=168). The gene-level association statistics for these
57 metabolitic biomarkers were used as inputs for the gene similarity tool *Gene-SCOUT*.
58

59
60 We first conducted a single variant analysis between the non-synonymous coding
61 variants (N=2,043,019 for the European ancestry subset) and the 198 metabolic biomarkers
62 (**Figure 1**). Excluding the MHC region, 19,351 significant variant-biomarker associations

63 ($p < 1 \times 10^{-8}$) were identified in the European subset of UK Biobank, which mapped to 12,217
 64 significant relationships between genes and biomarkers (**Tables S2A, S2B**). Pruning
 65 variants in linkage disequilibrium (r^2 threshold of 0.5) resulted in 9,738 significant gene-
 66 biomarker relationships. Notably, 243 distinct PTVs accounted for 1,366 significant
 67 associations, of which 602 (44%) were attributable to rare PTVs (MAF < 0.1%) with large
 68 effect sizes (> 0.5 SD) (**Figure 4a**) (**Tables S3A, S3B**). We identified 28 PTVs with MAF as
 69 low as 0.001% that achieved significance in the ExWAS: these include associations relating
 70 to several well-established and biologically plausible relationships such as *GOT1* and
 71 aspartate aminotransferase, *CST3* and cystatin C, *APOB* and cholesterol biomarkers, *ALPL*
 72 and alkaline phosphatase (**Table S3A**). Among other PTV findings that may provide new
 73 insights into important biology, a rare frameshift variant (MAF=0.03%) in *PLIN1* – a gene
 74 known to cause familial lipodystrophy¹⁸ – was associated with HDL-cholesterol (beta=0.40
 75 [0.27,0.53], $p = 1.6 \times 10^{-9}$), and a rare splice variant (MAF=0.09%) in *TNFRSF10B* – loss of
 76 which has been reported to promote survival of virus-infected liver cells¹⁹ – was associated
 77 with gamma glutamyltransferase (beta=0.21 [0.14,0.28], $p = 3.8 \times 10^{-9}$) (**Table S3A**).



78
 79 **Figure 2: Characteristics of metabolic blood biomarkers analysed in this study**
 80 The 198 metabolic blood biomarkers analysed in this study were grouped into the following 10
 81 biological classes: lipid, amino acid, liver, glucose, renal, hormone, ketone body, bone and joint,
 82 inflammatory, and immune. (a) The plot demonstrates that metabolic biomarkers belonging to the
 83 same biological class are correlated with each other. (b) Strong associations (plotted on the Y-axis)
 84 were observed between the metabolic biomarkers and 15,719 clinical traits (grouped by chapter)
 85 documented in the UKB²⁰.

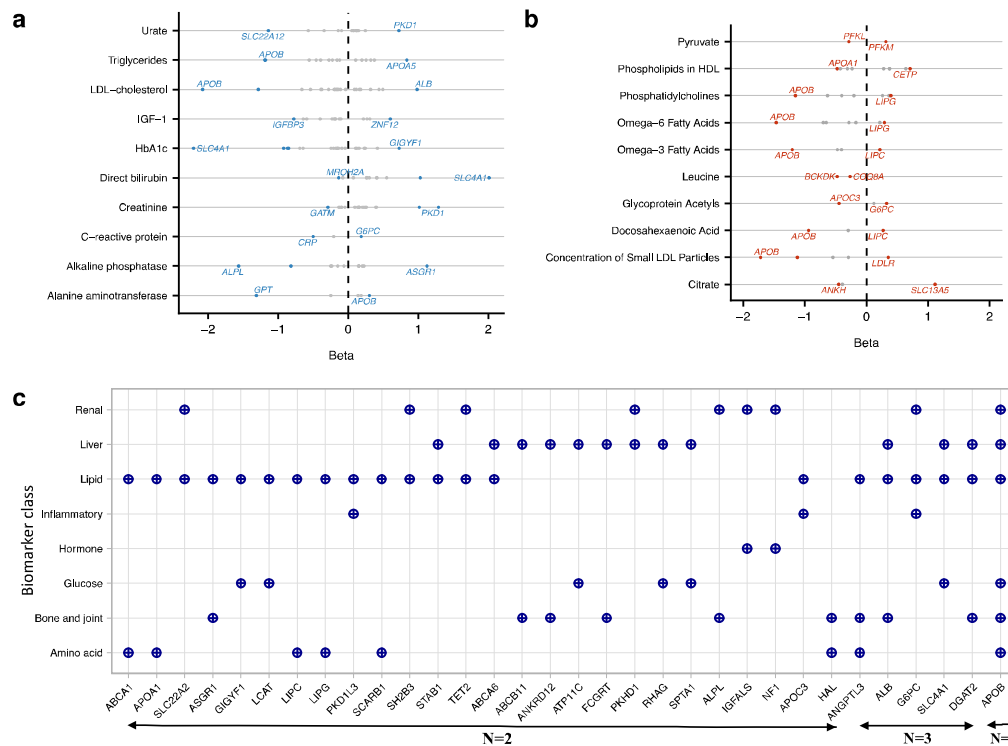
86 Next, we performed a gene-level collapsing analysis that tests the aggregate effect of
87 rare functional variants in each gene. We employed 10 different models to capture a diverse
88 range of genetic architectures (**Methods**). In the analysis involving individuals of European
89 ancestry alone, we identified 1,303 significant relationships between genes and metabolic
90 biomarkers ($p < 1 \times 10^{-8}$) (**Tables S4A, S4B; Figures 3a, 3b**). Most (68%, 880/1,303) gene-
91 biomarker relationships detected via the collapsing analysis were captured through models
92 that exclusively focused on PTV classes (“ptv” and “ptv5pcnt”), while the remaining 32%
93 were attributable to models that incorporated missense variants. We detected more
94 significant associations using our “ptv” and “ptv5pcnt” models than a prior study²¹ that also
95 performed gene-level collapsing analysis using the UKB exome sequence data, albeit with a
96 different analytical framework. For instance, associations between PTVs in 12 genes and
97 HbA1c that we detected were not reported in the other study: this includes the glucose
98 metabolism genes *HK1* and *G6PC2* (**Figure S1**). We also extended our gene-level
99 collapsing analysis to include all ancestral groups in the UKB (**Methods**). This detected an
100 additional 51 significant gene-biomarker relationships (**Tables S5A, S5B**). For the gene-
101 biomarker relationships that were significant only in the pan-ancestry analysis, we did not
102 observe a significant difference in the estimated effect size between the European-only and
103 the pan-ancestry analyses ($p=0.83$), suggesting that increased statistical power rather than
104 ancestry-specific effects is the more likely reason why these associations were identified in
105 the pan-ancestry analysis. One such association detected exclusively in the pan-ancestry
106 analysis was between recessive carriers of nonsynonymous variants in the membrane
107 transport gene *SYT7* and blood creatinine levels (number of QV carriers=5, $\beta=2.17$
108 [1.46,2.87], $p=1.6 \times 10^{-9}$). With 3 of the 5 carriers observed in the South Asian and African
109 ancestry participants, the pan-ancestry analysis facilitated detection of this association,
110 which was not study-wide significant in the European subset (number of QV carriers=2,
111 $\beta=1.17$ [0.06,2.28], $p=0.04$). Remarkably consistent with the biomarker findings,
112 recessive carriers of *SYT7* PTVs demonstrate a increased risk of glomerular disease in the

113 pan-ancestry analysis (OR=92.1 [12.1,713.2], $p=2.6 \times 10^{-5}$), but the clinical association on its
114 own is not yet study-wide significant.

115 The significant gene-level relationships from the collapsing analyses encompassed
116 207 distinct genes, of which 32 were associated with biomarkers across different biological
117 classes (**Figure 3c**). This includes *GIGYF1*, a tyrosine kinase receptor signalling protein, in
118 which rare PTVs were associated with biomarkers of glucose [glucose (beta=0.59
119 [0.42,0.76], $p=7.9 \times 10^{-12}$) and HbA1c (beta=0.73 [0.57,0.88], $p=4.5 \times 10^{-20}$)] and cholesterol
120 metabolism [total cholesterol (beta=-0.66 [-0.82,-0.50], $p=2.0 \times 10^{-15}$), LDL-cholesterol (beta=-
121 0.61 [-0.78,-0.45], $p=3.4 \times 10^{-13}$) and apolipoprotein B (beta=-0.60 [-0.77,-0.44], $p=1.3 \times 10^{-12}$)].
122 Additionally, among clinical traits documented in the UKB²⁰, significant associations were
123 observed for rare PTVs in *GIGYF1* with the risk of hypothyroidism (OR=4.2 [2.7,6.6],
124 $p=7.1 \times 10^{-9}$) and type 2 diabetes (OR=4.0 [2.7,5.8], $p=1.0 \times 10^{-10}$). Since hypothyroidism is
125 known to raise LDL-cholesterol levels, we subsequently tested the *GIGYF1*-LDL-cholesterol
126 association adjusted for a diagnosis of hypothyroidism. The signal between *GIGYF1* PTVs
127 and LDL-cholesterol (adjusted for the effect of statins) remained significant upon adjusting
128 for hypothyroidism (beta=-0.55 [-0.71,-0.38]; $p=6.2 \times 10^{-11}$), suggesting that the *GIGYF1* locus
129 likely influences cholesterol levels independent of solely thyroid hormone-mediated
130 pathways. Thus, by leveraging information from over 400,000 UKB exomes, our study
131 provides a more comprehensive picture regarding *GIGYF1*'s biomarker fingerprint and
132 associated clinical traits, expanding on previously reported common⁷ and rare variant
133 associations^{21,22} at this locus.

