

1 **Cross-platform genetic discovery of small molecule products of metabolism and application to**
2 **clinical outcomes**

3
4 Luca A. Lotta^{1#}, Maik Pietzner^{1#}, Isobel D. Stewart¹, Laura B.L. Wittemans^{1,2}, Chen Li¹, Roberto
5 Bonelli^{3,4}, Johannes Raffler⁵, Emma K. Biggs⁶, Clare Oliver-Williams^{7,8}, Victoria P.W. Auyeung¹, Jian'an
6 Luan¹, Eleanor Wheeler¹, Ellie Paige⁹, Praveen Surendran^{7,10,11,12}, Gregory A. Michelotti¹³, Robert A.
7 Scott¹, Stephen Burgess^{14,15}, Verena Zuber^{14,16}, Eleanor Sanderson¹⁷, Albert Koulman^{1,5,18}, Fumiaki
8 Imamura¹, Nita G. Forouhi¹, Kay-Tee Khaw¹⁵, MacTel Consortium, Julian L. Griffin¹⁹, Angela M.
9 Wood^{7,10,11,20,21}, Gabi Kastenmüller⁵, John Danesh^{7,10,11,20,22,23}, Adam S. Butterworth^{7,10,11,20,22,23}, Fiona
10 M. Gribble⁶, Frank Reimann⁶, Melanie Bahlo^{3,4}, Eric Fauman²⁴, Nicholas J. Wareham¹, Claudia
11 Langenberg^{1,11*}

12
13 *# these authors contributed equally*

- 14
15 1) MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
16 2) The Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford
17 3) Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research,
18 Parkville, Australia
19 4) Department of Medical Biology, The University of Melbourne, Parkville, Australia
20 5) Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for
21 Environmental Health, Neuherberg, Germany
22 6) Metabolic Research Laboratories, University of Cambridge, Cambridge, United Kingdom
23 7) British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary
24 Care, University of Cambridge, Cambridge, UK
25 8) Homerton College, University of Cambridge, Cambridge, UK
26 9) National Centre for Epidemiology and Population Health, The Australian National University, Canberra,
27 Australia
28 10) British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK
29 11) Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge,
30 Cambridge, UK
31 12) Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, UK
32 13) Metabolon Inc, Durham, North Carolina USA
33 14) MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom
34 15) Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom
35 16) Department of Epidemiology and Biostatistics, Imperial College London, UK
36 17) MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, UK
37 18) NIHR BRC Nutritional Biomarker Laboratory, University of Cambridge, UK
38 19) Biomolecular Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College
39 London, UK
40 20) National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics,
41 University of Cambridge, Cambridge, UK
42 21) The Alan Turing Institute, London, UK
43 22) National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge
44 and Cambridge University Hospitals, Cambridge, UK
45 23) Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK
46 24) Internal Medicine Research Unit, Pfizer Worldwide Research, Cambridge, MA 02142, USA

47
48 *Corresponding author:

49 Claudia Langenberg
50 MRC Epidemiology Unit
51 University of Cambridge School of Clinical Medicine
52 Institute of Metabolic Science
53 Cambridge, UK
54 claudia.langenberg@mrc-epid.cam.ac.uk

55 **Abstract**

56 Circulating levels of small molecules or metabolites are highly heritable, but the impact of
57 genetic differences in metabolism on human health is not well understood. In this cross-platform,
58 genome-wide meta-analysis of 174 metabolite levels across six cohorts including up to 86,507
59 participants (70% unpublished data), we identify 499 (362 novel) genome-wide significant
60 associations ($p < 4.9 \times 10^{-10}$) at 144 (94 novel) genomic regions. We show that inheritance of blood
61 metabolite levels in the general population is characterized by pleiotropy, allelic heterogeneity, rare
62 and common variants with large effects, non-linear associations, and enrichment for
63 nonsynonymous variation in transporter and enzyme encoding genes. The majority of identified
64 genes are known to be involved in biochemical processes regulating metabolite levels and to cause
65 monogenic inborn errors of metabolism linked to specific metabolites, such as *ASNS* (rs17345286,
66 MAF=0.27) and asparagine levels. We illustrate the influence of metabolite-associated variants on
67 human health including a shared signal at *GLP2R* (p.Asp470Asn) associated with higher citrulline
68 levels, body mass index, fasting glucose-dependent insulinotropic peptide and type 2 diabetes risk,
69 and demonstrate beta-arrestin signalling as the underlying mechanism in cellular models. We link
70 genetically-higher serine levels to a 95% reduction in the likelihood of developing macular
71 telangiectasia type 2 [odds ratio (95% confidence interval) per standard deviation higher levels 0.05
72 (0.03-0.08; $p = 9.5 \times 10^{-30}$)]. We further demonstrate the predictive value of genetic variants identified
73 for serine or glycine levels for this rare and difficult to diagnose degenerative retinal disease [area
74 under the receiver operating characteristic curve: 0.73 (95% confidence interval: 0.70-0.75)], for
75 which low serine availability, through generation of deoxysphingolipids, has recently been shown to
76 be causally relevant. These results show that integration of human genomic variation with
77 circulating small molecule data obtained across different measurement platforms enables efficient
78 discovery of genetic regulators of human metabolism and translation into clinical insights.

79

80 Introduction

81 Metabolites are small molecules that reflect biological processes and are widely measured in
82 clinical medicine as diagnostic, prognostic or treatment response biomarkers¹. Blood levels of
83 metabolites are highly heritable with twin studies reporting a median explained variance in plasma
84 levels of 6.9% and maximum of 50% depending on the metabolite^{2,3}. Several earlier studies have
85 started to characterise the genetic architecture of metabolite variation in the general population²⁻¹⁰,
86 but been limited in size and scope by focussing on metabolites assessed using a single method.
87 Integration of genetic association results for metabolites measured on different platforms can help
88 maximise the power for a given metabolite and provide a more refined understanding of genetic
89 influences on blood metabolite levels and human physiology.

90 To identify genomic regions regulating metabolite levels and systematically study their
91 relevance for disease, we designed and conducted a cross-platform meta-analysis of genetic effects
92 on levels of 174 blood metabolites measured in large-scale population-based studies. We included
93 metabolites covered by the targeted Biocrates AbsoluteIDQ™ p180 platform and measured in the
94 Fenland Study. We integrated unpublished data for any of these metabolites that were covered by
95 the Nightingale (¹H-NMR, Interval Study) or Metabolon (Discovery HD4™, EPIC-Norfolk and Interval
96 Studies) platforms, or had previously been reported^{2,4,5}. The focus on this targeted set of 'platform-
97 specific' metabolites enabled us to clearly map metabolites across platforms and maximise the
98 sample size for each of the 174 metabolites for this proof of concept cross-platform GWAS study. To
99 facilitate rapid sharing of our results, we developed a webserver
100 (<https://omicscience.org/apps/crossplatform/>) that allows flexible interrogation of our results.

101 Results

102 *Associations with blood metabolites at 144 genomic regions*

103 Genome-wide meta-analyses were conducted for 174 metabolites from 7 biochemical classes
104 (i.e. amino acids, biogenic amines, acylcarnitines, lyso-phosphatidylcholines, phosphatidylcholines,
105 sphingomyelins and hexose) commonly measured using the Biocrates p180 kit in up to 86,507
106 individuals, contributing over 3.7 million individual-metabolite data points (70% from unpublished
107 studies; **Fig. 1**). For each of the 174 metabolites, this was the largest genome-wide association
108 analyses (GWAS) to date, with at least a doubling of sample size (**Fig. 1A**). Sample sizes ranged from
109 8,569 to 86,507 individuals for metabolites depending on the platform used in each contributing
110 study. Using GWAS analyses we estimated the association of up to 10.2 million single nucleotide
111 variants with a minor allele frequency (MAF) >0.5%, including 6.1 million with MAF ≥ 5%.

112 We identified 499 variant-metabolite associations (362 novel) from 144 loci (94 novel) at a
113 metabolome-adjusted genome-wide significance threshold of $p < 4.9 \times 10^{-10}$ (correcting the usual
114 GWAS-threshold, $p < 5 \times 10^{-8}$, for 102 **principal** components explaining 95% of the variance in
115 metabolite levels using **principal** component analysis; **Fig. 1**). The vast majority of these associations
116 were consistent across studies and measurement platforms [median I^2 : 26.8 (interquartile range: 0 –
117 70.1) for 465 associations with at least two contributing studies] (**Supplementary Tab. S1-2**). To
118 identify possible sources of heterogeneity, we investigated the influence of differences by cohort,
119 measurement platform, metabolite class, and association strength in a joint meta-regression model
120 (**Supplementary Tab. S3**). This showed that heterogeneity was mainly due to the overall strength of
121 the signal, i.e. associations with higher z-scores showed greater heterogeneity ($p < 1.05 \times 10^{-9}$).
122 However, the majority of these statistically heterogeneous associations were directionally consistent
123 and nominally significant across and within each stratum for 146 of 170 associations with a z-score $>$
124 10, demonstrating the feasibility of pooling association estimates across metabolomics platforms for
125 the purpose of genetic discovery. Genetic variants at the *NLRP12* locus, e.g. rs4632248, were a
126 notable exception with large estimates of heterogeneity ($I^2 > 90\%$). The *NLRP12* locus is known to
127 affect the monocyte count¹¹ and has been shown to have pleiotropic effects on the plasma
128 proteome in the INTERVAL study¹². Monocytes, or at least a subpopulation subsumed under this cell
129 count measure, release a wide variety of biomolecules upon activation or may die during the sample
130 handling process and hence releasing intracellular biomolecules, such as taurine¹³, into the plasma.
131 In brief, one specific source of heterogeneity in mGWAS associations might relate to sample
132 handling differences across studies.

133 This highlights the utility of our genetic cross-platform approach to maximise power for a given
134 metabolite, substantially extending previous efforts for any given metabolite¹⁴. Previously reported
135 associations from platform-specific studies were also found to generally be consistent in our cross-
136 platform meta-analysis (**Supplementary Tab. S2**; <https://omicscience.org/apps/crossplatform/>).

137 *Insights in the genetic architecture of metabolite levels*

138 We identified a median of 2 (range: 1-67, **Fig. 2A**) associated metabolites for each locus and a
139 median of 3 (range: 1-20, **Fig. 2B**) locus associations for each metabolite, reflecting pleiotropy and
140 the extensive contribution of genetic loci to circulating metabolite levels. The number of associations
141 was proportional to the estimated heritability and the sample size of the meta-analysis for a given
142 trait (**Fig. 2C**).

143 We applied a multi-trait statistical colocalisation method¹⁵ and identified between 1-30
144 (median: 2) metabolites that did not meet the discovery p-value threshold, but showed high

145 posterior probability (>75%) of a shared genetic signal for 49 out of the 144 loci (**Supplemental Fig.**
146 **S1**). Two distinct variants (rs2414577 and rs261334) nearby *LIPC* showed the largest gain in
147 additionally associated metabolites, in line with previous reports of extensive pleiotropy and allelic
148 heterogeneity at this locus⁹. We note that a low posterior probability for the alignment of multiple
149 metabolites at other loci might be explained by the presence of multiple causal variants shared
150 across multiple metabolites.

151 To systematically classify pleiotropic variants taking into account the correlation structure
152 among metabolites we derived a data-driven metabolic network and performed community
153 detection (see **Methods** and **Supplemental Fig. S2**). A total of 129 (60.5%) of 214 variants
154 (associated with at least two metabolites at $p < 5 \times 10^{-8}$) were associated with metabolites from at
155 least two of the 14 communities (range: 2 – 11; **Supplemental Fig. S2**), i.e. showed evidence for
156 ‘horizontal’ or broad pleiotropy. The most extreme variants included those near *FADS1* (e.g.
157 rs17455) associated with 61 metabolites across 11 communities at $p < 5 \times 10^{-8}$. In contrast, rs2638315
158 (likely tagging a missense variant rs2657879 at *GLS2*) was associated with nine metabolites within a
159 single community and would therefore be considered as ‘horizontal pleiotropic’ for a well-defined
160 group of correlated metabolites (**Supplemental Fig. S2**).

161 Similar to what is routinely observed in GWAS literature, effect size estimates increased with
162 decreasing minor allele frequency (MAF) (**Fig. 3A**). However, there were 26 associations (**Tab. 1**) for
163 common lead variants with per-allele differences in metabolites levels greater than 0.25 standard
164 deviations (SD), a per-allele effect size that is >3-fold larger than the strongest common variants
165 associated with SDs of body mass index at the *FTO* locus.

166 Variants identified in this study explained up to 23% of the variance (median: 1.4%; interquartile
167 range: 0.5% - 2.8%) and up to 99.8% of the chip-based heritability (median 9.2%; interquartile range:
168 4.7% - 17.1%) for the 141 metabolites with at least one genetic association (**Fig. 2D**). The 26 common
169 variants with large effect sizes (>0.25 SD per allele) were identified for metabolites with higher
170 heritability (**Fig. 2D**) and accounted for up to 74% of the heritability explained in those metabolites.

171 GWAS analyses generally assume a linear relationship between genotypes and phenotypes, i.e.
172 an additive dose-response model. The identification of several metabolite-associated variants with
173 large effect sizes and availability of individual-level data in the Fenland cohort allowed us to test
174 whether the metabolite-associated variants showed evidence of deviation from a linear model. Of
175 499 associations tested, 9 showed evidence of departure from a linear association (**Fig. 2E-M**).
176 Modelling actual genotypes rather than assuming ‘additive’ linear associations in these instances
177 explained a median of 7.4% more (range: 1.4-15.2%) of the heritability in metabolite levels (**Fig. 2N**).

