

1 **Genome-wide association and Mendelian randomization analysis**
2 **prioritizes bioactive metabolites with putative causal effects on common**
3 **diseases**

4
5 Youwen Qin^{1,2}, Guillaume Méric^{1,3}, Tao Long^{4,5}, Jeramie D. Watrous⁵, Stephen Burgess^{6,7},
6 Aki S. Havulinna⁸, Scott C. Ritchie^{1,9-11}, Marta Brożyńska¹, Pekka Jousilahti⁸, Markus
7 Perola⁸, Leo Lahti¹², Teemu Niiranen^{8,12}, Susan Cheng^{13,14}, Veikko Salomaa⁸, Mohit Jain⁵,
8 Michael Inouye^{1,2,7,9-11,15,16}

9
10 ¹Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria,
11 Australia

12 ²School of BioSciences, University of Melbourne, Melbourne, Victoria, Australia

13 ³Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria 3004,
14 Australia

15 ⁴Bioinformatics and Structural Biology Program, Sanford Burnham Prebys Medical Discovery Institute, La
16 Jolla, California, USA

17 ⁵Departments of Medicine and Pharmacology, University of California, San Diego, CA, USA

18 ⁶MRC Biostatistics Unit, Institute of Public Health, University of Cambridge, Cambridge

19 ⁷British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care,
20 University of Cambridge, Cambridge, UK

21 ⁸Department of Public Health Solutions, Finnish Institute for Health and Welfare, Helsinki, Finland

22 ⁹Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of
23 Cambridge, Cambridge, UK

24 ¹⁰British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

25 ¹¹National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and
26 Cambridge University Hospitals, Cambridge, UK

27 ¹²Department of Future Technologies, University of Turku, Turku, Finland

28 ¹³Division of Cardiology, Brigham and Women's Hospital, Boston, USA

29 ¹⁴Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

30 ¹⁵Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge,
31 UK

32 ¹⁶The Alan Turing Institute, London, UK.

33
34 * Corresponding author: MI (minouye@baker.edu.au or mi336@medschl.cam.ac.uk)

35 **Abstract**

36

37 Bioactive metabolites are central to numerous pathways and disease pathophysiology, yet
38 many bioactive metabolites are still uncharacterized. Here, we quantified bioactive
39 metabolites using untargeted LC-MS plasma metabolomics in two large cohorts (combined
40 $N \approx 9,300$) and utilized genome-wide association analysis and Mendelian randomization to
41 uncover genetic loci with roles in bioactive metabolism and prioritize metabolite features for
42 more in-depth characterization. We identified 118 loci associated with levels of 2,319 distinct
43 metabolite features which replicated across cohorts and reached study-wide significance in
44 meta-analysis. Of these loci, 39 were previously not known to be associated with blood
45 metabolites. Loci harboring *SLCO1B1* and *UGT1A* were highly pleiotropic, accounting for
46 $>40\%$ of all associations. Two-sample Mendelian randomization found 46 causal effects of
47 31 metabolite features on at least one of five common diseases. Of these, 15, including
48 leukotriene D4, had protective effects on both coronary heart disease and primary sclerosing
49 cholangitis. We further assessed the association between baseline metabolite features and
50 incident coronary heart disease using 16 years of follow-up health records. This study
51 characterizes the genetic landscape of bioactive metabolite features and their putative causal
52 effects on disease.

53

54 **Introduction**

55

56 The circulating metabolome reflects the compendium of metabolic compounds which operate
57 as inputs, intermediaries and/or outputs for the molecular pathways in blood which sustain or
58 detract from an individual's health¹. While many important metabolites have been identified
59 and their role(s) in human diseases characterized, there are thousands, potentially millions, of
60 as yet unidentified metabolites which may be causal or predictive of future health status.
61 Since metabolite identification, the defining of molecular composition and structure, typically
62 requires multiple costly technologies run on valuable biospecimens, there is a need for
63 principled and efficient prioritization of metabolites for targeted, resource-intensive follow-
64 up.