134 We observed that adjusting biomarkers for medications that influence their levels can
135 also improve detection of associations: 31/84 (37%) significant gene-biomarker relationships
136 for apolipoprotein B, LDL-cholesterol, total cholesterol, and urate from the collapsing
137 analysis were detected only after we adjusted their values for commonly prescribed
138 medications (**Table S4A**). This includes association between putatively damaging missense
139 variants and PTVs in *HMGCR* ("flexdmg" model) and LDL-cholesterol (medication-adjusted:
140 beta=-0.19 and $p=1.7 \times 10^{-11}$; medication-unadjusted: beta=-0.15 and $p=6.1 \times 10^{-8}$), which

141 validates the value of medication adjustment to untangle the effects of therapeutic
 142 intervention vs natural aberration of *HMGCR*. Moreover, for gene-biomarker relationships
 143 that were significantly associated in both the medication-unadjusted and the medication-
 144 adjusted analyses (N=52), the absolute effect sizes were observably higher in the latter
 145 (**Figure S2**), but the difference was not statistically significant in the current sample (Mann
 146 Whitney $p=0.28$).



147 **Figure 3. Significant relationships between genes and metabolic blood biomarkers**
 148 **identified in the collapsing analysis**

149 (a, b) Significant gene relationships ($p < 1 \times 10^{-8}$) identified for select clinical biomarkers and metabolites
 150 in the collapsing analysis have been shown. The genes with the highest absolute effect sizes for each
 151 have been labelled.

152 (c) The plot lists the 32 genes that were significantly associated ($p < 1 \times 10^{-8}$) with metabolic biomarkers
 153 across two or more biological classes in the collapsing analysis. For each such gene, the
 154 corresponding biological classes have been indicated.

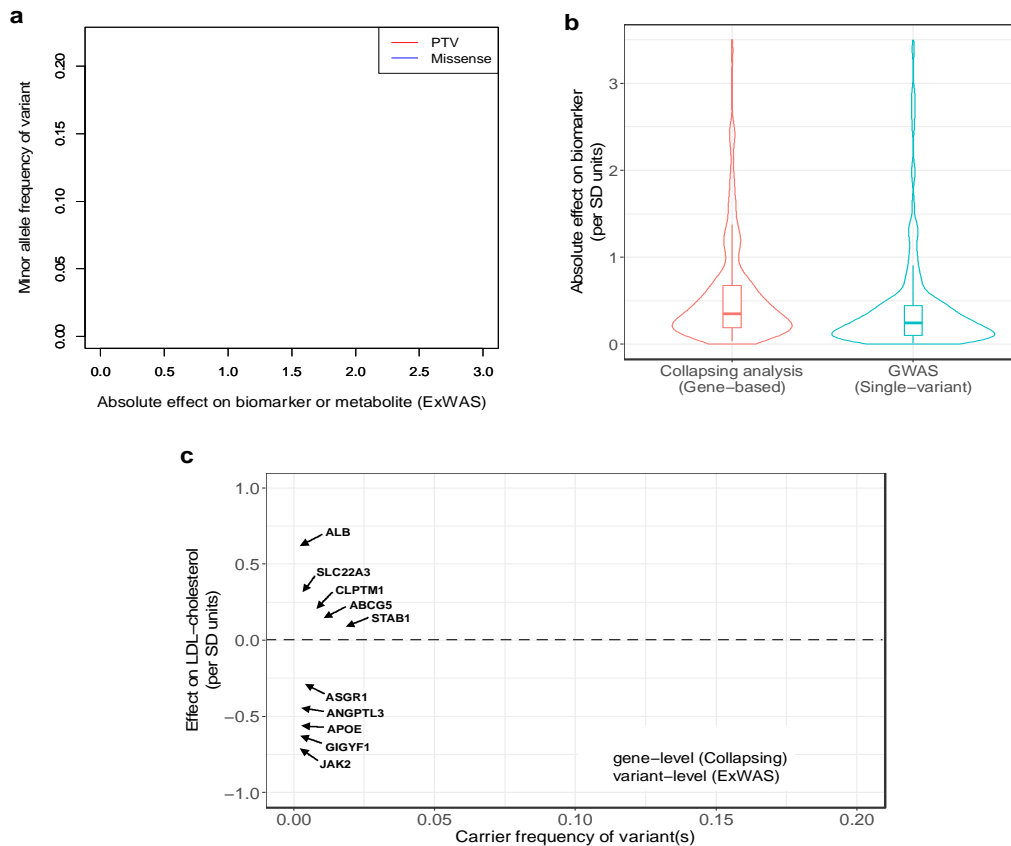
155 **Gene-level collapsing analysis: capturing allelic series**

156 We observed that 17% (215/1,303) of significant relationships between genes and
 157 metabolic biomarkers from the collapsing analysis did not achieve significance in the
 158 respective variant-level ExWAS (**Table S6**). Next, we also compared the gene-biomarker
 159 relationships that achieved significance in the collapsing analysis (**Tables S4A, S7**) and the
 160
 161

162 microarray-based GWAS¹⁴ (as per a less stringent significance threshold: $p < 1 \times 10^{-7}$) for the
163 32 biomarkers (28 blood and 4 urinary biomarkers) analysed in both studies. Of the
164 significant gene-biomarker relationships identified in the collapsing analysis, 40% (142/357)
165 were not detected in the microarray-based GWAS (**Table S8**). These include associations
166 for well-known drug target genes such as *HMGCR* (with LDL-cholesterol) and *PPARG* (with
167 HDL-cholesterol). Furthermore, the effect size estimates were significantly higher in the
168 collapsing analysis than in microarray-based GWAS for 215 gene-biomarker relationships
169 detected via both approaches (Mann-Whitney $p = 8.0 \times 10^{-6}$) (**Table S9; Figure 4b**). One likely
170 explanation for this is that by testing aggregate effects of rare putative functional variants in
171 a gene, associations arising from collapsing analysis are enriched for larger effects (**Figure**
172 **4c**). Collectively, these results highlight that application of a gene-based rare variant
173 collapsing analysis to large-scale exome sequencing can increase power to capture
174 associations that are driven by an allelic series, and thus expand our understanding of the
175 genetic architecture of traits, especially where a lot of success has already been achieved
176 through traditional microarray-based GWAS.

177 *SLC4A1*, which encodes a chloride/bicarbonate anion exchange protein in the red cell
178 membrane, represents one such gene for which multiple signals were detected in the gene-
179 level collapsing analysis but not in the ExWAS. We observed 32 carriers for 28 distinct
180 *SLC4A1* PTVs, of which 25 (89%) were private (i.e., observed in a single carrier) (**Figure S3**).
181 Overall, *SLC4A1* PTVs were significantly associated with a strong reduction in HbA1c (beta=
182 2.2 [-2.6,-1.8], $p = 1.4 \times 10^{-25}$) and LDL-cholesterol (beta=-1.0 [-1.4,-0.7], $p = 8.0 \times 10^{-9}$), while also
183 showing strong increases in total bilirubin (beta=1.7 [1.3,2.0], $p = 1.1 \times 10^{-22}$) and direct
184 bilirubin (beta=2.0 [1.7,2.4], $p = 1.8 \times 10^{-28}$). Among clinical phenotypes, *SLC4A1* PTVs are
185 significantly associated with disorders of reduced red cell membrane stability such as
186 hereditary spherocytosis and hereditary haemolytic anaemia, but not with any phenotype
187 related to glucose or lipid metabolism ($p < 1 \times 10^{-5}$). Similarly, in ClinVar, several missense and

188 loss-of-function mutations in this gene are reported as pathogenic for hereditary
 189 spherocytosis. Therefore, we further tested the *SLC4A1*-biomarker associations after
 190 adjusting for the diagnosis of hereditary spherocytosis or hereditary haemolytic anaemia
 191 and relevant blood cell indices, including red cell distribution width (RDW) and mean
 192 corpuscular haemoglobin



193
194
195
196
197
198
199
200
201
202
203
204
205
206
207

Figure 4: Effects of coding variants on metabolic blood biomarkers

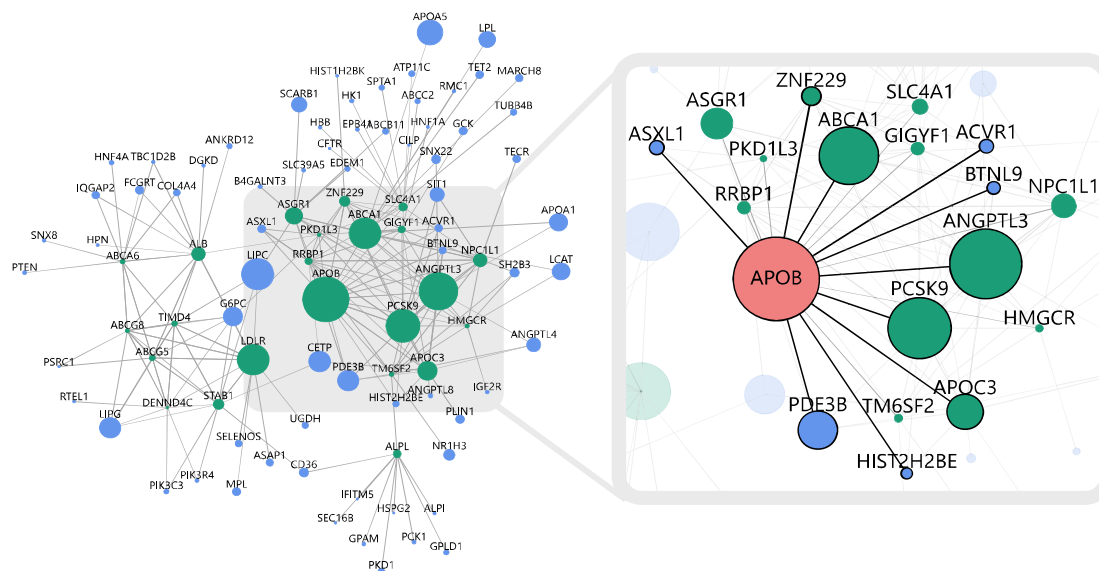
(a) Absolute effect sizes for missense variants and PTVs significantly associated ($p < 1 \times 10^{-8}$) with metabolic biomarkers in the single variant analysis (ExWAS) as a function of their minor allele frequency (in cases where a missense variant or a PTV was significantly associated with more than one biomarker, the association with the highest absolute effect size was selected). (b) The effect sizes estimated in the gene-based collapsing analysis and the Sinnott-Armstrong *et al*/microarray-based GWAS¹⁴ were compared for the gene-biomarker relationships that were significantly associated in both (N=215). For each significant gene-biomarker relationship, the collapsing model (from the collapsing analysis) and the individual variant (from the microarray-based GWAS) with the highest absolute effect sizes were selected. The effect sizes estimated in the collapsing analysis were significantly higher than that in the GWAS (Mann-Whitney $p = 8.0 \times 10^{-6}$). (c) Comparing effect sizes for individual variants and aggregate of rare variants (in a gene) that were significantly associated ($p < 1 \times 10^{-8}$) with LDL-cholesterol. Some examples of genes significantly associated with LDL-cholesterol have been highlighted. The Y-axis has been capped at 1 SD units for visual clarity.