178 Associations better described by an autosomal recessive or dominant model of inheritance might be
179 the most likely explanation for this. Variant rs3916, for example, which showed a more than additive
180 positive effect on butyrylcarnitine, is in perfect LD with a missense variant within *ACADS*
181 (rs1799958, MAF=26%), which encodes for short-chain acyl-CoA dehydrogenase (SCAD). SCAD
182 deficiency is an autosomal recessive disease diagnosed by elevated butyrylcarnitine concentrations
183 in blood and homozygous carrier status for established pathogenic variants¹⁶.

184 In 61 of the 499 associations the lead association signal was a nonsynonymous variant, a 40-fold
185 enrichment compared to what would be expected by chance given the annotation of ascertained
186 genetic variants (two-tailed binomial test, $p=5\times 10^{-30}$, **Fig. 3D**). For a further 59 associations, the lead
187 variant was in high LD with a nonsynonymous variant ($r^2>0.8$). Lead variants that were
188 nonsynonymous, or variants in high LD with a nonsynonymous variant, generally had lower MAF,
189 larger effect sizes, and smaller 99%-credible sets (**Supplemental Tab. S4**) than variants that were not
190 in these categories (**Fig 3B-D**).

191 We identified 22 loci harbouring two ($n=21$) or three ($n=1$) independent signals, i.e. different
192 plasma metabolites were associated with distinct genetic variants within the same genomic region
193 (**Supplementary Tab. S2**). For six regions, our two different annotations approaches assigned only
194 one causal gene (see below and **Methods**), including *ACADM*, *GLDC*, *ARG1*, *MARCH8*, *SLC7A2*, and
195 *LIPC* (**Supplementary Tab. S2**). We found evidence that allelic heterogeneity, i.e. conditionally
196 independent variants at a locus for a specific metabolite, explains the association pattern at 3 of
197 those loci (*ACADM*, *ARG1*, and *LIPC*; **Supplementary Tab. S5**). We identified another 16 loci
198 harbouring at least one (range: 2–6) additional conditionally independent variant(s) in exact
199 conditional analyses (see **Methods**, **Supplementary Tab. S5**).

200 *Effector genes, tissues, pathways*

201 We used two complementary strategies to prioritize likely causal genes for the observed
202 associations: (1) a hypothesis-free genetic approach based on physical distance, genomic annotation
203 and integration of expression quantitative trait loci (eQTLs) to prioritize genes in a systematic and
204 standardised way (see **Methods**), and (2) a biological knowledge-based approach integrating existing
205 knowledge about specific metabolites or related pathways to identify biologically plausible
206 candidate genes from the 20 genes closest to the lead variant (**Fig. 4A**). Using the hypothesis-free
207 genetic approach, we identified 249 unique likely causal genes for the 499 associations, with at least
208 one gene per association and some genes prioritized as likely causal for multiple metabolite
209 associations. The knowledge-based approach identified 130 biologically plausible genes for 349 out
210 of 499 associations. We asked whether the hypothesis-free genetic approach identified biologically

211 plausible genes (prioritized by strategy 2) more often than expected by chance. Amongst 9,980
212 possible gene-metabolite pairs (20 genes x 499 associations), 420 (4.2%) were biologically plausible,
213 condensed to 350 gene(s)-metabolite assignments after accounting for overlapping annotations. Of
214 the latter, 126 pairs (36%) were identical to genetically-prioritized gene-metabolite pairs,
215 representing a significant enrichment of biologically plausible genes among those prioritised by the
216 hypothesis-free algorithm (~8-fold more than expected by chance; two-tailed binomial test,
217 $p=2.3\times 10^{-80}$; **Fig. 4B**). Among the consistently assigned genes between both approaches, assignment
218 of the nearest gene (124 times out of 126, X^2 -test, $p<2.5\times 10^{-45}$) was the strongest shared factor, as
219 might be expected, followed by being (or in LD with) a missense variant ($R^2>0.8$, 30 times out of 126,
220 X^2 -test, $p<1.3\times 10^{-07}$) and only a minor contribution of eQTL data (20 times out of 126, X^2 -test,
221 $p<0.001$). Over 70% of genetically prioritized genes were enzymes or transporters (**Fig. 4C**).
222 Inconsistencies between the approaches might be explained by non-consideration of information on
223 biological pathways in the hypothesis-free genetic approach, as well as variants acting more distal to
224 the biological determinants of plasma metabolite levels not being considered in the knowledge-
225 based approach. The missense variant rs1260326 within *GCKR*, for example, colocalised with 49
226 metabolites across diverse biochemical classes (**Supplemental Fig. S1**) and likely confers its effects on
227 glucose metabolism through impaired inhibition of glucokinase by glucokinase regulatory protein
228 and might hence be considered as putative causal candidate by the knowledge-driven approach for
229 plasma glucose only. However, impairments in glucose metabolism result in numerous downstream
230 consequences including more distal metabolic branches such as amino acid and lipid metabolism.

231 In addition to being enriched in genes previously implicated in the biology of these metabolites,
232 the genetically prioritized genes were also enriched in genes known for mutations to cause rare
233 inborn errors of metabolism (IEMs), i.e. monogenic defects in the metabolism of small molecules
234 with very specific metabolite changes (**Fig. 4B**).

235 Integrating GWAS statistics across cohorts and platforms allowed us to identify three genes that
236 have never been associated with any metabolite level so far. At the *CERS6* locus, rs4143279
237 associates with levels of sphingomyelin (d18:1/16:0) ($p = 4.2\times 10^{-10}$). *CERS6* encodes a ceramide
238 synthase facilitating formation of ceramide, a precursor of sphingomyelins¹⁷. At the *ASNS* locus,
239 rs17345286 associates with levels of asparagine ($p = 4.7\times 10^{-20}$). The lead variant is in high LD ($R^2=1$)
240 with a missense mutation in *ASNS* (rs1049674, p.Val210Glu). *ASNS* encodes an asparagine
241 synthase¹⁸. Finally, at the *SLC43A1* locus, rs2649667 associates with levels of phenylalanine ($p =$
242 3.6×10^{-13}). *SLC43A1* encodes a liver-enriched transporter of large neutral amino acids, including
243 phenylalanine¹⁹.

244 *Insights into the causes of common and rare diseases from metabolite-associated loci*

245 The phenotypic consequences of metabolite-associated variants are currently not well
246 characterized. Below, we investigate the contribution of individual loci and polygenic predisposition
247 associated with differences in metabolite levels to the risk of common and rare diseases.

248 *A citrulline-raising functional variant in GLP2R increases type 2 diabetes risk*

249 Because several of the metabolites captured in this GWAS have been associated with incident
250 type 2 diabetes (T2D), we sought to investigate whether the association between metabolite-
251 associated loci and diabetes could provide insights into underlying pathophysiologic mechanisms.
252 Using estimates of effect for association with T2D based on a meta-analysis of 80,983 cases and
253 842,909 controls (see **Methods**), we observed a significant enrichment for associations with type 2
254 diabetes ($p\text{-value}=2.8\times 10^{-7}$) of metabolite-associated variants compared to a matched control set of
255 variants (**Fig. 5A**).

256 Amongst the diabetes- and metabolite-associated loci was a missense p.Asp470Asn
257 (rs17681684) variant in the *GLP2R* gene encoding the receptor for glucagon-like peptide 2, a 33
258 amino acid peptide hormone encoded by the proglucagon gene (*GCG*) that stimulates the growth of
259 intestinal tissue. Common variants at *GLP2R* are associated with an increased risk of T2D²⁰. The
260 previously reported lead variant for T2D (rs78761021) is in high LD ($r^2>0.87$) with our lead citrulline
261 association signal at *GLP2R* (rs17681684), which was associated with a 4% higher type 2 diabetes risk
262 (per-allele odds ratio, 1.04; 95% confidence interval, 1.02, 1.05; $p=1.1\times 10^{-8}$), comparable to
263 previous reports²⁰. Considering eleven phenotypes related to glucose homeostasis and metabolic
264 health²¹⁻²³, the A-allele of rs17681684 was significantly associated with insulin disposition index
265 ($\beta=-0.067$, $p<0.002$)²², corrected insulin response ($\beta=-0.061$, $p<0.004$)²², glycated haemoglobin
266 1c (HbA1c) ($\beta=0.006$, $p<0.0003$)²¹, and body mass index ($\beta=0.010$, $p<5.3\times 10^{-9}$), in addition to
267 the previously reported positive association with fasting glucose-dependent insulinotropic peptide
268 (GIP) and the suggestive inverse association with post-glucose load GLP-1 ($\beta=-0.035$, $p<4.6\times 10^{-4}$)²⁴.
269 While sample sizes and hence significance levels for insulin traits were not sufficient to support
270 formal colocalisation analysis, we still obtained a high posterior probability (PP>75%) for a shared
271 genetic signal across plasma citrulline, T2D risk, body mass index, and fasting levels of GIP (**Fig. 5B**).
272 We noted, that the *GLP2R* p.Asp470Asn variant was the only of 6 independent genome-wide
273 significant citrulline-raising loci that was associated with a higher risk of T2D, which indicates that
274 the association does not reflect a general effect of blood citrulline levels on T2D risk but rather a
275 locus-specific association at *GLP2R* (**Fig. 5C**). Plasma citrulline levels have been shown to reflect the
276 volume of intestinal cells and are a marker of GLP2R target engagement in the treatment of short-
277 bowel syndrome with glucagon-like peptide 2 analogues²⁵. Taken together, this suggests that
278 genetically higher GLP2R signalling, indicated by the higher citrulline levels among *GLP2R* 470Asn

279 carriers, may lead to chronically elevated GIP (though increased enteroendocrine mass and number
280 of GIP-secreting K-cells), which has been shown to downregulate GIP receptors on pancreatic beta
281 cells²⁶, thereby contributing to the observed reduction in the insulin secretory response and increase
282 in T2D risk.

283 G-protein coupled receptors like GLP2R may signal via G-protein-dependent cyclic adenosine
284 monophosphate (cAMP) production or via G-protein-independent beta-arrestin mediated
285 signalling²⁷. To investigate if the *GLP2R* p.Asp470Asn variant affects signalling via either of these
286 pathways, we expressed the *GLP2R* p.Asp470Asn variant in different *in vitro* models (see **Methods**).
287 We show that the variant allele is significantly associated with reduced recruitment of beta-arrestin
288 to GLP2R upon glucagon-like peptide 2 stimulation, but not with cAMP signalling, which suggests a
289 potential role for impaired beta-arrestin recruitment to GLP2R in the pathophysiology of type 2
290 diabetes (**Fig. 5E-G**).

291 *Serine and glycine levels play a critical role in the aetiology of a rare eye disease*

292 A recent GWAS of macular telangiectasia type 2 (MacTel), a rare neurovascular degenerative
293 retinal disease, identified three genome-wide susceptibility loci (*PHGDH*, *CPS1*, and *TMEM161B*–
294 *LINC00461*) of which the same variants at *PHGDH* and *CPS1* were associated with levels of the amino
295 acids serine and glycine in this GWAS²⁸. More recently, it was shown that low serine availability is
296 linked to both MacTel as well as hereditary sensory and autonomic neuropathy type 1 through
297 elevated levels of atypical deoxyshingolipids²⁹. Whether genetic predisposition to low serine and
298 glycine levels affects MacTel more generally or has predictive utility has not been investigated. To
299 test this and to explore the specificity of associations between genetic influences on metabolite
300 levels and the risk of MacTel, we generated genetic scores (GS) using the sentinel variants for each
301 of the 141 metabolites with at least one significantly associated locus identified in this GWAS and
302 tested their associations with the risk of MacTel. GS's for serine and glycine were the only scores
303 associated with risk for MacTel after removal of the known highly pleiotropic *GCKR* variant (**Fig. 6A**).
304 Each standard deviation higher serine levels via the serine GS was associated with a 95% lower risk
305 of MacTel (odds ratio (95% confidence interval), 0.05 (0.03-0.08); $p=9.5\times 10^{-30}$; **Fig. 6A**). Each of five
306 serine associated variants was individually associated with lower MacTel risk, with a clear dose-
307 response relationship and no evidence of heterogeneity (**Fig. 6B**). The association was unchanged
308 when removing the *GCKR* locus. To disentangle the effect of these two highly correlated metabolites
309 on MacTel risk, we used multivariable Mendelian randomization analysis, which allowed us to test
310 for a causal effect of both measures simultaneously. In this analysis, the effect of serine remained
311 strong, while the effect of glycine was attenuated (**Tab. 2**). Glycine and serine can be interconverted
312 and these results provide genetic evidence that the link between glycine and MacTel is via serine

313 levels through glycine conversion. This hypothesis is supported by the evidence of a log-linear
314 relationship between associations with serine and risk of MacTel among glycine-associated variants
315 (**Fig. 6B**). These findings provide strong evidence that pathways indexed by genetically higher serine
316 levels are strongly and causally associated with protection against MacTel.