65

66 Genome-wide association studies (GWAS) have uncovered hundreds of genetic variants
67 associated with levels of circulating metabolites²⁻¹⁰, the vast majority of which are chemically
68 identified. The genes implicated by metabolite-associated genetic variants also show
69 evidence of a functional link between metabolite levels and complex diseases⁴. For instance,
70 the associations between fibrinogen A- α phosphorylation (FA α P) and three genetic loci
71 (*ABO*, *ALPL* and *FUT2*) support the role of FA α P as a biomarker for acute myocardial
72 infarction, as *ABO* and *ALPL* are also associated with coronary artery disease⁴. Furthermore,
73 Mendelian randomization (MR)¹¹ techniques have identified metabolites with evidence of
74 causal effects on disease risk⁷.

75

76 Recent advances in untargeted metabolomic profiling, especially liquid chromatography
77 mass-spectrometry (LC-MS), have opened the possibility of detecting and quantifying tens of
78 thousands of discrete chemical features from human blood samples¹². Our group has
79 pioneered chemical profiling of human plasma using an LC-MS/MS approach specifically
80 developed for profiling of small, polar lipophilic bioactive metabolites in human plasma¹³.
81 These approaches enable detection and relative quantitation of bile acids, sterols, free fatty
82 acids, polyunsaturated fatty acids and oxylipins (e.g. eicosanoids, docosanoids, resolvins,
83 etc), as well as thousands of related metabolites^{13,14}. These metabolites are sensed through
84 cell surface and nuclear hormone receptors and serve as critical signaling agents among
85 cellular pathways and systems, mediating a wide variety of processes including host
86 inflammatory response¹⁵⁻¹⁷, cellular development^{18,19}, and nutrient absorption^{20,21}. A number
87 of these metabolites have been implicated in the pathobiology of human diseases, including
88 autoimmune disease, cancer, metabolic disease / diabetes, and cardiovascular disease²²⁻²⁴.
89 Given their placement in key biological pathways, monitoring of bioactive lipids yields
90 unique insight into the current and future health of monitored individuals and, with likely
91 many yet undiscovered functions of these compounds, will allow for expansion of known
92 biology and discovery of new therapeutic strategies.

93

94 While only a minority of bioactive metabolites are chemically identified, their prioritization
95 via biological and clinical utility is under intense investigation. The integration of GWAS,
96 MR and untargeted metabolomics allows for an unbiased, wide-angle approach to prioritize
97 identified and unidentified metabolites, metabolite fragments and metabolic compounds (here
98 termed 'metabolite features') for chemical characterization. In the context of common
99 diseases, such an approach may highlight predictive biomarkers and causal metabolic factors
100 which have therapeutic potential.

101

102 In this study, we used untargeted high-throughput LC-MS metabolomics to quantify ~11,000
103 metabolite features in blood plasma samples of nearly 10,000 individuals with matched

104 genome-wide genotype data. We conducted a discovery GWAS, replication and meta-
105 analysis, then evaluated an extensive set of genetic loci with known and previously unknown
106 roles in metabolism. An ensemble two-sample MR approach was used to assess the predicted
107 causal effects of both identified and unidentified bioactive metabolites on 45 common
108 diseases. We then evaluated the effect of lifetime exposure of bioactive metabolite levels to
109 disease-free survival from baseline using ~16 years of electronic health record follow-up.
110 Taken together, this study presents a series of findings, including genome-wide maps of
111 bioactive metabolite associations and pleiotropy as well as an unbiased MR-based
112 prioritization of identified and unidentified metabolites with putative causal effects on
113 disease.