208
209 concentration (MCHC). The gene-based *SLC4A1* PTV signals remained significant in the
210 adjusted analyses (**Table S10**). Although PTVs in this gene may be independently associated
211 with biomarkers of glucose, lipid and bilirubin metabolism, we cannot rule out the
212 possibility of under-reporting of hereditary spherocytosis and hereditary haemolytic
213 anaemia in the UKB that explains these observations. The *SLC4A1* enigma is consistent with
214 previous reports of other red blood cell loci that have also been significantly associated with
215 HbA1c²³.

216 **Gene-SCOUT: estimating gene similarity based on cohort statistics from collapsing analysis**

217 We considered the opportunity to leverage this new and rich catalogue of gene-level
218 association statistics from the collapsing analysis to determine genes with similar biomarker
219 fingerprints. To achieve this, we developed a gene similarity tool 'Gene-SCOUT'²⁴, that
220 solely uses the gene-level collapsing analysis statistics across the studied biomarkers to
221 identify genes with the most comparable biomarker genetic associations as a given gene of
222 interest. No other information is used in constructing the gene similarity scores. Since this
223 tool estimates gene similarity for an index gene by selecting features based on the
224 significance cut-off of $p < 1 \times 10^{-5}$, gene neighbours could not be determined for genes that did
225 not achieve association $p < 1 \times 10^{-5}$ with any biomarker feature. Accordingly, for our feature set
226 comprising of 198 biomarkers, we were able to determine gene similarity for 3%
227 (536/18,762) of human protein-coding genes. To illustrate Gene-SCOUT's application, we
228 selected the 24 genes that were significantly associated ($p < 1 \times 10^{-8}$) with LDL-cholesterol in
229 the collapsing analysis. We used each gene in this set as a seed gene to construct a
230 network figure that demonstrates their respective gene neighbours (**Figure 5**). Using *APOB*
231 as an example, we observe that genes with the most comparable biomarker fingerprint as
232 *APOB* include: *ABCA1*, *ACVR1*, *APOC3*, *ANGPTL3*, *ASGR1*, *ASXL1*, *BTNL9*, *GIGYF1*,
233 *HIST2H2BE*, *HMGCR*, *NPC1L1*, *PCSK9*, *PDE3B*, *PKD1L3*, *RRBP1*, *SLC4A1*, *TM6SF2* and
234 *ZNF229*. For some of these genes (e.g., *ZNF229*, *ACVR1*), the links with lipid metabolism

235 appear to be novel, in addition to the recently described relationships for *GIGYF1*^{21,22}.
236 Inhibition of *APOB*, such as through mipomersen, is known to be clinically effective in
237 reducing blood cholesterol levels. Remarkably, 5 (namely, *APOC3*, *ANGPTL3*, *HMGCR*,
238 *NPC1L1* and *PCSK9*) of the 18 genes (28%) determined to have similar cohort-level genetic
239 associations for biomarkers as *APOB* are also targets of lipid-lowering drugs that are already
240 approved or in various stages of development (<https://www.fda.gov/drugs>).



241
242 **Figure 5: Network figure demonstrating the gene neighbours i.e., genes with most**
243 **similar biomarker genetic signals, as the set of genes that were significantly**
244 **associated with LDL-cholesterol in the collapsing analysis**

245 The 24 genes that were significantly associated ($p < 1 \times 10^{-8}$) with LDL-cholesterol in the collapsing
246 analysis were used as seed genes (green nodes) to construct a network figure demonstrating
247 respective gene neighbours (edges). Non-seed genes are represented using blue nodes. The size of
248 a gene node corresponds to the number of features (of total 198) that the gene is associated with at
249 $p < 1 \times 10^{-5}$. The inset demonstrates the genes with most similar biomarker signature as *APOB* – these
250 include the ten closest genes for *APOB* as the seed gene (black edges) and other seed genes that
251 have *APOB* among their ten closest genes (grey edges).

252 **Discussion**

253 We used the 454,796 UK Biobank exome sequences to explore the contribution of private-
254 to-rare-to-common protein-coding variation for 30 clinical biomarkers and 168 metabolite
255 measurements. By adopting variant- and gene-level analysis frameworks and assessing the
256 full allelic frequency spectrum, we have expanded our understanding of the genetic
257 architecture of metabolic biomarkers that have previously been studied through microarray
258 data. The finding that 17% of gene-biomarker relationships detected in the gene-level
259 collapsing analysis were not identified in the single variant analysis demonstrates the power
260 of testing an aggregate effect of rare variants in a gene encompassing a range of genetic
261 architectures. We also illustrated how adjusting biomarker values for commonly prescribed
262 medications can improve signal detection.

263 There are several strengths of our study that might have implications for identifying or
264 validating drug targets. First, by virtue of focusing on coding variants, the observed
265 associations could provide a more causal link between a gene and a blood biomarker^{25–28}.
266 Moreover, association signals emerging from collapsing analysis are driven by an aggregate
267 effect of multiple rare variants (allelic series) that tend to be less impacted by local LD
268 structure. This contrasts with associations identified in microarray-based GWAS that often
269 map to non-coding regions of the genome or to regions of extensive LD, making it more
270 challenging to pinpoint the underlying causal variants.

271 Associations involving putative functional variants can also indicate the desired
272 modulation of the target gene e.g., upregulation or downregulation of the target gene
273 product, required to mitigate the risk of the disease related to the associated biomarker. For
274 instance, we observed a total of 182 associations for rare (MAF<0.1%) PTVs with the 30
275 blood biomarkers, which is >3-fold more than the 53 conditionally independent PTV
276 associations (for the same set of blood biomarkers) reported in the microarray-based
277 analysis¹⁴.

278 We also introduce a novel tool (Gene-SCOUT) that utilises all the gene-level
279 collapsing analysis statistics across the 198 studied biomarkers to estimate a ‘similarity’
280 metric between genes. With the aid of specific examples, we were able to demonstrate that
281 this approach can successfully identify genes with similar biomarker fingerprints.

282 While there are certain advantages of using blood biomarkers to dissect the genetics
283 of complex human diseases, including greater statistical power offered by quantitative traits
284 and better insights into biological pathways underlying associations, further work is
285 necessary to establish the causal relationship between genetic loci identified using
286 biomarkers or metabolites and the related disease(s). For instance, we observed
287 associations between certain biomarkers and variants in genes that encode them (e.g., *ALB*
288 with albumin, and *CST3* with cystatin C) – although such associations serve as excellent
289 positive controls that demonstrate the robustness of our analysis framework, they may not
290 offer novel insights into disease pathophysiology.

291 Using the largest collection of exome sequences linked to a diverse set of circulating
292 metabolic biomarkers, we demonstrate the value of this resource to enhance our
293 understanding of human diseases, and potentially, provide novel therapeutic targets focused
294 on mimicking natural human genetic discoveries. Our study also strongly supports the use of
295 a gene-based collapsing framework to uncover gene-biomarker relationships that are driven
296 by an aggregate effect of multiple rare, non-synonymous variants.

297

298 **Methods**

299

300 **UK Biobank (UKB) Resource**

301 The UKB resource¹³ is a prospective cohort study of ~500,000 individuals from across the
302 United Kingdom, aged between 40 and 69 years. The average age at recruitment for the
303 sequenced participants was 56.5 years and 54% of the sequenced cohort are females.
304 Participant data, obtained through questionnaires and assessment visits, include health
305 records that are periodically updated by the UKB, self-report survey information, linkage to
306 death and cancer registries, urine and blood biomarkers, imaging data, accelerometer data
307 and various other phenotypic endpoints¹³. All study participants provided informed consent.
308 For this study, data from the UKB resource was accessed under the application number
309 26041.

310

311 **Metabolic blood biomarkers**

312 Routine clinical blood biomarkers related to glucose and lipid metabolism, renal and liver
313 function, among others (N=30), were measured in the majority of the ~500,000 UKB
314 participants (**Table S1A**). Additionally, 168 blood metabolites, including lipoprotein lipids,
315 fatty acids and their compositions, and various low-molecular weight metabolites, were
316 profiled in a subset of ~120,000 UKB participants by Nightingale Health using nuclear
317 magnetic resonance spectroscopy¹⁷ (**Table S1B**). Samples with a 'quality control (QC) flag'
318 for the blood metabolites were excluded. In total, we analysed 198 metabolic blood
319 measures: 30 clinical biomarkers and 168 metabolites. We applied rank-based inverse-
320 normal transformation to the measurements prior to performing association analyses.

321 For four blood biomarkers (LDL-cholesterol, total cholesterol, apolipoprotein B and
322 urate) we adjusted for the effect of commonly prescribed medications known to influence
323 their levels. For LDL-cholesterol, total cholesterol and apolipoprotein B, we adjusted for the
324 effect of statins based on their 'statin adjustment factors', previously estimated in the UKB as
325 0.684, 0.749 and 0.719, respectively¹⁴. Similarly, we adjusted urate for the effect of

326 allopurinol based on an ‘allopurinol adjustment factor (0.810)’, calculated using an approach
327 identical to that described for statins¹⁴.