317 Given the large observed effect size, we estimated whether using serine and glycine-associated
318 loci might improve the prediction of this rare disease. **Adding genetically predicted glycine and**
319 **serine levels, based on newly discovered metabolite instruments from the present study and**
320 **previous MacTel variants linked to glycine and serine metabolism, substantially improved prediction**
321 **of MacTel based on an area under the receiver operating characteristic curve from 0.65 (CI 95%:**
322 **0.626-0.682) to 0.73 (0.702-0.753) (Fig. 6).**

323 *Common variation at inborn error of metabolism (IEM) associated genes influences the risk of*
324 *common manifestations of diseases related to the phenotypic presentation of those IEMs*

325 In his seminal 1902 work on alkaptonuria³⁰, also known as dark or black urine disease, Archibald
326 Garrod was the first to hypothesise that inborn errors of metabolism are “extreme examples of
327 variations of chemical behaviour which are probably everywhere present in minor degrees”.
328 Previous studies have shown enrichment of metabolite quantitative trait loci in genes known to
329 cause IEMs³¹. **Whether or not common variants at IEM causing loci translate into clinically manifest**
330 **disease remains unknown.** The identification of several metabolite-associated variants at IEM-linked
331 genes in this GWAS meta-analysis allows an investigation of the health consequences of genetically
332 determined differences in metabolism for more frequently occurring variants, representing
333 potentially milder forms of the metabolic and other clinical symptoms of IEMs, and providing new
334 candidate genes for rare extreme metabolic disorders that currently lack a genetic basis (**Fig. 7A**). In
335 this study, there were 153 locus-metabolite associations for which 53 unique IEM-associated genes
336 were prioritized as likely causal using either the hypothesis-free genetic approach or the knowledge-
337 based approach on the basis of the Orphanet database³². In 89% of these associations (136 of 153)
338 the metabolite associated with a given GWAS locus perfectly matched, or was closely related to, the
339 metabolite affected in patients with the corresponding IEM (**Fig. 7B**).

340 To test whether IEM-mirroring lead variants from our metabolite GWAS may increase the risk of
341 common manifestations of diseases known to exist in patients with the corresponding IEM (**Fig. 7A**)
342 we obtained a list of electronic health record diagnosis codes (International Statistical Classification
343 of Diseases and Related Health Problems 10th Revision [ICD-10]) and mapped those based on
344 symptoms seen in both, IEM patients and patients with common, complex disease manifestations
345 (see **Methods**). **We identified 93 ICD-10 codes with at least 500 cases within the UK Biobank study**

346 that aligned with the symptoms or presentations seen in patients with IEMs caused by mutations in
347 genes specifically associated with metabolites observed in the present study. We obtained the
348 association statistics of 85 unique metabolite-associated lead variants at the 136 locus-metabolite
349 associations with these 93 clinical diagnoses and observed 36 associations that met statistical
350 significance (false discovery rate < 5%, **Supplemental Table S6 and Fig. 7B**). For 15 out of those we
351 obtained strong evidence of a shared genetic metabolite-phenotype signal using colocalisation
352 analyses (posterior probability of a shared signal >80%; **Fig. 7D and Supplemental Fig. S3**). These
353 instances linked common genetic variants in or near *APOE*, *PCSK9*, *LPL*, and *LDLR* associated with
354 sphingomyelins (SM 16:0, SM 18:0, and SM-OH 24:1) with atherosclerotic heart disease diagnosis
355 codes (I21, I25), mirroring what is observed in rare familial forms of dyslipidaemia in which these
356 sphingomyelins are elevated and the risk of ischemic heart disease is greatly increased^{33,34}. These
357 results provide further evidence that common variation at IEM genes can lead to clinical phenotypes
358 and diseases that correspond to those that patients with rare mutations in those same genes are
359 severely affected by. Further studies with detailed follow-up for specific outcomes may provide
360 greater power and help clarify the medical consequences of genetic differences in metabolism
361 caused by metabolite altering variants in the general population.

362 **Discussion**

363 This large-scale genome-wide meta-analysis has integrated genetic associations for 174
364 metabolites across different measurement platforms, an approach that has resulted in a three-fold
365 increase in our knowledge of genetic loci regulating levels of these metabolites. We assign likely
366 causal genes for many of the identified associations using a dual approach that combined automated
367 database mining with manual curation.

368 Previous platform-specific genetic studies of blood metabolites have been substantially smaller
369 in size due to being restricted to a single platform and/ or study²⁻¹⁰. We build on these earlier studies
370 to identify and demonstrate enrichment of rare and low-frequency coding variants in enzyme and
371 transporter genes with large effects and reveal the importance of non-linear associations at several
372 loci.

373 Our results not only provide detailed insight into the genetic determinants of human
374 metabolism but consider their relevance for disease aetiology and prediction. We explore both
375 locus-specific and polygenic score effects and provide tangible examples with clear translational
376 potential. We discovered a strong link between GLP2R, citrulline metabolism and T2D, and
377 demonstrate that the p.Asp470Asn variant underlying the citrulline and T2D associations leads to

378 significantly reduced recruitment of beta-arrestin to GLP2R in various cellular models, providing an
379 explanation for a possible pathological mechanism of a variant previously predicted to be benign²⁴.

380 The finding that a standard deviation increase in serine levels via a genetic score is associated
381 with 95% lower risk of MacTel shows that genetic differences resulting in very specific metabolic
382 consequences can have profound effects on health. Our results suggest that inclusion of genetic
383 scores for metabolite levels can improve identification of high risk individuals. Serine and glycine
384 supplementation and/ or pharmacologic modulation of serine metabolism may help to reduce
385 development or alter the prognosis of this rare, severe eye disease, specifically if targeted to people
386 genetically with a genetic susceptibility to low serine levels. It is important to note, that randomized
387 control trials are needed testing this hypothesis before any recommendations on supplementations
388 could be made.

389 We finally show specific examples where common genetic variation in IEM-related genes is
390 associated with phenotypes that are also caused by rare highly penetrant mutations. These results
391 suggest that rare variants in metabolite regulating genes newly identified in our study may be
392 valuable candidate genes in patients without a genetic diagnosis but severe alterations in the
393 corresponding or related metabolites. Hence these results provide a new starting point for further
394 investigations into the relationships between human metabolism and common and rare disorders.

395

396 **Methods**

397 **Study design and participating cohorts**

398 We performed genome-wide meta-analyses of the levels of 174 metabolites from 7 biochemical
399 categories (amino acids, biogenic amines, acylcarnitines, phosphatidylcholines,
400 lysophosphatidylcholines, sphingomyelins, and sum of hexoses) captured by the Biocrates p180 kit
401 measured using mass spectrometry (MS). As described in more detail below, a total of 174
402 metabolites were successfully measured in up to 9,363 plasma samples from genotyped participants
403 of the Fenland study³⁵.

404 To maximise sample size and power, we meta-analysed genome-wide association (GWAS)
405 results from the Fenland cohort with those run in the EPIC-Norfolk³⁶ and INTERVAL³⁷ studies, in
406 which metabolites were profiled using MS (Metabolon Discovery HD4 platform) or protein nuclear
407 magnetic resonance (¹H-NMR) spectrometry³⁸³⁹ (**Supplementary Tab. 1**). Ten of the 174 Biocrates
408 metabolites were covered across all platforms, while 38 were available on the Biocrates and
409 Metabolon platforms and 126 were unique to Biocrates (**Fig. 1**). We integrated publicly available
410 summary statistics from genome-wide meta-analyses of the same metabolites measured using MS
411 (with Biocrates or Metabolon platforms) or ¹H-NMR spectrometry (**Supplementary Tab. 1**).
412 Metabolites were matched across platforms by comparing metabolite names and biochemical
413 formulas. Mapping across different Metabolon platforms was done based on retention time/index
414 (RI), mass to charge ratio (m/z), and chromatographic data (including MS/MS spectral data).
415 Scientists at Metabolon Inc. independently reviewed and confirmed metabolite matches.

416 A summary of the characteristics of participating cohorts is given in **Supplemental Table S1**. The
417 Fenland study is a population-based cohort study of 12,435 participants without diabetes born
418 between 1950 and 1975³⁵. Participants were recruited from general practice surgeries in Cambridge,
419 Ely and Wisbech (United Kingdom) and underwent detailed metabolic phenotyping and genome-
420 wide genotyping. Ethical approval for the Fenland study was given by the Cambridge Local Ethics
421 committee (ref. 04/Q0108/19) and all participants gave their written consent prior to entering the
422 study. The European Prospective Investigation of Cancer (EPIC)-Norfolk study is a prospective cohort
423 of 25,639 individuals aged between 40 and 79 and living in the county of Norfolk in the United
424 Kingdom at recruitment³⁶. The study was approved by the Norfolk Research Ethics Committee (REC
425 ref. 98CN01) and all participants gave their written consent before entering the study. INTERVAL is a
426 randomised trial of approximately 50,000 whole blood donors enrolled from all 25 static centres of
427 NHS Blood and Transplant, aiming to determine whether donation intervals can be safely and
428 acceptably decreased to optimise blood supply whilst maintaining the health of donors³⁷. All

429 participants of the study gave written informed consent and the study was approved by NRES
430 Committee East of England - Cambridge East (ref. 11/EE/0538).

431 **Metabolomics measurements**

432 The levels of 174 metabolites were measured in the Fenland study by the AbsoluteIDQ®
433 Biocrates p180 Kit (Biocrates Life Sciences AG, Innsbruck, Austria) as reported elsewhere in
434 detail^{39,40}. We used a Waters Acquity ultra-performance liquid chromatography (UPLC; Waters Ltd,
435 Manchester, UK) system coupled to an ABSciex 5500 Qtrap mass spectrometer (Sciex Ltd,
436 Warrington, UK). Samples were derivatised and extracted using a Hamilton STAR liquid handling
437 station (Hamilton Robotics Ltd, Birmingham, UK). Flow injection analysis coupled with tandem mass
438 spectrometry (FIA-MS/MS) using multiple reaction monitoring (MRM) in positive mode ionisation
439 was performed to measure the relative levels of acylcarnitines, phosphatidylcholines,
440 lysophosphatidylcholines and sphingolipids. The level of hexose was measured in negative ionisation
441 mode. Ultra-performance liquid chromatography coupled with tandem mass spectrometry using
442 MRM was performed to measure the concentration of amino acids and biogenic amines. The
443 chromatography consisted of a 5-minute gradient starting at 100% aqueous (0.2% Formic acid)
444 increasing to 95% acetonitrile (0.2% Formic acid) over a Waters Acquity UPLC BEH C18 column (2.1 x
445 50 mm, 1.7 µm, with guard column). Isotopically labelled internal standards are integrated within
446 the Biocrates p180 Kit for quantification. Data was processed in the Biocrates Met/DQ software. Raw
447 metabolite readings underwent extensive quality control procedures. Firstly, we excluded from any
448 further analysis metabolites for which the number of measurements below the limit of
449 quantification (LOQ) exceeded 5% of measured samples. Excluded metabolites were carnosine,
450 dopamine, putrescine, asymmetric dimethyl arginine, dihydroxyphenylalanine, nitrotyrosine,
451 spermine, sphingomyelins SM(22:3), SM(26:0), SM(26:1), SM(24:1-OH), phosphatidylcholine acyl-
452 alky 44:4, and phosphatidylcholine diacyl C30:2. Secondly, in samples with detectable but not
453 quantifiable peaks, we assigned random values between 0 and the run-specific LOQ of a given
454 metabolite. Finally, we corrected for batch-effects with a “location-scale” approach, i.e. with
455 normalization for mean and standard deviation of batches.

456 The levels of up to 38 metabolites were measured in EPIC-Norfolk and INTERVAL using the
457 Metabolon HD4 Discovery platform. Measurements were carried out using MS/MS instruments. For
458 these measurements, instrument variability, determined by calculating the median relative standard
459 deviation, was of 6%. Data Extraction and Compound Identification: raw data was extracted, peak-
460 identified and quality control-processed using Metabolon’s hardware and software. Compounds
461 were identified by comparison to library entries of purified standards or recurrent unknown entities.
462 Metabolon maintains a library, based upon authenticated standards, that contains the retention

463 time/index (RI), mass to charge ratio (m/z), and chromatographic data (including MS/MS spectral
464 data) of all molecules present in the library. Identifications were based on three criteria: retention
465 index, accurate mass match to the library +/- 10 ppm, and the MS/MS forward and reverse scores
466 between the experimental data and authentic standards. Metabolite Quantification and Data
467 Normalization: Peaks were quantified using area-under-the-curve. A data normalization step was
468 performed to correct variation resulting from instrument inter-day tuning differences. Essentially,
469 each compound was corrected in run-day blocks by registering the medians to equal one (1.00) and
470 normalizing each data point proportionately (termed the “block correction”).

471 The serum levels of 230 metabolites were measured in the INTERVAL study using ¹H-NMR
472 spectroscopy^{38,41}. Among those, 10 metabolites (creatinine, alanine, glutamine, glycine, histidine,
473 isoleucine, leucine, valine, phenylalanine, and tyrosine) overlapped with what is captured by the
474 Biocrates p180 Kit and were used in the present study. Further details of the ¹H-NMR spectroscopy,
475 quantification data analysis and identification of the metabolites have been described previously^{38,42}.
476 Participants with >30% of metabolite measures missing and duplicated individuals were removed.
477 Metabolite data more than 10 SD from the mean was also removed.