114
115

116 **Results**

117

118 **Genetic variants in 118 loci associate with levels of 2,319 circulating metabolite features**

119

120 We first performed a discovery genome-wide association scan on the abundances of 11,067
121 circulating bioactive metabolite features and 7,979,834 genotyped or imputed genetic
122 variants from 7,013 individuals in the population-based FINRISK02 cohort (**Figure 1A;**
123 **Methods**). Of these metabolite features, 818 were chemically identified metabolites (414
124 confirmed or putative eicosanoids¹¹, 100 free fatty acids (FFA), 66 polar compounds, 47
125 very-long-chain dicarboxylic acids (VLCDA), 47 fatty acids esters of hydroxy fatty acids
126 (FAHFA), 35 bile acids, 14 docosanoids, 14 endocannabinoids and 4 sterols). A further
127 10,249 metabolite features were, as yet, unidentified (**Table S1**). Stringent quality control
128 and normalization procedures were applied to both LC-MS and genetic data (**Methods**).
129 After the univariate genome-wide scan for each bioactive metabolite feature, a joint model
130 was fitted to all genome-wide significant genetic variants to identify a final set of
131 conditionally independent variants reaching significance together with their conditional effect
132 size estimates (**Methods**).

133

134 A majority of metabolite features (5,874, 57%) had at least one genetic variant reaching
135 genome-wide significance ($p < 5 \times 10^{-8}$) and were taken forward for replication in the
136 Framingham Heart Study (FHS, N=2,886). Of these associations, the vast majority (4,799)
137 were replicated in FHS at statistical significance ($p < 0.01$; **Table S2 & S3**). At a stringent
138 study-wide significance threshold accounting for the number of independent metabolite
139 features ($p < 1.45 \times 10^{-11}$; **Methods**), 3,602 SNP-metabolite associations were detected in
140 FINRISK02, replicated in FHS at $p < 0.01$, and maintained study-wide significance in meta-
141 analysis (**Table S2**). A set of suggestive associations reaching genome-wide but not study-
142 wide significance is given in **Table S3**.

143

144 This final set of 3,602 SNP-metabolite associations passing study-wide significance
145 ($p < 1.45 \times 10^{-11}$) comprised 2,319 metabolite features (161 chemically identified) and 118
146 genetic loci, of which 39 loci had no previously reported association with blood metabolites
147 (**Figure 1B, Table S4**). Of the 118 genetic loci, 75% were associated with multiple
148 metabolite features, and 31% of metabolite features were associated with multiple genetic
149 loci (**Table S5 & S6**).

150

151 Of the 265 associations with chemically identified metabolites, 152 involved eicosanoids
152 (**Table S2**). The strongest associations were between levels of dihydrotestosterone (putative
153 5α -androstane-17 β -ol-3-one) and a variant proximal to *SLC22A8* and *SLC22A24* (rs17713514;

154 meta-analysis $p < 1 \times 10^{-300}$). *SLC22A24* encodes a solute carrier that has recently been linked
155 to the transport of steroid conjugates²⁵, and *SLC22A8* (encoding OAT3) is a recently
156 identified metabolite QTL⁹ that has also been reported in functional experiments to be
157 associated with a wide range of metabolites, including bile acids, flavonoids, nutrients, amino
158 acids and lipids²⁶. In addition, levels of dihydrotestosterone were independently associated
159 with variants mapping to *SLC22A9* (rs147394024, meta-analysis $p = 1.2 \times 10^{-48}$) and *SRD5A2*
160 (rs2208532, meta-analysis $p = 7.2 \times 10^{-26}$), which itself encodes 3-oxo-5 α -steroid 4-
161 dehydrogenase-2. Furthermore, tetrahydrocortisol levels were associated with variant
162 (rs9994887, meta-analysis $p = 1.2 \times 10^{-28}$) proximal to *UGT2B15*, encoding UDP-
163 glucuronosyltransferase 2B15, an enzyme involved in the catabolism of xenobiotic
164 compounds and the metabolism of androgens²⁷.

165

166 Loci previously reported to be associated with blood metabolites showed diverse associations
167 with chemically identified and unidentified metabolite features, and were enriched for
168 pleiotropy as compared to loci without a previous reported metabolite association ($p < 0.0001$,
169 Wilcoxon rank-sum test) (**Figure 1C & 1D, Table S5**). Novel loci were associated with 13
170 identified (including 5 eicosanoids) and 305 unidentified metabolite features (**Table S6**),
171 indicating a sizable component of uncharacterized but genetically controlled metabolism.
172 These findings raised a series of hypotheses.