328

329 **Whole-exome sequencing and bioinformatics pipeline**

330 Whole-exome sequences for 454,988 UKB participants were generated at the Regeneron
331 Genetics Center as part of a pre-competitive data generation collaboration between AbbVie,
332 Alnylam Pharmaceuticals, AstraZeneca, Biogen, Bristol-Myers Squibb, Pfizer, Regeneron
333 and Takeda²⁹. The exome sequencing procedure and the relevant QC steps have been
334 detailed previously in Szustakowski *et al* (2021)²⁹ and Wang *et al* (2021)²⁰. The FASTQ
335 sequences that were made available were first aligned, following which, single nucleotide
336 variants (SNVs) and small indels were called using Illumina’s DRAGEN Bio-IT Platform
337 Germline Pipeline v3.0.7 on the Amazon Web Services cloud compute platform available at
338 AstraZeneca’s Centre for Genomics Research. SNPEff v4.3³⁰ was used to annotate the
339 ‘most damaging effect’ predicted for each protein coding variant. In addition, we used certain
340 other bioinformatic tools such as missense tolerance ratio (MTR) scores³¹ to identify regions
341 of protein coding genes under constraint for missense variants, and REVEL³² to prioritise
342 coding variants based on their predicted deleteriousness. Further details on how these tools
343 were applied to the UKB exome sequencing dataset have been previously described²⁰.

344

345 **Selection of UKB samples for the association analyses**

346 Prior to performing the association analyses, we excluded samples from the available UKB
347 exome sequencing dataset (N=454,796) based on the following QC measures²⁰ (**Figure S4**):
348 (i) *DNA contamination*: VerifyBAMID freemix (measure of DNA contamination) >4%.
349 (ii) *Coverage depth*: $\geq 10x$ for <94.5% of the consensus coding sequence (CCDS release
350 22).
351 (iii) *Relatedness*: 2nd-degree relatives or closer (equivalent to kinship coefficient >0.0884), as
352 estimated using the --kinship function in KING v2.2.2³³.

353 Additionally, to perform analyses accounting for differing genetic ancestry, we
354 assigned samples to one of the four major ancestral groups (minimum 1,000 participants):

355 European (N=394,695), South Asian (N=8,078), East Asian (N=2,209) and African
356 (N=7,412). This was done by excluding participants: (i) with predicted genetic ancestry
357 <0.99 (for European ancestry) or <0.95 (for the remaining ancestries), as estimated using
358 PEDDY v0.4.2; or (ii) lying outside four standard deviations for the top four principal
359 components for each of the genetic ancestry collections.

360

361 **Association analysis for metabolic blood biomarkers**

362 A number of stringent variant-level QC steps, detailed previously²⁰, were applied to select
363 variant calls with highest confidence for association testing. Briefly, the variant-level QC
364 criteria included coverage depth, genotype and mapping quality scores, DRAGEN variant
365 status, read position rank sum score (RPRS), mapping quality rank sum score (MQRS),
366 alternate allele read proportion for heterozygous calls, proportion of samples failing any of
367 these QC criteria, and gnomAD-related filters.

368 Association testing between the metabolic blood biomarkers and the variants in the
369 exome sequencing dataset was conducted using two complementary analytical approaches
370 **(Figure 1):**

371 (i) Single variant exome-wide association study (ExWAS)

372 (ii) Gene-level collapsing analysis

373 We conducted the association analyses separately in the European ancestry
374 participants as this comprised the single largest ancestral group in this resource and for all
375 four ancestries combined ('pan-ancestry' analysis).

376

377 Single variant exome-wide association study (ExWAS)

378 In the single-variant analysis (hereafter referred to as 'ExWAS'), variants that passed the QC
379 steps were filtered further to include those that had a minimum of six carriers (equivalent to
380 MAF>0.0008% in the European ancestry subset). We additionally excluded variants that had
381 one of the following annotations as their most damaging effect as per SNPEff:

382 *3_prime_UTR, 5_prime_UTR, initiator_codon_variant, non_coding_transcript_exon_variant,*

383 and *synonymous_variant*. The remaining non-synonymous coding variants (N=2,043,019 in
384 the European ancestry subset) were used to perform the ExWAS.

385 The ExWAS was conducted by fitting a linear regression model adjusted for age, sex
386 and BMI (for blood metabolites only), using the tool PEACOK that was developed as a
387 modification of the R package PHESANT³⁴. For the pan-ancestry analysis, we additionally
388 included the categorical ancestral group and top five ancestry principal components as
389 covariates. For each of the 198 biomarkers, three different genetic models were evaluated in
390 the ExWAS: (i) genotypic (AA vs AB vs BB), (ii) dominant (AA+AB vs BB), and (iii) recessive
391 (AA vs AB+BB), where A and B denote the reference and alternative alleles, respectively. A
392 significance cut-off of $p < 1 \times 10^{-8}$ was adopted for the ExWAS³⁵.

393

394 Gene-level collapsing analysis

395 In order to boost power to detect associations for rare variants (including private mutations)
396 having the same direction of effect, we adopted a collapsing framework to test the aggregate
397 effect of rare functional variants in a gene. Overall, 10 different collapsing models (9
398 dominant and one recessive) were implemented per gene to evaluate a range of genetic
399 architectures. Additionally, a synonymous collapsing model was used for the purpose of
400 establishing an empirical negative control²⁰.

401 As outlined in **Table S11**, the criteria for qualifying variants (QVs)³⁶ for the collapsing
402 models were based on the following parameters: type of variant (missense, non-
403 synonymous or PTV), minor allele frequency, *in silico* deleteriousness predictors (REVEL
404 and MTR), and type of genetic model (dominant or recessive). The following variant
405 annotations were used to define PTVs: *exon_loss_variant*, *frameshift_variant*, *start_lost*,
406 *stop_gained*, *stop_lost*, *splice_acceptor_variant*, *splice_donor_variant*, *gene_fusion*,
407 *bidirectional_gene_fusion*, *rare_amino_acid_variant* and *transcript_ablation*. Hemizygous
408 genotypes for the X chromosome also qualified for the recessive model.

409 For a given collapsing model, the effect of QVs in each gene (N=18,762) was
410 calculated as the difference in the mean of a blood biomarker between carriers and non-

411 carriers of the QVs, using a linear regression model in PEACOCK. Covariates used in the
412 linear regression model were identical to that described for the ExWAS.

413 A significance cut-off of $p < 1 \times 10^{-8}$ was set for the collapsing analysis based on the
414 observed p-value distribution for the synonymous model and an n-of-1 permutation, as
415 described previously²⁰.

416

417 **Association analysis of clinical phenotypes documented in the UKB**

418 We harmonized and union mapped the clinical phenotypes available in the UKB, as
419 previously described²⁰. Phenome-wide collapsing analysis for 15,719 clinical phenotypes
420 was performed for the 11 collapsing models, as described in our previously published
421 study²⁰. We queried the results of this analysis for genes of interest that emerged from the
422 analysis of the metabolic biomarkers.

423 Additionally, we also performed an association analysis between the each of the 198
424 metabolic biomarkers and the clinical phenotypes using a linear regression model adjusted
425 for age and sex.

426

427 **Comparison of results from collapsing analyses to microarray-based genome-wide** 428 **association study**

429 We explored the hypothesis that the application of a collapsing framework – that tests the
430 aggregate effect of rare functional variants in a gene identified using exome sequencing –
431 detected gene-biomarker relationships that were previously not identified in microarray-
432 based studies. In order to do that, we compared our findings with the results from a recent
433 study¹⁴ that conducted single variant association analysis (GWAS) for clinical biomarkers in
434 the UKB using microarray data, including directly genotyped coding variants. Besides the
435 28/30 clinical blood biomarkers that we studied, seven other biomarkers (mainly, urine-
436 related) were analysed in the GWAS. These seven biomarkers comprised of four urinary
437 biomarkers that were directly measured in the UKB and an additional three derived
438 measurements. For the purpose of comparing findings, we additionally performed gene-level
439 collapsing analysis for the four urinary biomarkers for which data were directly available in

440 the UKB (i.e. 'sodium in urine', 'potassium in urine', 'microalbumin in urine', and 'creatinine
441 (enzymatic) in urine'). To be consistent with the microarray-based GWAS, we used the
442 statin-adjusted values for LDL-cholesterol, total cholesterol, and apolipoprotein B, and the
443 medication-unadjusted values for the remaining biomarkers. Thereafter, for the set of 32
444 biomarkers (28 blood and 4 urinary biomarkers) common to both studies, we compared
445 gene-biomarker relationships that achieved significance ($p < 1 \times 10^{-8}$) in the collapsing analysis
446 with gene-biomarker relationships corresponding to the significant coding variant
447 associations reported in the GWAS. We considered a comparatively relaxed significance
448 threshold of $p = 1 \times 10^{-7}$ for the GWAS results in order to be stringent when attributing a gene-
449 biomarker relationship as being specific to the collapsing analysis.

450 We also hypothesised that the various variant-level "purifying" filters implemented for
451 QV selection in the collapsing analysis can enable a more direct estimate for the effect of
452 gene aberrations (e.g., PTVs) on biomarker levels. To investigate this hypothesis, we
453 compared the effect sizes for gene-biomarker relationships that achieved significance in both
454 the gene-level collapsing analysis and the microarray-based GWAS. For each such gene-
455 biomarker relationship, we selected: (i) the *model* with the highest absolute beta in the
456 collapsing analysis, and (ii) the individual *variant* with the highest absolute beta as reported
457 in the Sinnott-Armstrong *et al* GWAS¹⁴. For the latter, we adopted the absolute beta
458 estimated in the genotypic model in our ExWAS (for the corresponding gene-biomarker
459 relationship) as a substitute, to account for possible differences in trait transformation,
460 association model or covariates between our study and the Sinnott-Armstrong *et al* GWAS.
461 Nonetheless, the absolute betas were highly correlated between the Sinnott-Armstrong *et al*
462 GWAS and our ExWAS (Spearman's $\rho = 0.99$) (**Figure S5**). We then compared the
463 absolute beta of the collapsing model [step (i)] with that of the individual variant [step (ii)].
464 This approach provides a means to compare the effect size of aberrations in genes on
465 biomarker levels estimated from individual coding variants captured by microarrays with that
466 estimated from an aggregate of rare coding variants identified using exome sequencing.
467

468 **Estimating gene similarity based on association signatures from collapsing analysis**

469 We aimed to leverage the rich catalogue of gene-level association statistics from the
470 collapsing analysis – ascertained for the set of studied metabolic biomarkers and under
471 different QV models – to identify genes that possess similar metabolic biomarker fingerprint
472 as a (target) gene of interest. Such a ‘gene similarity’ metric can provide opportunities to not
473 only expand our understanding of gene networks, but also offer alternative candidates in
474 cases of difficult-to-drug targets. Gene-SCOUT (Gene Similarity from Continuous Traits)²⁴,
475 the tool that we developed for this purpose, can also estimate “similarity” between genes
476 based on any set of quantitative traits of interest.