478 **GWAS and meta-analysis**

479 In Fenland and EPIC-Norfolk, metabolite levels were natural log-transformed, winsorised to
480 five standard deviations and then standardised to a mean of 0 and a standard deviation of 1.
481 Genotypes were measured using Affymetrix Axiom or Affymetrix SNP5.0 genotyping arrays. In brief,
482 genotyping in Fenland was done in two waves including 1,500 (Affymetrix SNP5.0) and 9,369
483 (Affymetrix Axiom) participants and imputation was done using IMPUTE2 to 1000 Genomes Phase
484 1v3 (Affymetrix SNP5.0) or phase 3 (Affymetrix Axiom) reference panels (**Supplemental Tab. S1**).
485 Plasma metabolite and genotype data was available for 8,714 (Affymetrix Axiom) and 1,022
486 (Affymetrix SNP5.0) unrelated individuals. In EPIC-Norfolk, 21,044 samples were forwarded to
487 imputation using 1000 Genomes Phase 3 (Oct. 2014) reference panels (**Supplemental Tab. S1**).
488 Imputed SNPs with imputation quality score less than 0.3 or minor allele account less than 2 were
489 removed from the imputed dataset. Genome-wide association analyses were carried out using BOLT-
490 LMM v2.2 adjusting for age, sex, and study-specific covariates in mixed linear models. Alternatively
491 (when the BOLT-LMM algorithm failed due to heritability estimates close to zero or one) analyses
492 were performed using SNPTTEST v2.4.1 in linear regression models, additionally adjusting for the top
493 4 genetic ancestry principal components and excluding related individuals (defined by proportion
494 identity-by-descent calculated in Plink⁴³ > 0.1875 as recommended⁴⁴). GWAS analyses in Fenland
495 were performed within genotyping chip, and associations meta-analysed.

496 In INTERVAL, genotyping was conducted using the Affymetrix Axiom genotyping array. Standard
497 quality control procedures were conducted prior to imputation. The data were phased and imputed
498 to a joint 1000 Genomes Phase 3 (May 2013)-UK10K reference imputation panel. After QC, a total of
499 40,905 participant remained with data obtained by ¹H-NMR spectroscopy. For variants with a MAF
500 of >1% and imputed variants with an info score of >0.4 a univariate GWAS for each of the ten
501 metabolic measures was conducted, after adjustment for technical and seasonal effects, including
502 age, sex, and the first 10 principal components, and rank-based inverse normal transformation. The
503 association analyses were performed using BOLT-LMM v2.2 and R. Data based on the Metabolon
504 HD4 platform was available for 8,455 participants. Prior to the Metabolon HD4 genetic analysis,
505 genetic data were filtered to include only variants with a MAF of >0.01% and imputed variants with
506 an info score of >0.3. Phenotype residuals corrected for age, gender, metabolon batch, INTERVAL
507 centre, plate number, appointment month, the lag time between the blood donation appointment
508 and sample processing, and the first 5 ancestry principal components were calculated for each
509 metabolite and the residuals were standardised prior to the genetic analyses in SNPTEST v2.5.1.

510 For all GWAS analysis within Fenland, EPIC-Norfolk and INTERVAL, variants with Hardy-
511 Weinberg equilibrium $p < 1 \times 10^{-6}$ and associations with absolute value of effect size >5 or standard
512 error (SE) >10 or <0 were excluded; insertions and deletions were excluded.

513 For each metabolite, we performed a meta-analysis of z-scores (betas divided by standard
514 errors) as a measure of association, signals and loci (see below), using METAL software.
515 Heterogeneity between studies for each association was estimated by Cochran's Q-test. For each
516 metabolite, we also performed a meta-analysis of beta and standard errors for the subset of studies
517 (Fenland and, when available, EPIC-Norfolk and/or INTERVAL) where we had access to individual
518 level data and standardised phenotype preparation to estimate effect sizes. Quality filters
519 implemented after meta-analysis included exclusion of SNPs not captured by at least 50% of the
520 participating studies and 50% of the maximum sample size for that metabolite and variants with a
521 minor allele frequency below 0.5%. As a result, meta-analyses assessed the associations of up to
522 13.1 million common or low-frequency autosomal SNPs. Chromosome and base pair positions are
523 determined referring to GRCh37 annotation. To define associations between genetic variants and
524 metabolites, we corrected the conventional threshold of genome wide significance for 102 tests (i.e.
525 $p < 4.9 \times 10^{-10}$), corresponding to the number of principal components explaining 95% of the variance
526 of the 174 metabolites in the Fenland cohort, as previously described⁴⁵.

527 **Signal selection**

528 For each metabolite, we ranked associated SNPs ($p < 4.9 \times 10^{-10}$) by z-score to select trait-sentinel
529 SNPs and defined an “association” region as the region extending 1 Mb to each side of the trait-
530 sentinel SNP. During forward selection of trait-sentinel SNPs and loci for each trait, adjacent and
531 partially overlapping association regions were merged by extending region boundaries to a further 1
532 Mb. After defining trait-sentinel SNPs and association regions we defined overall lead-sentinel SNP
533 and loci for any metabolite using a similar approach. Trait-sentinel SNPs were sorted by z-score for
534 the forward selection of lead-sentinel SNPs and a “locus” was defined as the region extending 1 Mb
535 each side of the lead-sentinel SNP. Regions larger than 2 Mb defined in the trait-sentinel association
536 region definition were carried over in the definition of lead-sentinel SNP loci. As a result, all lead-
537 sentinel SNPs were >1Mb apart from each other and had very low or no linkage disequilibrium ($R^2 <$
538 0.05).

539 For a given locus, independent signals across metabolites were determined based on linkage
540 disequilibrium (LD)-clumping of SNPs that reached the Bonferroni corrected p-value. SNPs with the
541 smallest p-values and an R^2 less than 0.05 were identified as independent signals. LD patterns were
542 estimated with SNP genotype data imputed using the haplotype reference consortium (HRC)
543 reference panel, with additional variants from the combined UK10K plus 1000 Genomes Phase 3
544 reference panel in the EPIC-Norfolk study ($n = 19,254$ after removing ancestry outliers and related
545 individuals).

546 Throughout the manuscript, the term “locus” indicates a genomic region (≥ 1 Mb each side) of a
547 lead-sentinel SNP harbouring one or more trait-sentinel SNPs; “signal” indicates a group of trait-
548 sentinel SNPs in LD with each other but not with other trait-sentinel SNPs in the locus ($R^2 < 0.05$);
549 “association” indicates trait-sentinel SNP to metabolite associations defined by a trait-lead SNP and
550 its surrounding region (≥ 1 Mb each side).

551 We tested at each locus for conditional independent variants using exact stepwise conditional
552 analysis in the largest Fenland sample ($n = 8,714$) using SNPTTEST v2.5 with the same baseline
553 adjustment as in the discovery approach. To refine signals at those loci we used a more recent
554 imputation for this analysis based on the HRC v1 reference panel and additional SNPs imputed using
555 UK10K and 1000G phase 3. We defined secondary signals as those with a conditional p-value $< 5 \times 10^{-8}$.
556 To avoid problems with collinearity we tested after each round if inclusion of a new variant
557 changed associations of all previous variants with the outcome using a joint model. If this model
558 indicated that one or more of the previously selected variants dropped below the applied
559 significance threshold we stopped the procedure, otherwise we repeated this procedure until no
560 further variant met the significance threshold in conditional models. We considered only locus-

561 metabolite associations meeting the GWAS-threshold for significance in the Fenland analysis
562 (n=228).

563 **Investigation of heterogeneity**

564 We used a meta-regression model to identify factors associated with larger I^2 values across all
565 499 identified SNP-metabolite associations. To this end, a vector of heterogeneity estimates, I^2 , from
566 the meta-analysis was obtained as outcome and the following explanatory variables were
567 considered: strength of effect (absolute Z-score of the SNP – metabolite association), biochemical
568 class, dummy variables indicating the study of origin (related to the measurement platform), and the
569 number of contributing studies as an estimate of sample size. A significant effect of any of those
570 terms in a linear regression model was taken to indicate a source of heterogeneity across SNP-
571 metabolite associations and hence identified systematic factors contributing to any observed cross-
572 platform heterogeneity.

573 **Statistical fine-mapping**

574 We used statistical fine mapping to determine 99%-credible intervals for all independently
575 associated SNPs using the R package 'corrcoverage'. Briefly, regional summary statistics (betas and
576 standard errors) were converted to approximate Bayes factors as described in Wakefield *et al.*⁴⁶ to
577 calculate the posterior probability (PP) for each variant driving the association. Credible sets are
578 subsequently defined as the ranked list of variants cumulatively covering 99% of the PP to cover the
579 true causal signal. For loci with evidence of independent secondary signals we used GCTA COJO-cond
580 algorithm to generate conditional association statistics conditioning on all other independent signals
581 in the locus. Since the calculation of approximate Bayes factors requires betas and standard errors
582 we used meta-analysis results across studies for which we had access to individual data (Fenland,
583 EPIC-Norfolk, and INTERVAL). However, out of 546 detected signals 473 reached genome-wide
584 significance ($p < 5 \times 10^{-8}$) in this smaller subset and we restricted fine-mapping to those associations.

585 **Multi-trait colocalisation across metabolites**

586 We used hypothesis prioritisation in multi-trait colocalisation (HyPrColoc)¹⁵ at each of the
587 identified 144 loci 1) to identify metabolites sharing a common causal variant over and above what
588 could be identified in the meta-analysis to increase statistical power, and 2) to identify loci with
589 evidence of multiple causal variants with distinct associated metabolite clusters. Briefly, HyPrColoc
590 aims to test the global hypothesis that multiple traits share a common genetic signal at a genomic
591 location and further uses a clustering algorithm to partition possible clusters of traits with distinct
592 causal variants within the same genomic region. HyPrColoc provides for each cluster three different
593 types of output: 1) a posterior probability (PP) that all traits in the cluster share a common genetic

594 signal, 2) a regional association probability, i.e. that all the metabolites share an association with one
595 or more variants in the region, and 3) the proportion of the PP explained by the candidate variant.
596 We considered a highly likely alignment of a genetic signal across various traits if the PP > 75% or the
597 regional association probability > 80% and the PP > 50%. The second criterion takes into account
598 that metabolites may share multiple causal variants at the same locus. We used the same set of
599 summary statistics as described for statistical fine-mapping, i.e. based on betas and standard errors
600 across studies for which we had access to individual level data. We further filtered metabolites with
601 no evidence of a likely genetic signal ($p > 10^{-5}$) in a region before performing HyPrColoc, which
602 improved clustering across traits by minimizing noise. We used the same workflow to test for the
603 alignment of a genetic signal at the *GLPR2* locus using summary statistics from T2D (see below), a
604 meta-analysis for body mass index across GIANT and UK Biobank, plasma GIP, and plasma citrulline.

605 **Testing for non-linear effects**

606 We tested each of the 499 identified SNP (*j*) – metabolite (*i*) pairs for the deviation from an
607 additive linear model by introducing a dummy variable encoding heterozygous carriers (*D*), i.e. $D = 1$
608 if heterozygous and 0 otherwise, in the following regression model:

$$609 \text{Metabolite}_i \sim \beta_1 + \beta_2 * \text{SNP}_j + \beta_3 * D + \dots \text{Confounder} \dots + \epsilon$$

610 A significant estimate β_3 indicates departure from linearity. In a more formal framework this test
611 allows to test for either a dominant negative or positive model of inheritance depending on the
612 coding of the effect allele. We implemented this test in STATA version 13 using individual level data
613 from the Fenland cohort.

614 **Metabolic network and community detection**

615 We used Gaussian graphical modelling (GGMs) to construct a metabolic network across all 174
616 metabolites in a data-driven manner². Briefly, GGMs are based on partial correlation minimizing
617 confounding and have been shown to recover tight biochemical dependencies from single spot
618 blood measurements. The final network comprised 167 metabolites and 554 significant ($p < 3.3 \times 10^{-6}$)
619 edges. We next performed community detection using the Girvan-Newman algorithm, which
620 successively removes edges with high edge betweenness creating a dendrogram of splits of the
621 network into communities, as implemented in the R package *igraph*. We obtained 14 distinct
622 communities including those covering metabolites of distinct biochemical species as well as
623 subdividing larger metabolite classes (**Supplemental Fig. S2**).

624 **Hypothesis-free (genetic) assignment of causal genes**

625 To assign likely causal genes to lead SNPs at each locus we generated a scoring system. We
626 identified the nearest gene for each variant by querying HaploReg⁴⁷. Next we integrated expression
627 quantitative trait loci (eQTL) studies (GTEx v6p) to identify genes whose expression levels are
628 associated with metabolite levels using TWAS/FUSION (Transcriptome-wide association study /
629 Functional summary-based imputation)⁴⁸. In doing so, we assigned to each variant-metabolite
630 association one or more associated genes using the variant as common anchor. We further assigned
631 higher impact for a causal gene if either the metabolite variant itself or a proxy in high linkage
632 disequilibrium ($R^2 > 0.8$) was a missense variant for a known gene again using the HaploReg database
633 to obtain relevant information. Based on those three criteria we ranked all possible candidate genes
634 and kept those with the highest score as putative causal gene.

635 **Knowledge-based (biological) assignment of causal genes**

636 Metabolite traits are unique among genetically evaluated phenotypes in that the functional
637 characterization of the relevant genes has often already been carried out using classic biochemical
638 techniques. The objective for the knowledge-based assignment strategy was to find the
639 experimental evidence that has previously linked one of the genes proximal to the GWAS lead
640 variant to the relevant metabolite. For many loci and metabolites this 'retrospective' analysis has
641 already been carried out³¹⁴⁹. For these cases, previous causal gene assignments were generally
642 adopted. For novel loci, we employed a dual strategy that combined automated database mining
643 with manual curation. In the automated phase, seven approaches were employed to identify
644 potential causal genes among the 20 protein-coding genes closest to each lead variant, as described
645 in detail below, using the shortest distance determined from the lead SNP to each gene's
646 transcription start site (TSS) or transcription end site (TES), with a distance value of 0 assigned if the
647 SNP fell between the TSS and TES.