173

174 **Previously unreported loci associated with bioactive metabolites**

175

176 Two loci *GLYATL2* (11q12.1) and *GLYATL3* (6p12.3) encode glycine N-acyltransferases
177 which showed the strongest associations with 3 chemically unidentified metabolite features
178 (**Figure 1B, Table S6**). Importantly, the *ABCC3* (17q21.33) locus harbored the greatest
179 number of associations with 111 chemically unidentified metabolite features, as well as four
180 identified metabolites (leukotriene D4, putative tauroursodeoxycholic acid, putative 11 α -
181 Hydroxyprogesterone b-D-glucuronide, and putative 1,3,5(10)-Estratrien-3,17b-diol
182 diglucosiduronate). *ABCC3* encodes an ATP-binding cassette transporter, multidrug
183 resistance-associated protein 3 (MRP3), known to interact with metabolites through its role as
184 a multidrug exporter, in particular via efflux of potentially toxic endogenous and exogenous
185 compounds from the cell^{28,29}. Leukotriene D4 is an endogenous compound derived
186 extracellularly from leukotriene C4, whose cellular release is mediated by MRP3^{30,31}.
187 Tauroursodeoxycholic acid is a taurine conjugated compound converted from
188 ursodeoxycholic acid, a secondary bile acid synthesized in liver. Previous studies have
189 revealed that MRP3 is involved in the liver regeneration by pumping out excessive bile
190 acids^{32,33}. In addition, mouse study revealed that MRP3 involved in the transport of
191 glucuronidated compounds³⁴. 1,3,5(10)-Estratrien-3,17b-diol is a major form of estrogen in
192 human. It has been reported that 17 β -Glucuronosyl estradiol is a substrate of MRP3^{27,35} and
193 ethynylestradiol (a synthetic estrogen) increases expression of MRP3 in a rat model³⁶. Taken
194 together, our findings support the diverse roles of MRP3 in influencing the systemic
195 metabolism and human health.

196

197 Furthermore, *UGT2B7* (UDP-Glucuronosyltransferase-2B7, 4q13.2) was associated with 26
198 metabolite features, including three identified metabolites (bilirubin, palmitoleic acid, and a
199 putative eicosanoid). These findings consistent with previous studies showing that UDP-
200 glucuronosyltransferase has a diverse set of unrelated endogenous substrates and regulates
201 metabolism^{37,38}. This locus was reported to be suggestively associated with plasma
202 metabolites³⁹; however the signal in our study was very strong (meta-analysis $p = 1.43 \times 10^{-213}$).

203

204 ***SLCO1B1* and *UGT1A* loci are highly pleiotropic**

205

206 *UGT1A* (2q37.1) and *SLCO1B1* (12p12.1) showed extensive pleiotropy, with the two loci
207 accounting for a combined 1,495 of the 3,602 total SNP-metabolite associations, and
208 covering 1,211 of the 2,319 metabolite features (**Figure 1D, Table S5 & S6**). Interestingly,
209 only 30 metabolite features (including one eicosanoid) showed associations with both loci.
210 *SLCO1B1* encodes the solute carrier organic anion transporter family member 1B1
211 (OATP1B1), which is specific to the liver and involved in the transport of various
212 compounds and drugs, including both statins and antibiotics, from the blood into the liver.
213 *SLCO1B1* has been linked to the circulation of several fatty acids as well as statin-induced
214 myopathy⁴⁰, where *SLCO1B1* variants (such as rs4149056) have been linked to statin
215 response and risk of heart failure⁴¹. Notably, we found that levels of leukotriene D4, a
216 vasoconstrictor eicosanoid, were strongly independently associated with two SNPs at the
217 *SLCO1B1* locus (rs4149056, meta-analysis $p=5.4\times 10^{-193}$; rs11045856, meta-analysis
218 $p=7.15\times 10^{-108}$; **Table S2**). Rs4149056 showed diverse associations with other eicosanoids,
219 palmitoleic acid, and unidentified metabolite features (**Table S2**). *UGT1A*, encoding UDP
220 glucuronosyltransferase 1 family polypeptide A cluster, is a complex locus of several UDP-
221 glucuronosyltransferases involved in glucuronidation and metabolism. This family of UDP-
222 glucuronosyltransferases are assembled via differential splicing of numerous exons^{42,43}. In
223 previous GWASs, variants in *UGT1A* have been shown to be associated with multiple
224 circulating metabolites, including bile pigments^{3,4,6,10}. Our findings were consistent with this
225 and indicated a systemic metabolite role for *UGT1A*.