477 Rather than calculating similarities between genes directly, Gene-SCOUT estimates
478 distances between genes, which it then uses as a proxy for their similarity. Based on that,
479 the set of genes having the smallest distance from a given seed gene represent those that
480 are most ‘similar’ to it. We applied the cosine distance method – which is commonly used in
481 natural language processing³⁷ – to calculate distances between genes³⁸ based on their
482 effects on the metabolic biomarkers (referred to as ‘features’) estimated in the collapsing
483 analysis. In order to minimise the impact of stochastic effects on the gene similarity
484 estimations, for a given seed gene of interest, only those features that the genes is
485 associated with at $p < 1 \times 10^{-5}$ are selected (‘feature selection’ step), guided by sensitivity
486 analyses performed for a range of p-value thresholds²⁴. Thus, distances from genes having
487 $p > 1 \times 10^{-5}$ for all features in common with the seed gene are not considered.

488 The feature set used to generate the Gene-SCOUT results comprised of the 198
489 metabolic blood biomarkers. Though there is a degree of correlation in our feature set
490 (**Figure 2a**), we have demonstrated through simulations that correlation between features
491 has minimal impact on gene similarity estimations²⁴.

492 To illustrate the tool’s utility, we generated a network figure showing the genes that were
493 most similar to each of the 24 genes that were significantly associated with LDL-cholesterol
494 in the collapsing analysis.

495 **Ethics Reporting**

496 The protocols for UKB are overseen by The UK Biobank Ethics Advisory Committee (EAC);
497 for more information see: <https://www.ukbiobank.ac.uk/ethics/> and
498 <https://www.ukbiobank.ac.uk/wp-content/uploads/2011/05/EGF20082.pdf>.

499
500 **Acknowledgements**

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

508
509 **Author Contributions**

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

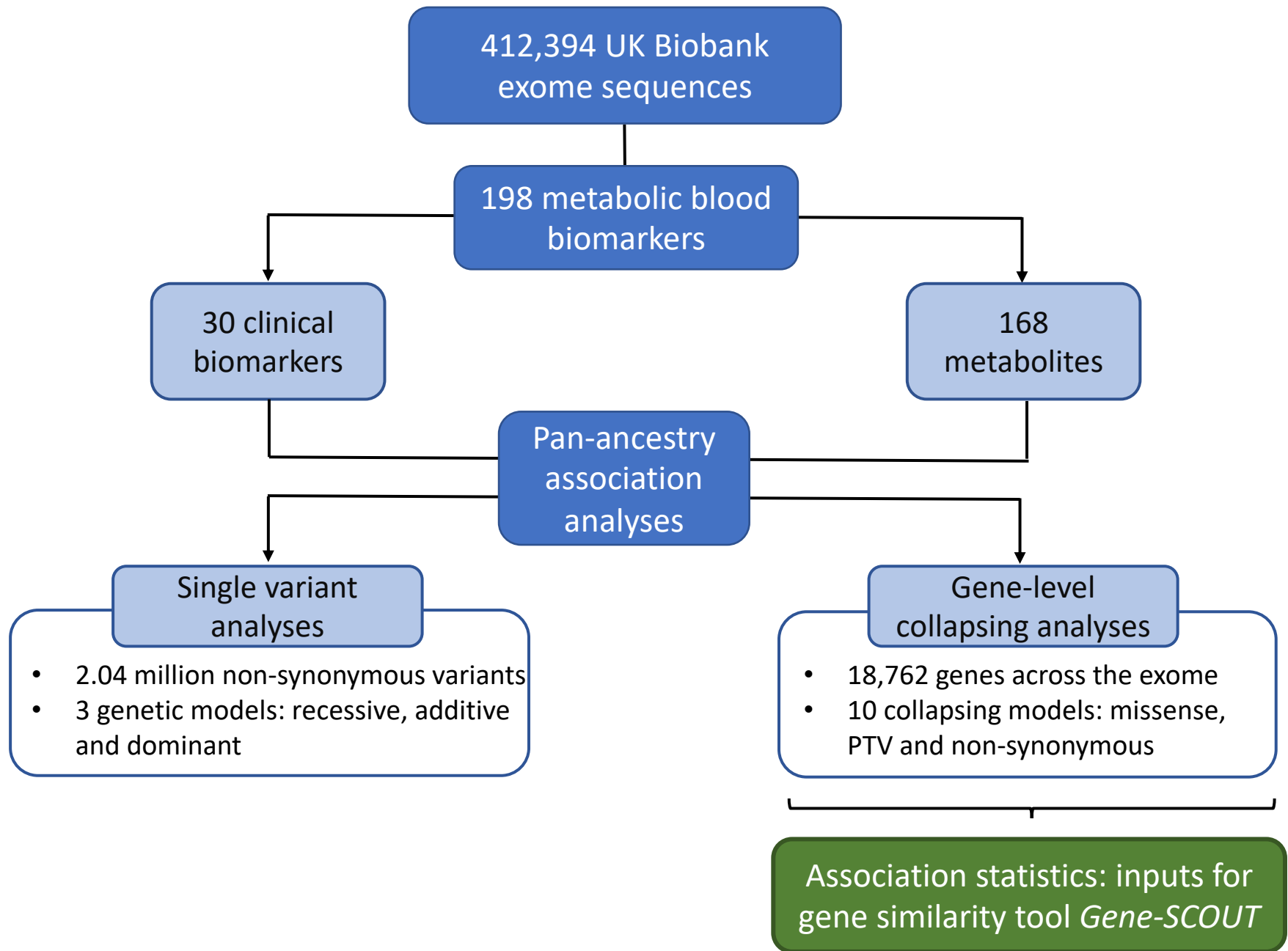
513
514 **Competing interests**

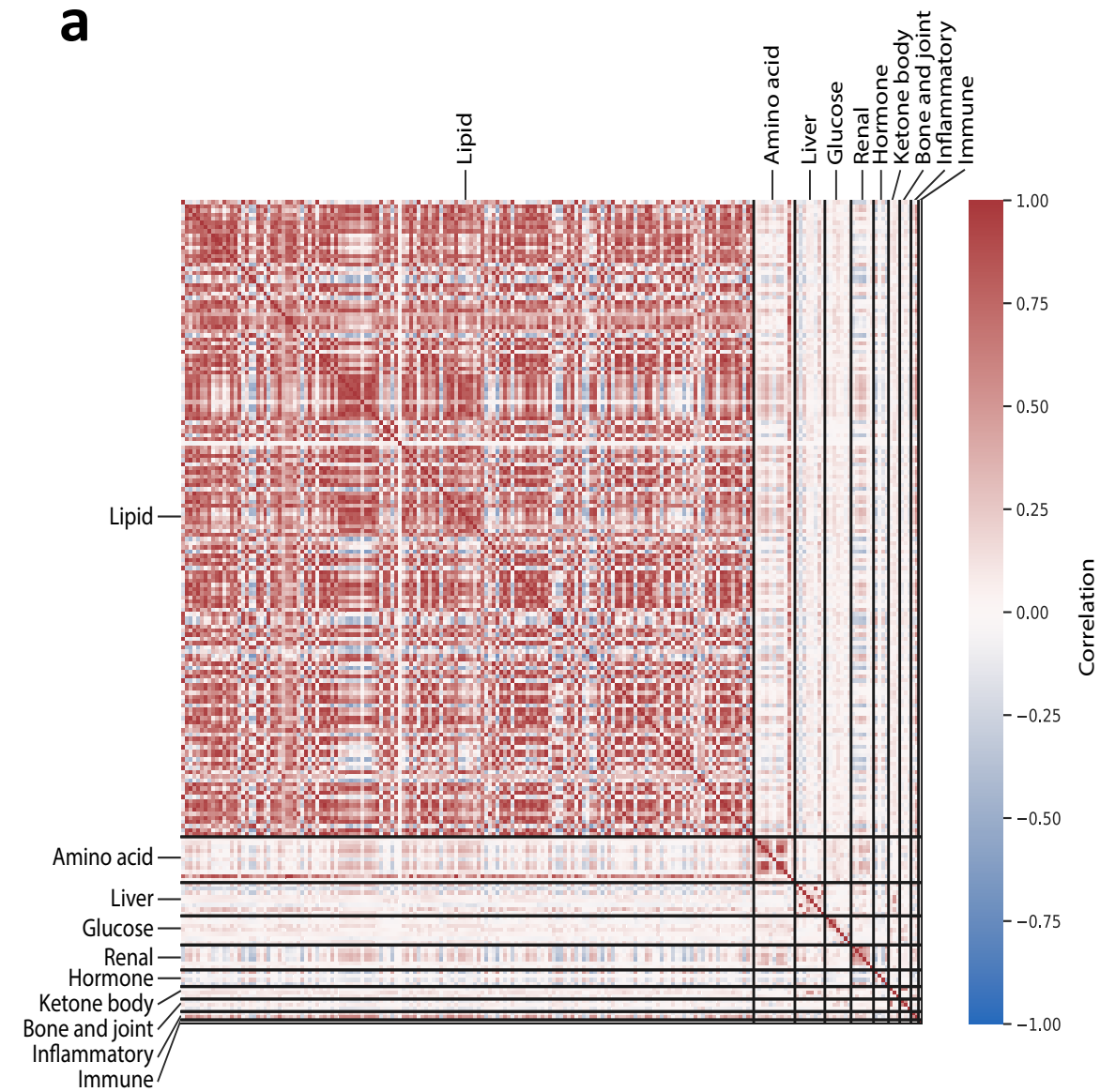
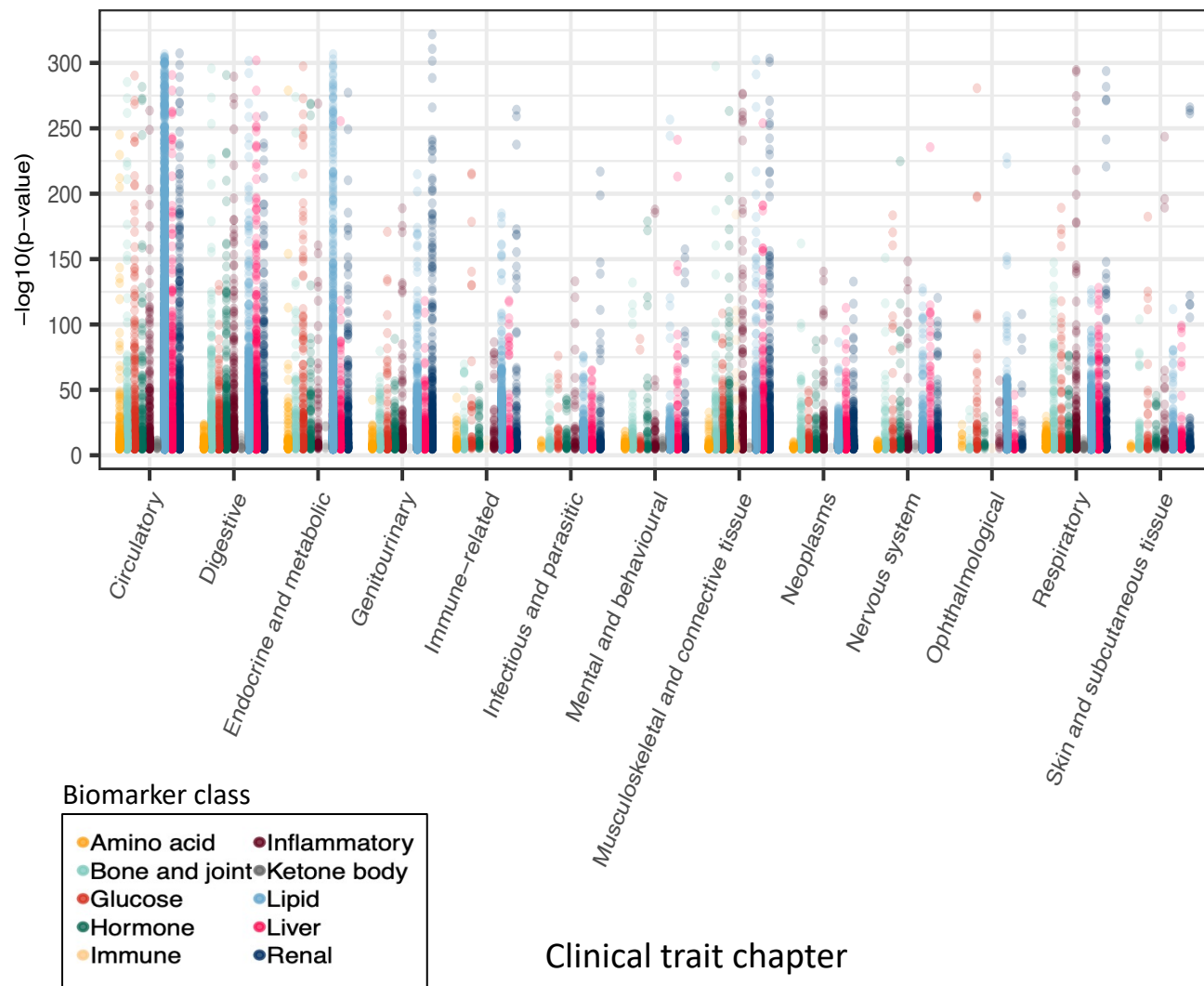
515 A.N., L.M., R.S.D., D.V., E.W., E.L.A., A.R., K.C., K.R.S., Q.W., B.C., D.S.P., A.R.H. and
516 S.P. are current employees and/or stockholders of AstraZeneca.