648 These 7 approaches were as follows:

- 649 1) HMDB metabolite names⁵⁰ were compared to each entrez gene name;
- 650 2) Metabolite names were compared to the name and synonyms of the protein encoded by each
651 gene⁵¹
- 652 3) HMDB metabolite names and their parent terms (class) were compared to the names for the
653 protein encoded by each gene (UniProt).
- 654 4) Metabolite names were compared to rare diseases linked to each gene in OMIM³² after
655 removing the following non-specific substrings from disease names: uria, emia, deficiency, disease,

656 transient, neonatal, hyper, hypo, defect, syndrome, familial, autosomal, dominant, recessive, benign,
657 infantile, hereditary, congenital, early-onset, idiopathic;

658 5) HMDB metabolite names and their parent terms were compared to all GO biological processes
659 associated with each gene after removing the following non-specific substrings from the name of the
660 biological process: metabolic process, metabolism, catabolic process, response to, positive
661 regulation of, negative regulation of, regulation of. For this analysis only gene sets containing fewer
662 than 500 gene annotations were retained.

663 6) KEGG maps⁵² containing the metabolite as defined in HMDB were compared to KEGG maps
664 containing each gene, as defined in KEGG. For this analysis the large “metabolic process” map was
665 omitted.

666 7) Each proximal gene was compared to the list of known interacting genes as defined in HMDB.
667 For each text-matching based approach, a fuzzy text similarity metric (pair coefficient) as encoded in
668 the ruby gem “fuzzy_match” was used with a score greater than 0.5 considered as a match.

669 In the next step, all automated hits at each locus were manually reviewed for plausibility. In
670 addition, other genes at each locus were reviewed if the Entrez gene or UniProt description of the
671 gene suggested it could potentially be related to the metabolite. If existing experimental evidence
672 could be found linking one of the 20 closest genes to the metabolite, that gene was selected as the
673 biologically most likely causal gene. If no clear experimental evidence existed for any of the 20
674 closest protein coding genes, no causal gene was manually selected. In a few cases multiple genes at
675 a locus had existing experimental evidence. This frequently occurs in the case of paralogs with
676 similar molecule functions. In these cases, all such genes were flagged as likely causal genes.

677 For each manually selected causal gene, the earliest experimental evidence linking the gene
678 (preferably the human gene) to the metabolite was identified. The median publication year for the
679 identified experimental evidence was 2000.

680 **Enrichment of type 2 diabetes associations among metabolite associated lead variants**

681 We examined whether the set of independent lead metabolite associated variants (N=168)
682 were enriched for associations with type 2 diabetes. We plotted observed versus expected $-\log_{10}(p$
683 values) for the 168 lead variants in a QQ-plot, using association statistics from a type 2 diabetes
684 meta-analysis including 80,983 cases and 842,909 non-cases from the DIAMANTE study⁵³ (55,005
685 T2D cases, 400,308 non-cases), UK Biobank⁵⁴ (24,758 T2D cases, 424575 non-cases, application
686 number 44448) and the EPIC-Norfolk study (additional T2D cases not included in DIAMANTE study:
687 1,220 T2D cases and 18,026 non-cases). This QQ-plot was compared to those for 1000 sets of

688 variants, where variants in each set were matched to the index metabolite variants in terms of MAF,
689 the number of variants in LD ($R^2 > 0.5$), gene density and distance to nearest gene (for all parameters
690 +/- 50% of the index variant value), but otherwise randomly sampled from across the autosome
691 excluding the HLA region. MAF and LD parameters for individual variants were determined from the
692 EPIC-Norfolk study (using the combined HRC, UK10K and 1000G imputation as previously described)
693 and gene information was derived from GENCODE v19 annotation⁵⁵. A one-tailed Wilcoxon rank sum
694 test was used to compare the distribution of association $-\log_{10}$ p-values for the metabolite
695 associated variants with that for the randomly sampled, matched, variants.

696 **Functional characterisation of D470N mutant GLP2R**

697 To investigate the functional differences between wild-type (WT) GLP2R and the D470N
698 mutant GLP2R we generated D470N GLP2R mutant constructs using site-directed mutagenesis and
699 characterised canonical GLP2R signalling pathways via cAMP as well as alternative signalling
700 pathways via β -arrestin and P-ERK.

701 *Generation of D470N GLP2R mutant expressing constructs*

702 Human GLP2R cDNA within the pcDNA3.1+ vector was purchased, and Gibson cloning was
703 completed to insert an internal ribosome entry site (IRES) and venus gene downstream of the GLP2R
704 sequence. Following this, QuikChange Lightning site directed mutagenesis was used to perform a
705 single base change from GAC (encoding aspartic acid) to AAC (encoding asparagine) at amino acid
706 position 470 (**Supplemental Fig. 4A-B**). Successful mutagenesis was confirmed by DNA Sanger
707 sequencing (**Supplemental Fig. 4C**), and the successful products were scaled up for use in functional
708 assays. The WT and mutant GLP2R constructs within the pcDNA3.1+ vector were used to assess
709 signalling by cAMP and P-ERK. To determine β -arrestin recruitment using NanoBiT[®] technology, an
710 alternative vector was required for lower expression of GLP2R, and fusion of GLP2R to the Large BiT
711 subunit of NanoBiT[®]. For this, GLP2R was cloned into the pBiT1.1_C[TK/LgBiT] vector using
712 restriction cloning and ligation. DNA Sanger sequencing was then used for confirmation of successful
713 cloning.

714 *Comparison of WT and D470N GLP2R signalling via cAMP*

715 After generation of WT and D470N GLP2R containing constructs, these were used to assess
716 differences in WT and mutant GLP2R signalling. The initial signalling pathway to be assessed was Gas
717 signalling via cAMP. CHO K1 cells were transiently transfected with WT or mutant GLP2R constructs,
718 then after 16-24 hours were treated with a dose response of GLP-2. cAMP levels were measured
719 following 30 minutes of GLP-2 treatment, in an end-point lysis HitHunter[®] cAMP assay. The presence
720 of IRES-Venus within the GLP2R expressing vectors allowed transfection efficiency to be determined

721 for each construct. Transfection efficiency was approximately 60-70%, with no differences between
722 the WT and mutant constructs. Comparison of the GLP-2 dose-response in WT and mutant GLP2R
723 expressing cells revealed no significant differences in signalling, with an almost overlapping dose
724 response curve (**Fig. 5E**).

725 *Comparison of β -arrestin recruitment to the WT and D470N GLP2R*

726 Both β -arrestin 1 and β -arrestin 2 recruitment were assessed using a Nano-Glo® live cell
727 assay in transiently transfected HEK293 cells. Briefly, the recruitment of β -arrestin to GLP2R brings
728 the large and small BiT subunit of NanoBiT® together, resulting in increased luciferase activity. The
729 top concentrations from the GLP-2 dose response in the cAMP assay (1–100 nmol/l GLP-2) were
730 chosen for stimulation of the GLP2R and observation of β -arrestin recruitment. Both β -arrestin 1 and
731 β -arrestin 2 were recruited to the WT GLP2R upon GLP-2 stimulation, in a dose-dependent manner
732 (**Supplemental Fig. 5a, c**). The maximal luciferase activity for both β -arrestin 1 and β -arrestin 2
733 recruitment to the mutant GLP2R was significantly decreased when compared to the WT GLP2R,
734 indicating the extent of β -arrestin recruitment was markedly decreased (**Supplemental Fig. 5b, d**).
735 The example traces indicate that neither β -arrestin 1 or β -arrestin 2 were recruited to the mutant
736 GLP2R upon stimulation with 1 nmol/l GLP-2, however the same concentration of GLP-2 induced β -
737 arrestin recruitment to the WT GLP2R. Overall there was a significant decrease in β -arrestin 1 and β -
738 arrestin 2 recruitment to the D470N GLP2R mutant (**Figure 5F-G**).

739 **Genetic score and Mendelian randomization analysis for macular telangiectasia type 2**

740 For each metabolite a genetic score (GS) was calculated using all variants meeting genome-
741 wide significance and their beta-estimates as weights obtained from the meta-analysis of studies for
742 which individual level data was available. We used fixed-effect meta-analysis to test for the effect of
743 the GS on MacTel risk using the summary statistics from the most recent GWAS. A conservative
744 Bonferroni-correction for the number of tested GS's was used to declare significance ($p < 3.5 \times 10^{-4}$).
745 Sensitivity analyses were performed where the pleiotropic *GCKR* variant was removed.

746 To test for causality between circulating levels of glycine and serine for MacTel we
747 performed two types of Mendelian randomization (MR) analysis. In a two-sample univariable MR⁵⁶
748 we tested for an individual effect of serine (n=4 SNPs) or glycine (n=15 SNPs) on the risk of MacTel
749 using independent non-pleiotropic (i.e. the variant in *GCKR*) genome-wide SNPs as instruments. To
750 this end, we used the inverse variance weighted method to pool SNP ratio estimates using random
751 effects as implemented in the R package *MendelianRandomization*. SNP effects on the risk for
752 MacTel were obtained from²⁸. To disentangle the individual effect of those two highly correlated
753 metabolites at the same time we used a multivariable MR model⁵⁷ including all SNPs related to

754 serine or glycine (n=15 SNPs). Beta estimates and standard errors for both metabolites and all SNPs
755 were obtained from the summary statistics and mutually used as exposure variables in multivariable
756 MR. Effect estimates were again pooled using a random effect model as implemented in the R
757 package *MendelianRandomization*. This procedure allowed us to obtain causal estimates for both
758 metabolites while accounting for the effect on each other. Estimates can be interpreted as increase
759 in risk for MacTel per 1 SD increase in metabolite levels while holding the other metabolite constant.

760 To estimate a potential clinical usefulness of the identified variants we constructed two
761 GRS's for MacTel using a) sex, the first genetic principal component, and the SNPs rs73171800 and
762 rs9820286 which were identified by the MacTel GWAS study²⁸ but not found to be related to either
763 glycine or serine in our study and b) all the previous but additionally including genetically predicted
764 serine and glycine at individual levels, via GS, to the model. An interaction between serine and sex at
765 birth was included to reflect the interaction between SNP rs715 and sex as previously identified²⁸.
766 To assess the predictive ability of both models, receiver operating characteristic curves were
767 computed based on prediction values in 1,733 controls and 476 MacTel cases.

768 **Identification of genes related to inborn errors of metabolism**

769 Biologically or genetically assigned candidate genes were annotated for IEM association
770 using the Orphanet database³². Using a binomial two-tailed test, enrichment of metabolic loci was
771 assessed by comparing the annotated list with the full list of 784 IEM genes in Orphanet against a
772 backdrop of 19,817 protein-coding genes⁵⁸. IEM-annotated loci for which the associated metabolite
773 matched or was closely biochemically related to the IEM corresponding metabolite(s) based on
774 IEMBase⁵⁹ were considered further for analysis.

775 We hypothesised that IEM-annotated loci with metabolite-specific consequences could also
776 have phenotypic consequences similar to the IEM. To test this, we first obtained terms describing
777 each IEM and translated them into IEM-related ICD-10 codes using the Human Phenotype Ontology
778 and previously-generated mappings^{60,61}. We obtained association statistics from the 85 IEM SNPs for
779 phenotypic associations with corresponding ICD-codes among UK Biobank restricting to diseases
780 with at least 500 cases (N=93, **Fig. 7B**, <http://www.nealelab.is/uk-biobank>). We tested locus-disease
781 pairs meeting statistical significance (controlling the false discovery rate at 5% to account for
782 multiple testing) for a common genetic signal with the corresponding locus-metabolite association
783 using statistical colocalisation. **Because of the hypothesis-driven nature of the approach, i.e. prior**
784 **knowledge of the causal gene and metabolite effect for a given IEM, we adopted an FDR-based**
785 **strategy to account for multiple testing. We further highlight only those examples with strong**
786 **evidence for a shared genetic signal (see below).**

787 **Colocalisation analyses**

788 We used statistical colocalisation⁶² to test for a shared genetic signal between a metabolite and
789 a disease of interest. We obtained posterior probabilities (PP) of: H0 – no signal; H1 – signal unique
790 to the metabolite; H2 – signal unique to the trait; H3 – two distinct causal variants in the same locus
791 and H4 – presence of a shared causal variant between a metabolite and a given trait. PPs above 80%
792 were considered highly likely. We used p-values and MAFs obtained from the summary statistics
793 with default priors to perform colocalisation.