226

227 **Cytochrome P450 loci**

228

229 Cytochrome P450 (CYP) enzymes are known to have diverse roles in the metabolism of both
230 endogenous and exogenous compounds, including drugs, eicosanoids, bile acids etc⁴⁴. In
231 total, 12 loci encoding seven CYP families were associated with 316 metabolite features in
232 meta-analysis, collectively accounting for 12% (433/3602) of total SNP-metabolite
233 associations (**Table S5 & S6**). Notably, the *CYP3A* subfamily locus (7q22.1) was associated
234 with the largest number of metabolite features, including 6 identified and 114 chemically
235 unidentified. The *CYP3A* subfamily is expressed in liver and gut and is known to metabolize
236 >120 commonly prescribed drugs⁴⁴. In addition, two loci *CYP2C19* and *CYP2C9* (10q23.33)
237 encoding the *CYP2C* enzyme family were together associated with over 113 metabolite
238 features (including 5 identified metabolites) (**Table S6**).

239

240 **Putative causal effects of metabolite features on disease**

241

242 To prioritize metabolite features using their evidence for causal effects on disease, we
243 utilized two-sample MR together with GWAS summary statistics for 45 common diseases
244 (**Table S7**). An ensemble of five different MR methods were used, and directionally-
245 consistent statistical significance from at least three methods was necessary to infer causality
246 (**Methods**). Given the frequency of pleiotropic effects, we also used a second step correction
247 using MR-PRESSO⁴⁵ to control for horizontal pleiotropic outliers. At least five genetic
248 instruments were required for a metabolite feature to be considered for causal inference
249 (**Methods, Table S8**). Of the 70 metabolite features meeting these criteria, 31 showed
250 evidence of causal effect on at least one disease (**Figure 2, Table S9**), and 30 of these were
251 robust after further correction with MR-PRESSO. The majority of metabolite features with
252 causal effects were for coronary heart disease (CHD) and primary sclerosing cholangitis
253 (PSC), with one metabolite feature each affecting risk of schizophrenia, bipolar disorder or

254 rheumatoid arthritis. Amongst the 31 metabolite feature levels, there were a wide range of
255 correlations with most exhibiting weak positive correlations as well as ~5 clusters of features
256 whose levels were in moderate-high correlation.

257

258 Of the 15 metabolite features with putative causal effects on both PSC and CHD, all were
259 inverse associations: lifetime exposure to low metabolite feature levels increased risk of CHD
260 and PSC, with the latter exhibiting a somewhat greater effect size (**Figure 2A**). Lifetime
261 exposure to elevated leukotriene D4, a basophil-secreted metabolite known to be involved in
262 the induction of smooth muscle contraction, vasoconstriction and vascular permeability⁴⁶,
263 was predicted to reduce risk of both PSC and CHD (**Figure 2B**). While evidence of a
264 pleiotropic effect (MR-Egger intercept=0.07, $p=0.02$) for PSC was observed, the causal
265 estimates were consistent across all testing methods. Leukotriene D4 was only weakly
266 correlated with other unidentified metabolite features (**Figure S2**), with metabolite IDs
267 627.3755_2.19225 (m/z ratio 627.3755), 480.2732_1.924, and 612.3843_2.670166 showing
268 similar or stronger causal effect sizes on risks of CHD and PSC (**Figure 2A**).