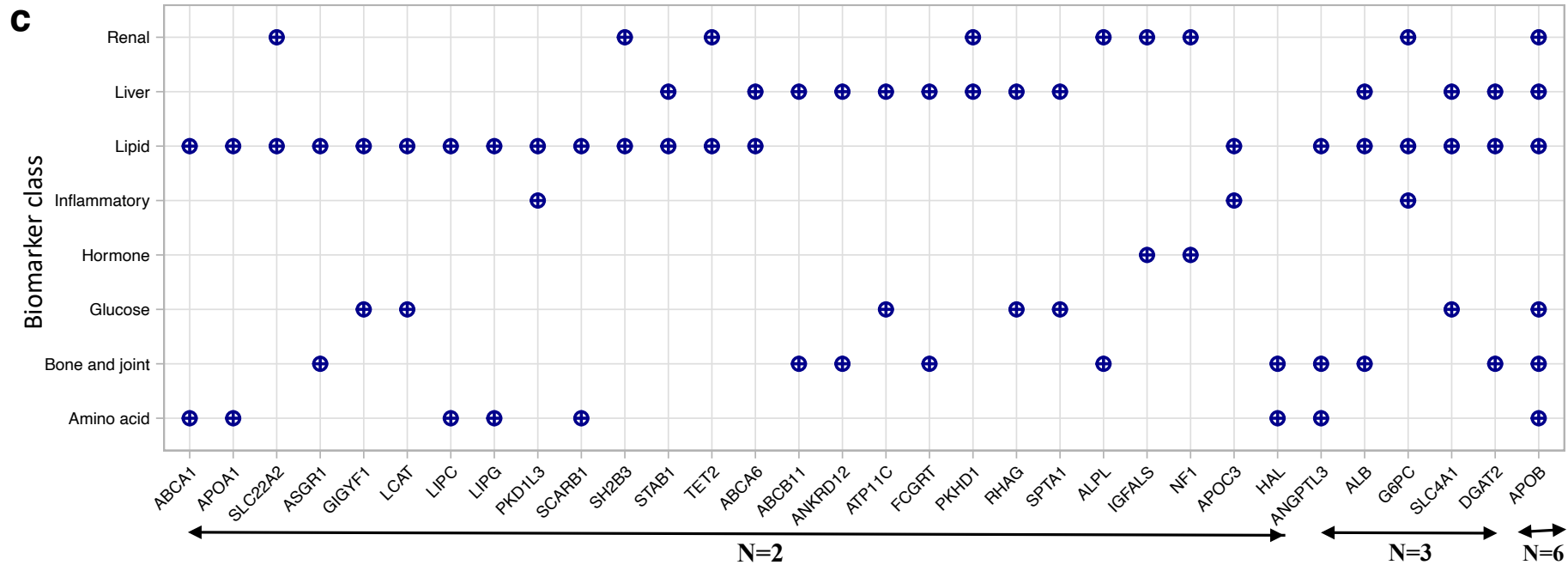
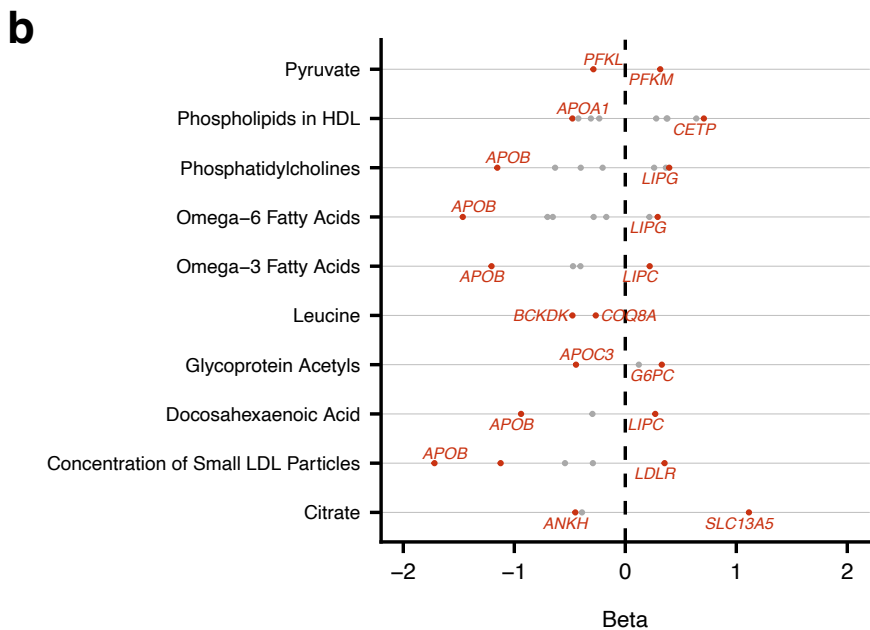
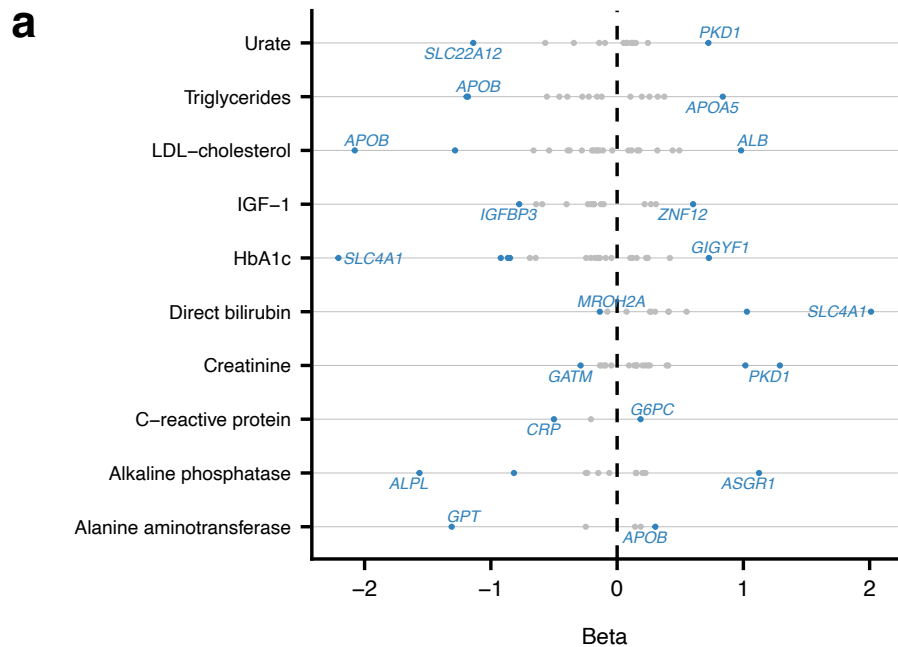
References