794 **Acknowledgement/Funding**

795 M.P. was supported by a fellowship from the German Research Foundation (DFG PI 1446/2-1). C.O.
796 was founded by an early career fellowship at Homerton College, University of Cambridge. L. B. L. W.
797 acknowledges funding by the Wellcome Trust (WT083442AIA). J.G. was supported by grants from
798 the Medical Research Council (MC_UP_A090_1006, MC_PC_13030, MR/P011705/1 and
799 MR/P01836X/1). Work in the Reimann/Gribble laboratories was supported by the Wellcome Trust
800 (106262/Z/14/Z and 106263/Z/14/Z), UK Medical Research Council (MRC_MC_UU_12012/3) and
801 PhD funding for EKB from MedImmune/AstraZeneca. Praveen Surendran is supported by a
802 Rutherford Fund Fellowship from the Medical Research Council grant MR/S003746/1. A. W. is
803 supported by a BHF-Turing Cardiovascular Data Science Award and by the EC-Innovative Medicines
804 Initiative (BigData@Heart). J.D. is funded by the National Institute for Health Research [Senior
805 Investigator Award] [*]. The EPIC-Norfolk study (<https://doi.org/10.22025/2019.10.105.00004>) has
806 received funding from the Medical Research Council (MR/N003284/1 and MC-UU_12015/1) and
807 Cancer Research UK (C864/A14136). The genetics work in the EPIC-Norfolk study was funded by the
808 Medical Research Council (MC_PC_13048). Metabolite measurements in the EPIC-Norfolk study
809 were supported by the MRC Cambridge Initiative in Metabolic Science (MR/L00002/1) and the
810 Innovative Medicines Initiative Joint Undertaking under EMIF grant agreement no. 115372. We are
811 grateful to all the participants who have been part of the project and to the many members of the
812 study teams at the University of Cambridge who have enabled this research. The Fenland Study is
813 supported by the UK Medical Research Council (MC_UU_12015/1 and MC_PC_13046). Participants
814 in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS
815 Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other
816 elements of the trial. DNA extraction and genotyping was co-funded by the National Institute for
817 Health Research (NIHR), the NIHR BioResource (<http://bioresource.nihr.ac.uk>) and the NIHR
818 [Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation
819 Trust] [*]. Nightingale Health NMR assays were funded by the European Commission Framework
820 Programme 7 (HEALTH-F2-2012-279233). Metabolite Metabolomics assays were funded by the NIHR

821 BioResource and the National Institute for Health Research [Cambridge Biomedical Research Centre
822 at the Cambridge University Hospitals NHS Foundation Trust] [*]. The academic coordinating centre
823 for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in
824 Donor Health and Genomics (NIHR BTRU-2014-10024), UK Medical Research Council
825 (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and the
826 NIHR [Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation
827 Trust] [*].The academic coordinating centre would like to thank blood donor centre staff and blood
828 donors for participating in the INTERVAL trial.

829 This work was supported by Health Data Research UK, which is funded by the UK Medical Research
830 Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council,
831 Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government
832 Health and Social Care Directorates, Health and Social Care Research and Development Division
833 (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and
834 Wellcome.

835 *The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or
836 the Department of Health and Social Care.

837 UK Biobank: This research has been conducted using the UK Biobank resource under
838 Application Number 44448.

839 **Author Contribution**

840 Concept and design: L.A.L. and C.L.

841 Generation, acquisition, analysis and/or interpretation of data: all authors.

842 Drafting of the manuscript: L.A.L., M.P., and C.L.

843 Critical review of the manuscript for important intellectual content and approval of the final version
844 of the manuscript: all authors.

845 **Competing Interests statement**

846 A.S.B. has received grants from AstraZeneca, Biogen, Bioverativ, Merck, Novartis, and Sanofi. J. D.
847 sits on the International Cardiovascular and Metabolic Advisory Board for Novartis (since 2010), the
848 Steering Committee of UK Biobank (since 2011), the MRC International Advisory Group (ING)
849 member, London (since 2013), the MRC High Throughput Science 'Omics Panel Member, London
850 (since 2013), the Scientific Advisory Committee for Sanofi (since 2013), the International

851 Cardiovascular and Metabolism Research and Development Portfolio Committee for Novartis and
852 the Astra Zeneca Genomics Advisory Board (2018). E.B.F. is an employee and stock holder of Pfizer.

853 **Data Availability**

854 All genome-wide summary statistics will be made available through an interactive webserver upon
855 publication of the manuscript.

856 **Code Availability**

857 Each use of software programs has been clearly indicated and information on the options that were
858 used is provided in the Methods section. Source code to call programs is available upon request.

859 **REFERENCES**

- 860 1. Wishart, D. S. Metabolomics for investigating physiological and pathophysiological processes.
861 *Physiol. Rev.* **99**, 1819–1875 (2019).
- 862 2. Shin, S.-Y. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**,
863 543–550 (2014).
- 864 3. Long, T. *et al.* Whole-genome sequencing identifies common-to-rare variants associated with
865 human blood metabolites. *Nat. Genet.* **49**, 568–578 (2017).
- 866 4. Draisma, H. H. M. *et al.* Genome-wide association study identifies novel genetic variants
867 contributing to variation in blood metabolite levels. *Nat. Commun.* **6**, 7208 (2015).
- 868 5. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and
869 reveals novel systemic effects of LPA. *Nat. Commun.* **7**, 11122 (2016).
- 870 6. Illig, T. *et al.* A genome-wide perspective of genetic variation in ... [Nat Genet. 2010] -
871 PubMed result. *Nat. Genet.* **42**, 137–41 (2010).
- 872 7. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research.
873 *Nature* **477**, 54–60 (2011).
- 874 8. Rhee, E. P. P. *et al.* A genome-wide association study of the human metabolome in a
875 community-based cohort. *Cell Metab.* **18**, 130–43 (2013).
- 876 9. Gallois, A. *et al.* A comprehensive study of metabolite genetics reveals strong pleiotropy and
877 heterogeneity across time and context. *Nat. Commun.* **10**, 1–13 (2019).
- 878 10. Rhee, E. P. *et al.* An exome array study of the plasma metabolome. *Nat. Commun.* **7**, 12360
879 (2016).
- 880 11. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to
881 Common Complex Disease. *Cell* **167**, 1415–1429.e19 (2016).
- 882 12. Bansal, N. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
- 883 13. Learn, D. B., Fried, V. A. & Thomas, E. L. Taurine and hypotaurine content of human
884 leukocytes. *J. Leukoc. Biol.* **48**, 174–182 (1990).
- 885 14. Yet, I. *et al.* Genetic Influences on Metabolite Levels: A Comparison across Metabolomic
886 Platforms. *PLoS One* **11**, e0153672 (2016).
- 887 15. Foley, C. N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic
888 risk factors across multiple traits. **44**, 1–47 (2019).
- 889 16. Pedersen, C. B. *et al.* The ACADS gene variation spectrum in 114 patients with short-chain
890 acyl-CoA dehydrogenase (SCAD) deficiency is dominated by missense variations leading to
891 protein misfolding at the cellular level. *Hum Genet* **124**, 43–56 (2008).
- 892 17. Lahiri, S. *et al.* Kinetic characterization of mammalian ceramide synthases: Determination of
893 Km values towards sphinganine. *FEBS Lett.* **581**, 5289–5294 (2007).
- 894 18. Horowitz, B. *et al.* Asparagine synthetase activity of mouse leukemias. *Science (80-.).* **160**,

- 895 533–535 (1968).
- 896 19. Babu, E. *et al.* Identification of a Novel System L Amino Acid Transporter Structurally Distinct
897 from Heterodimeric Amino Acid Transporters. *J. Biol. Chem.* **278**, 43838–43845 (2003).
- 898 20. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in
899 Europeans. *Diabetes* **66**, 2888–2902 (2017).
- 900 21. Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2
901 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide
902 meta-analysis. *PLoS Med.* **14**, (2017).
- 903 22. Prokopenko, I., Poon, W., Mägi, R., Prasad, B. R. & Salehi, S. A. A Central Role for GRB10 in
904 Regulation of Islet Function in Man. *Claire Levy-Marchal* **17**,.
- 905 23. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies
906 genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–
907 669 (2012).
- 908 24. Almgren, P. *et al.* Genetic determinants of circulating GIP and GLP-1 concentrations. (2017)
909 doi:10.1172/jci.insight.93306.
- 910 25. Fragkos, K. C. & Forbes, A. *Citrulline as a marker of intestinal function and absorption in
911 clinical settings: A systematic review and meta-analysis. United European Gastroenterology
912 Journal* vol. 6 181–191 (SAGE Publications Ltd, 2018).
- 913 26. Tseng, C. C. & Zhang, X. Y. The cysteine of the cytoplasmic tail of glucose-dependent
914 insulinotropic peptide receptor mediates its chronic desensitization and down-regulation.
915 *Mol. Cell. Endocrinol.* **139**, 179–186 (1998).
- 916 27. Estall, J. L., Koehler, J. A., Yusta, B. & Drucker, D. J. The glucagon-like peptide-2 receptor C
917 terminus modulates β -arrestin-2 association but is dispensable for ligand-induced
918 desensitization, endocytosis, and G-protein-dependent effector activation. *J. Biol. Chem.* **280**,
919 22124–22134 (2005).
- 920 28. Scerri, T. S. *et al.* Genome-wide analyses identify common variants associated with macular
921 telangiectasia type 2. *Nat. Genet.* **49**, 559–567 (2017).
- 922 29. Gantner, M. L. *et al.* Serine and lipid metabolism in macular disease and peripheral
923 neuropathy. *N. Engl. J. Med.* **381**, 1422–1433 (2019).
- 924 30. Garrod, A. E. The incidence of alkaptonuria: a study in chemical individuality. *Lancet* **160**,
925 1616–1620 (1902).
- 926 31. Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**,
927 543–550 (2014).
- 928 32. Rath, A. *et al.* Representation of rare diseases in health information systems: The orphanet
929 approach to serve a wide range of end users. *Hum. Mutat.* **33**, 803–808 (2012).
- 930 33. Stübiger, G. *et al.* Targeted profiling of atherogenic phospholipids in human plasma and
931 lipoproteins of hyperlipidemic patients using MALDI-QIT-TOF-MS/MS. *Atherosclerosis* **224**,
932 177–186 (2012).
- 933 34. Van Der Graaf, A., Kastelein, J. J. P. P. & Wiegman, A. *Heterozygous familial
934 hypercholesterolaemia in childhood: Cardiovascular risk prevention. Journal of Inherited
935 Metabolic Disease* vol. 32 (2009).
- 936 35. Lindsay, T. *et al.* Descriptive epidemiology of physical activity energy expenditure in UK adults
937 (The Fenland study). *Int. J. Behav. Nutr. Phys. Act.* **16**, 126 (2019).
- 938 36. Day, N. *et al.* EPIC-Norfolk: study design and characteristics of the cohort. European
939 Prospective Investigation of Cancer. *Br. J. Cancer* **80 Suppl 1**, 95–103 (1999).
- 940 37. Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations
941 can be safely and acceptably decreased to optimise blood supply: Study protocol for a
942 randomised controlled trial. *Trials* **15**, (2014).
- 943 38. Soinen, P. *et al.* High-throughput serum NMR metabolomics for cost-effective holistic
944 studies on systemic metabolism. *Analyst* **134**, 1781–5 (2009).
- 945 39. Wittemans, L. B. L. *et al.* Assessing the causal association of glycine with risk of cardio-

- 946 metabolic diseases. *Nat. Commun.* **10**, (2019).
- 947 40. Lotta, L. A. *et al.* Genetic Predisposition to an Impaired Metabolism of the Branched-Chain
948 Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis. *PLoS Med.*
949 (2016) doi:10.1371/journal.pmed.1002179.
- 950 41. Di Angelantonio, E. *et al.* Efficiency and safety of varying the frequency of whole blood
951 donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet* **390**, 2360–2371 (2017).
- 952 42. Inouye, M. *et al.* Metabonomic, transcriptomic, and genomic variation of a population cohort.
953 *Mol. Syst. Biol.* **6**, 441 (2010).
- 954 43. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage
955 analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 956 44. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat.*
957 *Protoc.* **5**, 1564–1573 (2011).
- 958 45. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a
959 correlation matrix. *Heredity (Edinb).* **95**, 221–227 (2005).
- 960 46. Wakefield, J. Bayes Factors for Genome-Wide Association Studies : Comparison with P -
961 values. **86**, 79–86 (2009).
- 962 47. Ward, L. D. & Kellis, M. HaploReg: A resource for exploring chromatin states, conservation,
963 and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*
964 **40**, (2012).
- 965 48. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies.
966 *Nat. Genet.* **48**, 245–252 (2016).
- 967 49. Stacey, D. *et al.* ProGeM: A framework for the prioritization of candidate causal genes at
968 molecular quantitative trait loci. *Nucleic Acids Res.* **47**, (2019).
- 969 50. Wishart, D. S. *et al.* HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.*
970 **46**, D608–D617 (2018).
- 971 51. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**,
972 D158–D169 (2017).
- 973 52. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives
974 on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
- 975 53. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-
976 density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
- 977 54. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide
978 Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, (2015).
- 979 55. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes.
980 *Nucleic Acids Res.* **47**, D766–D773 (2019).
- 981 56. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with
982 multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
- 983 57. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: The use of pleiotropic
984 genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).
- 985 58. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE
986 project. *Genome Res.* **22**, 1760–1774 (2012).
- 987 59. Lee, J. J. Y., Wasserman, W. W., Hoffmann, G. F., Van Karnebeek, C. D. M. & Blau, N.
988 Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism.
989 *Genet. Med.* **20**, 151–158 (2018).
- 990 60. Köhler, S. *et al.* The human phenotype ontology in 2017. *Nucleic Acids Res.* **45**, D865–D876
991 (2017).
- 992 61. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM codes to phecodes: Workflow development and
993 initial evaluation. *J. Med. Internet Res.* **21**, (2019).
- 994 62. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association
995 studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
- 996

998 **FIGURE LEGENDS**
999

1000 **Figure 1A** Sample size by contributing study and technique for each of the 174 metabolites included.
1001 **B** A three-dimensional Manhattan plot displaying chromosomal position (x-axis) of significant
1002 associations ($p < 4.9 \times 10^{-10}$, z-axis) across all metabolites (y-axis). Colours indicate metabolite groups.
1003 **C** A top view of the 3D-Manhattan plot. Dots indicate significantly associated loci. Colours indicate
1004 novelty of metabolite – locus associations. Loci with indication for pleiotropy have been annotated.