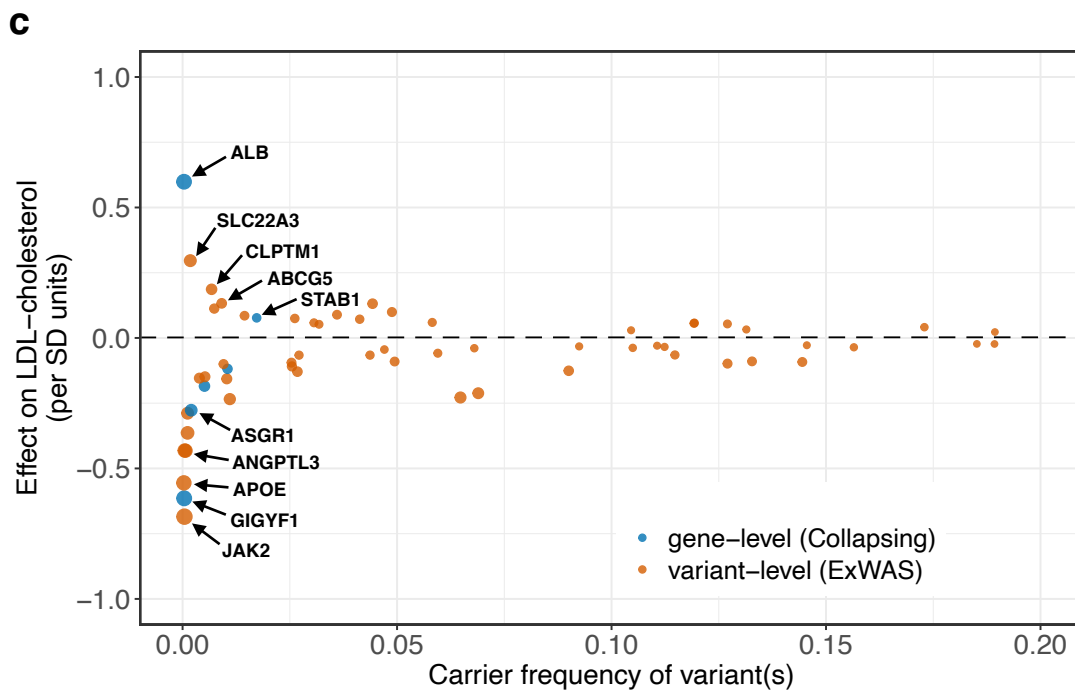
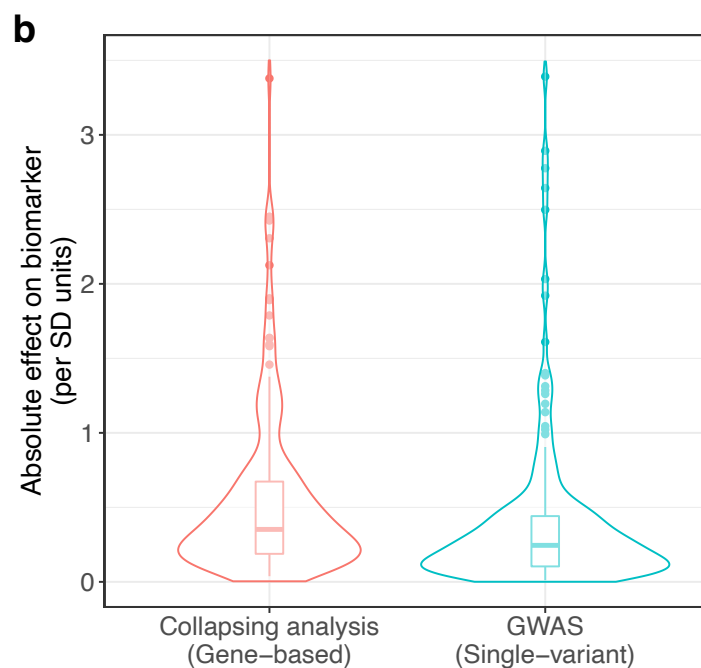
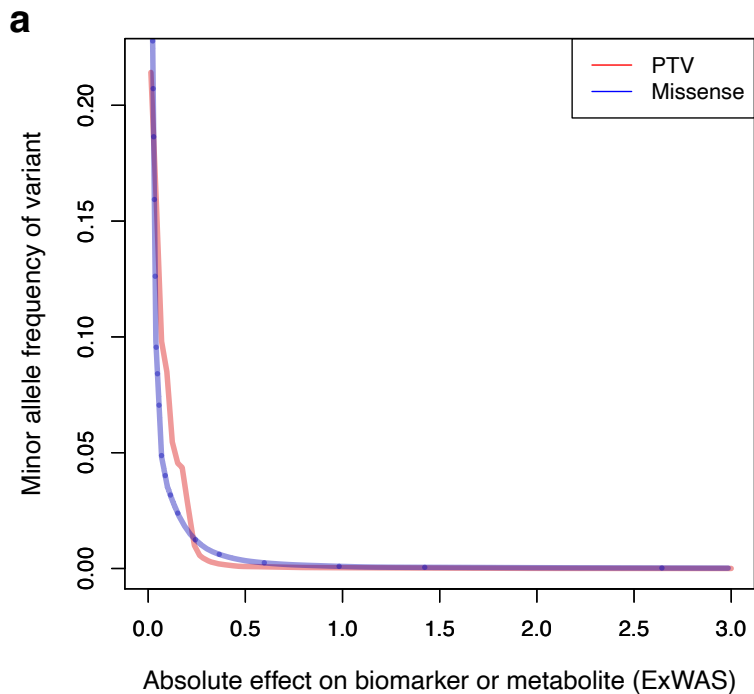
- 517 1. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature*
518 *genetics* **45**, (2013).
- 519 2. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from
520 analyses of a million individuals. *Nature genetics* **51**, (2019).
- 521 3. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing
522 human serum metabolite levels. *Nature genetics* **44**, (2012).
- 523 4. Yet, I. *et al.* Genetic Influences on Metabolite Levels: A Comparison across
524 Metabolomic Platforms. *PLoS one* **11**, (2016).
- 525 5. Suhre, K. *et al.* A genome-wide association study of metabolic traits in human urine.
526 *Nature genetics* **43**, (2011).
- 527 6. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nature*
528 *genetics* **46**, (2014).
- 529 7. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of
530 the Million Veteran Program. *Nature genetics* **50**, (2018).
- 531 8. Chambers, J. C. *et al.* Genome-wide association study identifies loci influencing
532 concentrations of liver enzymes in plasma. *Nature genetics* **43**, (2011).
- 533 9. Prins, B. P. *et al.* Genome-wide analysis of health-related biomarkers in the UK
534 Household Longitudinal Study reveals novel associations. *Scientific reports* **7**, (2017).
- 535 10. Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type
536 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic
537 genome-wide meta-analysis. *PLoS medicine* **14**, (2017).
- 538 11. Long, T. *et al.* Whole-genome sequencing identifies common-to-rare variants
539 associated with human blood metabolites. *Nature genetics* **49**, (2017).
- 540 12. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to
541 Function. *American journal of human genetics* **102**, (2018).
- 542 13. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
543 *Nature* **562**, 203–209 (2018).
- 544 14. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK
545 Biobank. *Nature genetics* **53**, (2021).
- 546 15. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and
547 disease. *Nature* **526**, (2015).
- 548 16. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human
549 protein-coding genes. *Science (New York, N.Y.)* **335**, (2012).
- 550 17. Ritchie, S. C. *et al.* Quality control and removal of technical variation of NMR
551 metabolic biomarker data in ~120,000 UK Biobank participants. *medRxiv*
552 2021.09.24.21264079 (2021) doi:10.1101/2021.09.24.21264079.
- 553 18. Gandotra, S. *et al.* Perilipin deficiency and autosomal dominant partial lipodystrophy.
554 *The New England journal of medicine* **364**, (2011).
- 555 19. Shin, G.-C., Kang, H. S., Lee, A. R. & Kim, K.-H. Hepatitis B virus-triggered
556 autophagy targets TNFRSF10B/death receptor 5 for degradation to limit
557 TNFSF10/TRAIL response. *Autophagy* **12**, (2016).
- 558 20. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank
559 exomes. *Nature* (2021) doi:10.1038/s41586-021-03855-y.
- 560 21. Aimee M. Deaton *et al.* Gene-level analysis of rare variants in 379,066 whole exome
561 sequences identifies an association of GIGYF1 loss of function with type 2 diabetes.
562 *Scientific Reports* **11**, (2021).
- 563 22. Jurgens, S. J. *et al.* Rare Genetic Variation Underlying Human Diseases and Traits:
564 Results from 200,000 Individuals in the UK Biobank. *bioRxiv* 2020.11.29.402495
565 (2020) doi:10.1101/2020.11.29.402495.
- 566 23. Chen, J. *et al.* The trans-ancestral genomic architecture of glycemc traits. *Nature*
567 *genetics* **53**, (2021).

- 568 24. Lawrence Middleton *et al.* Gene-SCOUT: identifying genes with similar continuous
569 trait fingerprints from phenome-wide association analyses. *Nucleic Acids Res (in*
570 *submission)* (2021).
- 571 25. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbs, H. H. Sequence variations in
572 PCSK9, low LDL, and protection against coronary heart disease. *The New England*
573 *journal of medicine* **354**, (2006).
- 574 26. Abul-Husn, N. S. *et al.* A Protein-Truncating HSD17B13 Variant and Protection from
575 Chronic Liver Disease. *The New England journal of medicine* **378**, (2018).
- 576 27. Akbari, P. *et al.* Sequencing of 640,000 exomes identifies GPR75 variants associated
577 with protection from obesity. *Science (New York, N.Y.)* **373**, (2021).
- 578 28. Nag, A. *et al.* Human genetic evidence supports MAP3K15 inhibition as a therapeutic
579 strategy for diabetes. *medRxiv* 2021.11.14.21266328 (2021)
580 doi:10.1101/2021.11.14.21266328.
- 581 29. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery
582 through exome sequencing of the UK Biobank. *Nature genetics* **53**, (2021).
- 583 30. Cingolani, P. *et al.* A program for annotating and predicting the effects of single
584 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*
585 strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
- 586 31. Traynelis, J. *et al.* Optimizing genomic medicine in epilepsy through a gene-
587 customized approach to missense variant interpretation. *Genome research* **27**, 1715–
588 1729 (2017).
- 589 32. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity
590 of Rare Missense Variants. *American journal of human genetics* **99**, 877–885 (2016).
- 591 33. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association
592 studies. *Bioinformatics* **26**, (2010).
- 593 34. Millard, L. A. C., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software
594 Application Profile: PHESANT: a tool for performing automated phenome scans in UK
595 Biobank. *International journal of epidemiology* **47**, (2018).
- 596 35. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value
597 threshold revisited and updated for low-frequency variants. *European journal of*
598 *human genetics* □: *EJHG* **24**, (2016).
- 599 36. Petrovski, S. *et al.* An Exome Sequencing Study to Assess the Role of Rare Genetic
600 Variation in Pulmonary Fibrosis. *American journal of respiratory and critical care*
601 *medicine* **196**, (2017).
- 602 37. Huang A. Similarity Measures for Text Document Clustering. *NZCSRSC* (2008).
- 603 38. Kittipong Chomboon, Pasapitch Chujai, Pongsakorn Teerarassamee, Kittisak
604 Kerdprasop & Nittaya Kerdprasop. An Empirical Study of Distance Metrics for k-
605 Nearest Neighbor Algorithm. *Proceedings of the 3rd International Conference on*
606 *Industrial Application Engineering* (2015).
- 607



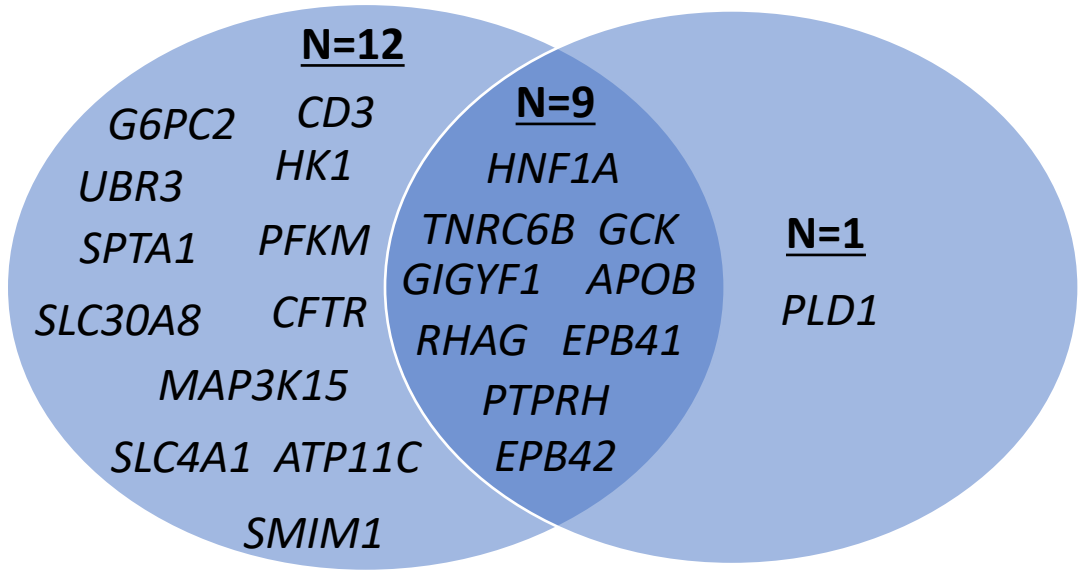
a**b**

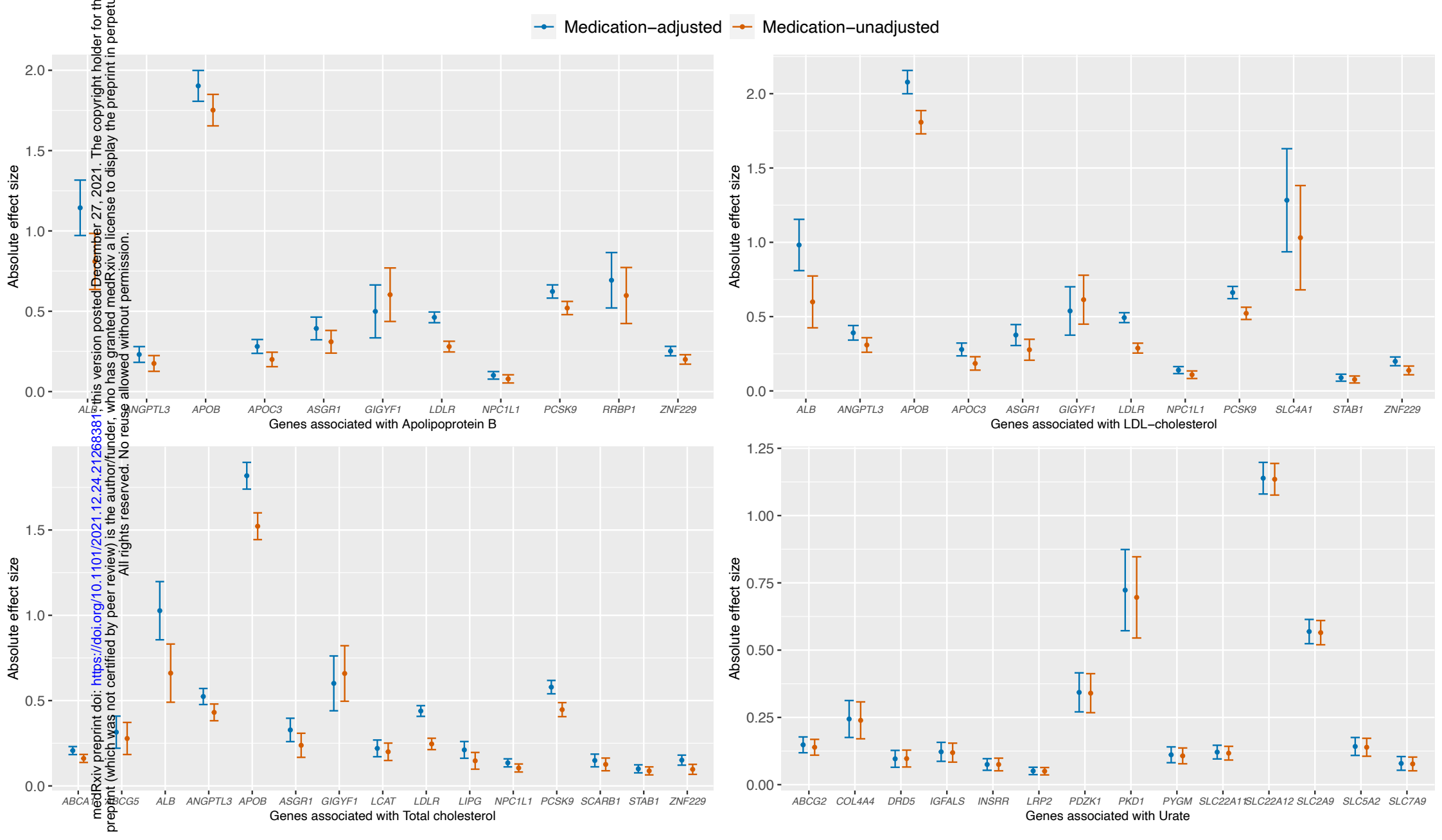




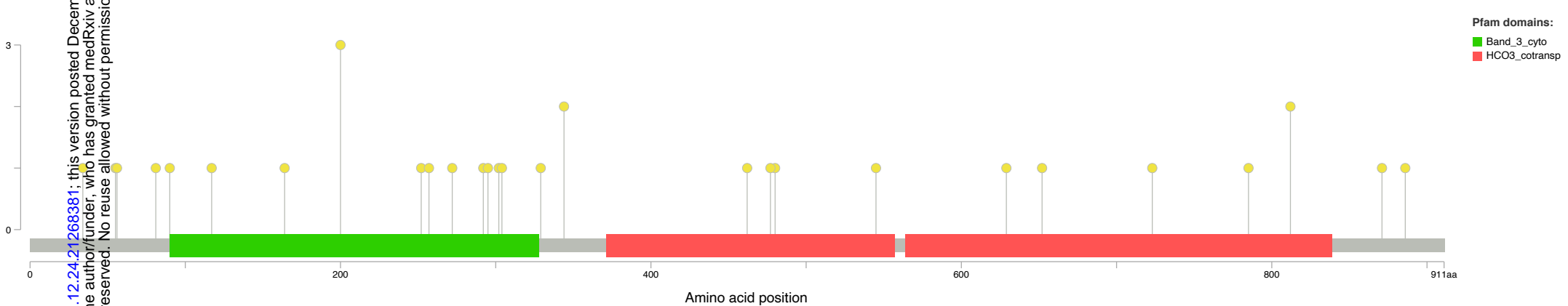
Current study

Deaton *et al*





SLC4A1 Mutations



454,796 UK Biobank exome sequences

DNA contamination
VerifyBAMID freemix >4%

(N=122)

Coverage depth
10x for <94.5% of the CCDS

(N=5)

Relatedness
2nd-degree relatives or closer
(kinship coefficient >0.0884)

(N=32,181)

Genetic ancestry
(i) Predicted ancestry <0.99
(ii) >4 SDs for top four PCs for the genetic ancestry

Europeans
(N=394,695)

East Asians
(N=2,209)

South Asians
(N=8,078)

Africans
(N=7,412)