1005

1006 **Figure 2A** Distribution of pleiotropy, i.e. number of associated metabolites, among loci identified in
1007 the present study. **B** Distribution of polygenicity of metabolites, i.e. number of identified loci for
1008 each metabolite under investigation. **C** Scatterplot comparing the estimated heritability of each
1009 metabolite against the number of associated loci. Size of the dots indicates samples sizes. **D**
1010 Heritability estimates for single metabolites. Colours indicate the proportion of heritability
1011 attributed to single nucleotide polymorphisms (SNPs) with large effect sizes ($\beta > 0.25$ per allele). **E** –
1012 **M** SNP – metabolite association with indication of non-additive effects. Beta is an estimate from the
1013 departure of linearity. **N** Barplot showing the increase in heritability and explained variance for each
1014 SNP – metabolite pair when including non-additive effects.

1015

1016 **Figure 3A** Scatterplot comparing the minor allele frequencies (MAF) of associated variants with
1017 effect estimates from linear regression models (N loci=499). Colours indicate possible functional
1018 consequences of each variant: maroon – nonsynonymous variant; blue – in strong LD ($r^2 > 0.8$) with a
1019 nonsynonymous variant and grey otherwise. **B-D** Distribution of effect sizes (B), allele frequencies
1020 (C), and width of credible sets (D) based on the type of single nucleotide polymorphism (SNP) (0 –
1021 non-coding or synonymous, 1 – in strong LD with nonsynonymous, 2 - nonsynonymous). **E**
1022 Distribution of functional annotations of metabolite associated variants (red), trait-associated
1023 variants (blue – continuous, purple – diseases) obtained from the GWAS catalogue, and all SNPs
1024 included in the present genome-wide association studies. The inlet for exonic variants distinguishes
1025 between synonymous (syn) and nonsynonymous variants (nsyn).

1026

1027 **Figure 4A** Comparison between the hypothesis-free genetically prioritized versus biologically
1028 plausible approaches used in the present study to assign candidate genes to metabolite associated
1029 single nucleotide polymorphisms. The Venn-diagram displays the overlap between both approaches.
1030 **B** Enrichment of genetically prioritized genes among biologically plausible or genes linked to inborn
1031 errors of metabolism (IEM). **C** Proportion of genetically prioritized genes encoding for either
1032 enzymes or transporters.

1033

1034 **Figure 5A** Enrichment of associations with type 2 diabetes (T2D: 80,983 cases, 842,909 controls)
1035 among metabolite-associated SNPs. Blue dots indicate metabolite-SNPs and grey dots indicate a
1036 random selection of matched control SNPs. **B** Regional association plots for plasma citrulline, type 2
1037 diabetes, body mass index, and fasting levels of glucose-dependent insulinotropic peptide (GIP)
1038 focussing on the *GLPR2* gene. Variants are coloured based on linkage disequilibrium with the lead
1039 variant (rs17681684) for plasma citrulline. *Summary statistics for GIP were obtained from the more
1040 densely genotyped study included in Almgren et al.²⁴ (to increase coverage of genetic variants for
1041 multi-trait colocalisation). **C** Individual association summary statistics for all citrulline associated
1042 SNPs (coded by the citrulline increasing allele) for T2D and an inverse-variance weighted (IVW)
1043 estimate pooling all effects. **D** Schematic sketch for the location of the missense variant induces
1044 amino acid substitution in the glucagon-like peptide-2 receptor (GLP2R). **E** GLP-2 dose response
1045 curves in cAMP assay for GLP2R wild-type and mutant receptors. The dose response curves of cAMP
1046 stimulation by GLP-2 in CHO K1 cells transiently transfected with either GLP2R wild-type or mutant

1047 constructs. Data were normalised to the wild-type maximal and minimal response, with 100% being
1048 GLP-2 maximal stimulation of the wild-type GLP2R, and 0% being wild-type GLP2R cells with buffer
1049 only. Mean \pm standard errors are presented (n=4). **F-G** Summary of wild-type and mutant GLP2R
1050 beta-arrestin 1 and beta-arrestin 2 responses. Area under the curve (AUC) summary data (n=3-4)
1051 displayed for beta-arrestin 1 recruitment (E) and beta-arrestin 2 recruitment (F). AUCs were
1052 calculated using the 5 minutes prior to ligand addition as the baseline value. Mean \pm standard errors
1053 are presented. Normal distribution of log₁₀ transformed data was determined by the D'Agostino &
1054 Pearson normality test. Following this statistical significance was assessed by one-way ANOVA with
1055 post hoc Bonferroni test. ***p<0.001, *p<0.05.

1056

1057 **Figure 6A** Results from genetic scores for each metabolite on risk for macular telangiectasia type 2
1058 (MacTel). The dotted line indicates the level of significance after correction for multiple testing. The
1059 inset shows the same results but after dropping the pleiotropic variants in *GCKR* and *FADS1-2*. **B**
1060 Effect estimates of serine-associated genetic variants on the risk for MacTel. **C** Comparison of effect
1061 sizes for lead variants associated with plasma serine levels and the risk for MacTel. **D** Receiver
1062 operating characteristic curves (ROC) comparing the discriminative performance for MacTel using a)
1063 sex, the first genetic principal component, and two MacTel variants (*rs73171800* and *rs9820286*) not
1064 associated with metabolite levels, and b) additionally including genetically predicted serine and
1065 glycine at individual levels as described in the methods. The area under the curve (AUC) is given in
1066 the legend.

1067

1068 **Figure 7A** Scheme of the workflow to link common variation in genes causing inborn errors of
1069 metabolism (IEM) to complex diseases. **7B** Flowchart for the systematic identification of metabolite-
1070 associated variants to genes and diseases related to inborn errors of metabolism (IEM). **C** P-values
1071 from phenome-wide association studies among UK Biobank using variants mapping to genes
1072 knowing to cause IEMs and binary outcomes classified with the ICD-10 code. Colours indicate
1073 disease classes. The dotted line indicates the significance threshold controlling the false discovery
1074 rate at 5%. **D** Posterior probabilities (PPs) from statistical colocalisation analysis for each significant
1075 triplet consisting of a metabolite, a variant, and a ICD-10 code among UK Biobank. The dotted line
1076 indicates high likelihood (>80%) for one of the four hypothesis tested: H0 – no signal; H1 – signal
1077 unique to the metabolite; H2 – signal unique to the trait; H3 – two distinct causal variants in the
1078 same locus and H4 – presence of a shared causal variant between a metabolite and a given trait.

1079

1080 TABLES
 1081
 1082
 1083

Table 1 Genomic loci with effect sizes larger than 0.25 units in standard deviation of metabolite levels per allele.

rsID	Position*	Metabolite	EA/OA	EA	N	MA p-value	Beta (se)**	Candidate genes	Expl. var. (%)
rs13538	2:73868328	Acetylcarnitine	A/G	0.78	30692	1.99E-1984	0.85 (0.01)	NAT8, ACTG2	18.4
rs3916	12:121177272	Butyrylcarnitine	C/G	0.26	30694	1.67E-2010	0.81 (0.01)	ACADS,	16.9
rs12587599	14:104575130	Asparagine	T/C	0.14	23606	8.98E-294	0.49 (0.013)	ASPG, ADSSL1	8.2
rs3970551	22:18906839	Proline	G/A	0.11	23618	1.10E-224	0.48 (0.015)	PRODH	5.0
rs174547	11:61570783	lysoPC a C20:4	T/C	0.67	16829	4.42E-398	0.47 (0.015)	FADS1, DAGLA	9.9
rs174545	11:61569306	PC aa C38:4	C/G	0.67	16828	1.37E-361	0.45 (0.015)	FADS1,	9.2
rs715	2:211543055	Glycine	C/T	0.31	80000	3.00E-1632	0.44 (0.006)	CPS1, IDH1	12.9
rs174564	11:61588305	PC ae C42:3	A/G	0.66	9363	5.72E-183	0.44 (0.015)	FADS1, DAGLA	8.9
rs174547	11:61570783	PC aa C36:4	T/C	0.67	16830	3.25e-313	0.43 (0.015)	FADS1, DAGLA	8.6
rs1171617	10:61467182	Carnitine	T/G	0.77	31001	2.06E-444	0.43 (0.011)	SLC16A9,	7.0
rs102275	11:61557803	PC ae C40:5	T/C	0.67	16839	8.23E-202	0.43 (0.015)	C11orf10, DAGLA	8.7
rs7157785	14:64235556	PC aa C28:1	T/G	0.16	16833	4.60E-136	0.35 (0.019)	SGPP1,SYNE2	3.3
rs174547	11:61570783	PC ae C36:5	T/C	0.67	16828	2.48E-185	0.33 (0.015)	FADS1, DAGLA	5.1
rs102275	11:61557803	PC aa C38:5	T/C	0.67	16836	8.31E-198	0.33 (0.015)	C11orf10, DAGLA	5.0
rs174564	11:61588305	PC ae C42:2	A/G	0.66	9363	7.04E-99	0.32 (0.015)	FADS1, DAGLA	4.8
rs174564	11:61588305	lysoPC a C26:1	A/G	0.66	9363	1.38E-91	0.32 (0.016)	FADS1, DAGLA	4.6
rs7157785	14:64235556	SM (OH) C14:1	T/G	0.16	16833	1.65E-96	0.29 (0.019)	SGPP1	2.2
rs174546	11:61569830	PC aa C24:0	C/T	0.67	13184	4.16E-89	0.29 (0.016)	FADS1, DAGLA	3.6
rs174546	11:61569830	PC ae C38:5	C/T	0.67	16839	8.98E-146	0.29 (0.015)	FADS1, DAGLA	3.9
rs7552404	1:76135946	Octanoylcarnitine	A/G	0.69	31969	2.30E-260	0.28 (0.01)	ACADM	2.8
rs1171615	10:61469090	Propionylcarnitine	T/C	0.77	32590	7.09E-185	0.27 (0.011)	SLC16A9	3.1
rs1171617	10:61467182	Acetylcarnitine	T/G	0.77	31008	1.92E-156	0.27 (0.011)	SLC16A9	3.3
rs2286963	2:211060050	Nonacylcarnitine	G/T	0.36	13925	5.46E-159	0.26 (0.016)	ACADL	3.2
rs12210538	6:110760008	Octadecandienylcarnitine	A/G	0.77	30227	1.69E-144	0.26 (0.011)	SLC22A16	1.0
rs102275	11:61557803	PC aa C36:5	T/C	0.66	16835	2.09E-120	0.25 (0.015)	C11orf10, DAGLA	3.0
rs174550	11:61571478	PC ae C36:3	C/T	0.33	16830	2.05E-105	0.25 (0.015)	FADS1, DAGLA	2.7

EA = effect allele; OA = other allele; MA = meta-analysis; se = standard error; *Chromosome:Position based on Genome Reference Consortium Human Build 37; **based on meta-analysis across cohorts for which individual-level data was available (more information is provided in Supplementary Tab. S2).

1084
 1085
 1086
 1087

1088

1089 **Table 2** Results from Mendelian randomisation (MR) analysis between metabolite levels and risk of
1090 macular telangiectasia type 2.

1091

Metabolite	Univariable MR	Multivariable MR
<i>Serine</i> (4 SNPs)		
Odds ratio per SD increase	0.06 (0.03; 0.13)	0.10 (0.05; 0.21)
p-value	9.45×10^{-12}	2.95×10^{-9}
<i>Glycine</i> (15 SNPs)		
Odds ratio per SD increase	0.17 (0.08; 0.37)	0.50 (0.29; 0.87)
p-value	9.99×10^{-6}	1.35×10^{-2}

1092

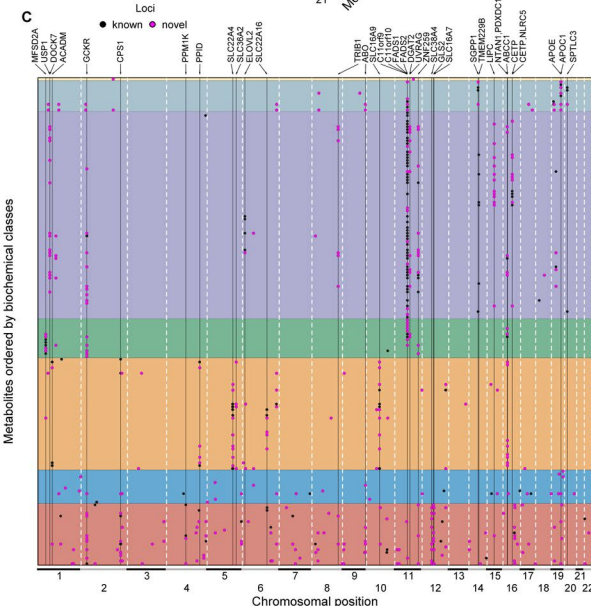
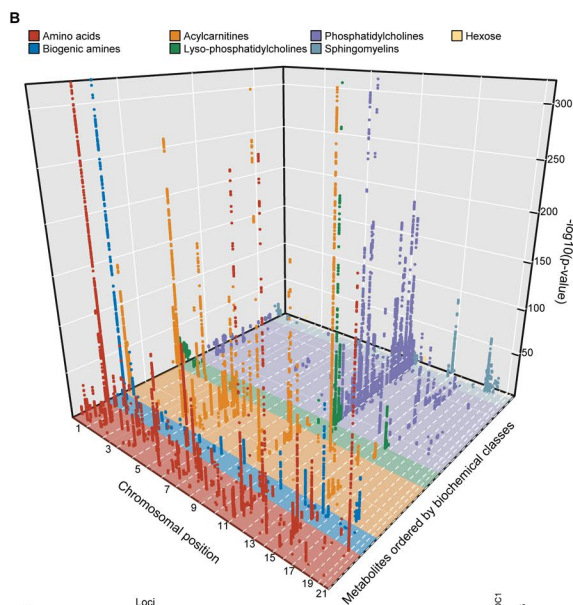
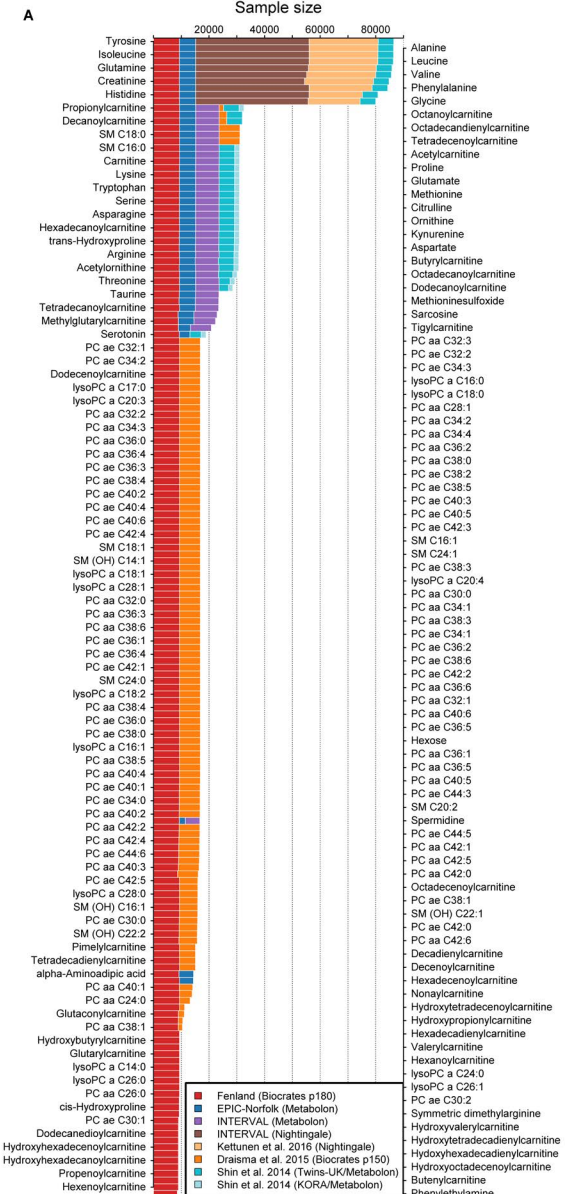
1093

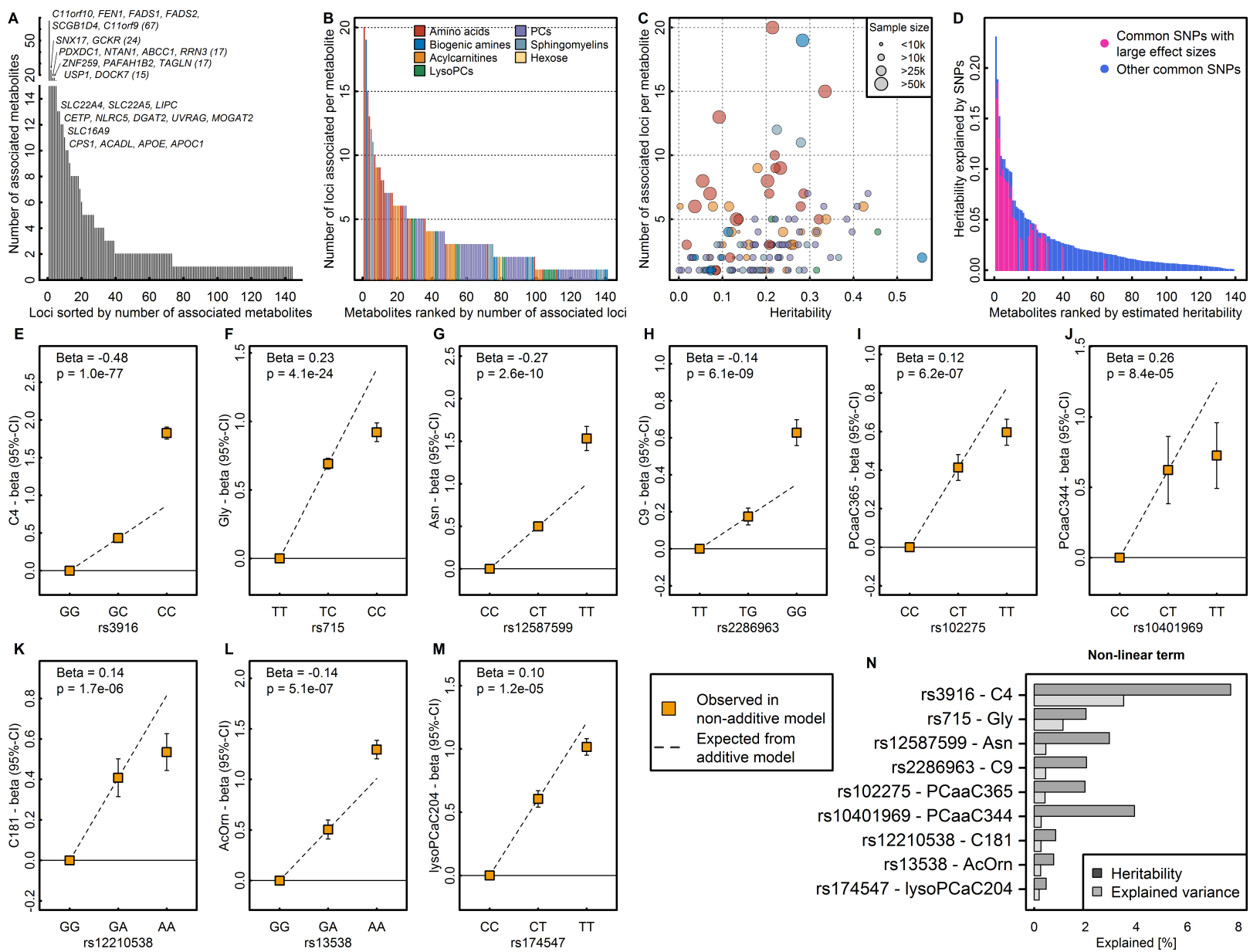
1094

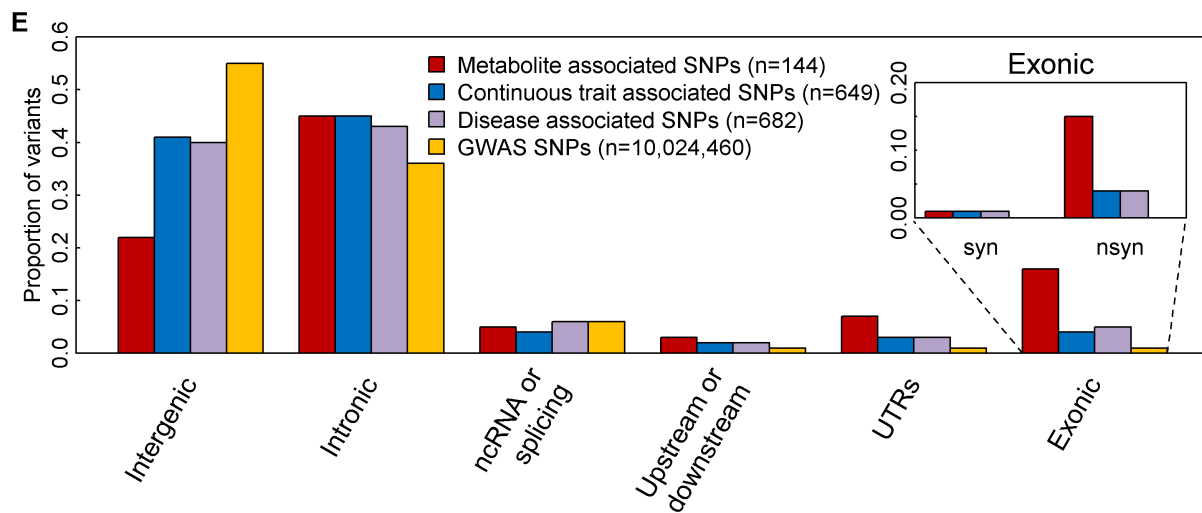
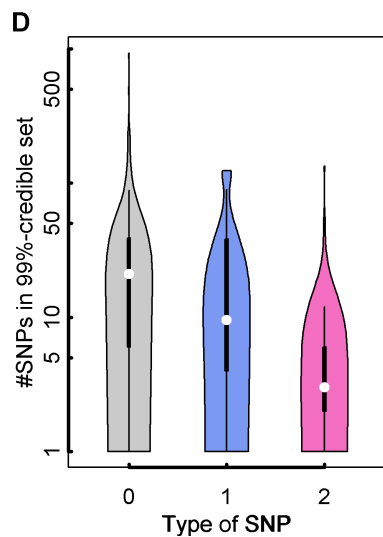
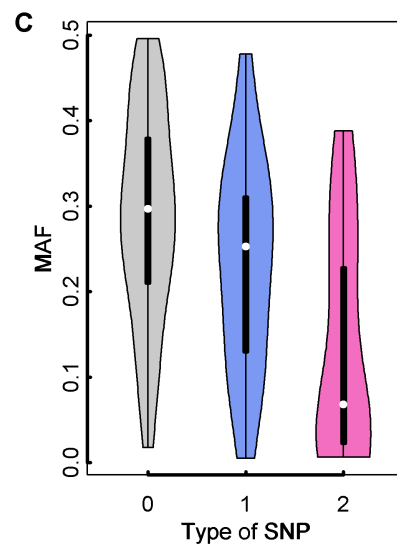
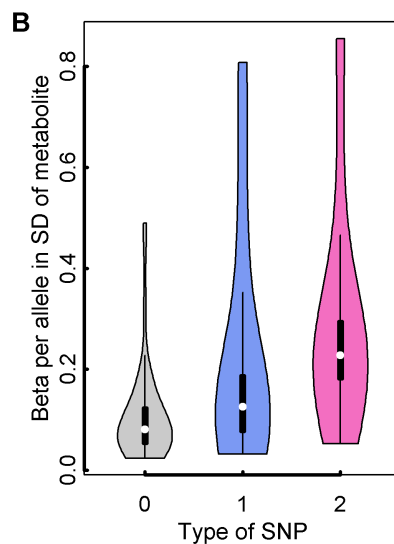
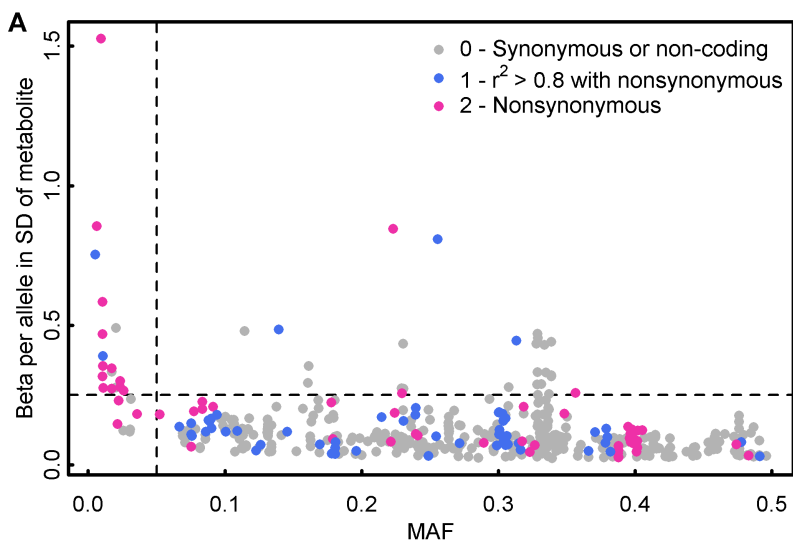
MR estimates are based on the inverse variance-weighted method using random effects to pool estimates. All single nucleotide polymorphisms (SNPs) significantly associated with either serine or glycine have been included in multivariable MR analysis. SD = standard deviation

1095

Sample size







A

Genetic prioritization approach

- Physically proximity
- Genetically predicted gene expression associated with metabolite
- Nonsynonymous variant is the lead or in LD with the lead variant

Biological knowledge-based approach

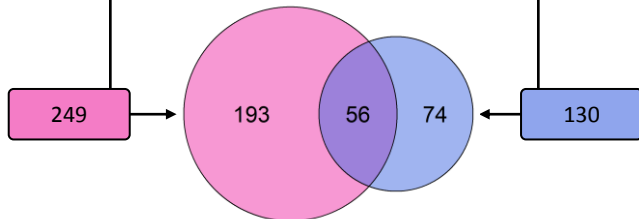
- Metabolite mapped to HMDB
- Genes in locus considered for relevance to metabolite
- Entrez genes
- Uniprot
- OMIM
- Go-terms
- KEGG

Number of potential causal genes

249

Number of potential causal genes

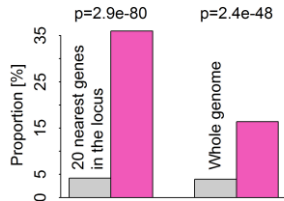
130



B

Biologically plausible genes

IEM genes



C

Genetically prioritized genes

Enzyme Transporter Both Other