269

270 **Association of metabolite features with incident coronary heart disease**

271

272 Using the baseline bioactive metabolite features, we next assessed CHD-free survival for
273 incident CHD via the linked EHR data available in both cohorts (**Methods**). While PSC is
274 relatively rare in the population (with only 9 incident cases in FINRISK02), we were
275 powered to detect baseline metabolite feature associations with CHD (541 and 97 incident
276 CHD cases in FINRISK02 and FHS, respectively). Time-to-event Cox proportional hazards
277 models were used in both cohorts and then meta-analyzed (**Methods**).

278

279 Of the 19 metabolite features with putative causal effects on CHD, six were associated with
280 incident CHD at FDR-adjusted significance (**Table S10, Figure 3**), with an additional
281 metabolite feature associated at nominal significance ($p<0.05$). For these seven metabolite
282 features, there was opposing direction of effect for the MR-based lifetime exposure estimate
283 and time-to-event Cox model estimate, with the former being negative and the latter being
284 positive (**Figure 3**). The corresponding hazards ratios from the Cox models of each
285 metabolite feature in FINRISK02 and FHS were consistently positive in both cohorts (**Table**
286 **S10**).

287

288

289 **Discussion**

290

291 In this study, we investigated the genetic associations of over 10,000 bioactive metabolite
292 features and their relationships with common diseases. We identified 118 genetic loci
293 harboring variants robustly associated with the levels of 2,319 metabolite features, 91% of
294 which were chemically unidentified compounds, suggesting a largely unexplored reservoir of
295 genetic control of the circulating metabolome. We identified 39 genetic loci previously
296 unlinked to blood metabolites and highlighted loci with extensive pleiotropy for bioactive
297 metabolites. We found causal effects for multiple identified and unidentified bioactive
298 metabolites on diverse common diseases, and investigated the baseline relationship of
299 putatively causal metabolite features with incident coronary heart disease.

300

301 Our findings were consistent with known metabolic pathways and indicate potentially new
302 gene functions. First, membrane transporters play important role in homeostasis by regulating
303 the transcellular movement of solutes between body fluid compartments. The communication

304 of small-molecule substrates between cells requires the activity of both SLC (generally
305 influx) and ABC (efflux) transporters⁴⁷, of which SLC22A8 and SLCO1B1 are two major
306 transporters with particular relevance to drug compounds⁴⁷. Loci encoding *SLC22A8* and
307 *SLCO1B1* were associated with diverse metabolite features, indicating broad substrate
308 specificity. Second, although the role of ABCC3 (MRP3) in metabolism has been
309 established^{28,47}, to our knowledge genetic associations have not been previously reported.
310 *ABCC3*, a hepatocyte efflux pump for bilirubin, was associated with four identified bioactive
311 metabolites with previously known relationships to MRP3 as well as over 100 unidentified
312 metabolite features which may be part of these pathways or lead to novel metabolic roles for
313 MRP3. Third, our findings suggest the *UGT1A* and *SLCO1B1* loci have central roles in
314 bioactive metabolic pathways, accounting for ~40% of our study-wide significant
315 associations. Given the splicing complexity of the *UGT1A* locus and its propensity to encode
316 a diversity of UDP-glucuronosyltransferases, this may indicate genetic control of exon usage
317 and downstream enzymatic functions.

318

319 The combination of untargeted metabolomics, GWAS and MR detected a set of 31
320 metabolite features, most of which as yet are unidentified, which showed causal effects on
321 disease risk and which now may be prioritized for further investigation. These included 15
322 bioactive metabolite features with shared causal effects on both CHD and PSC. While there
323 has been little to link between CHD and PSC in the existing literature. Our findings indicate
324 that bioactive metabolites, including leukotriene D4, may comprise previously unknown
325 causal metabolic pathways modulating risk of both diseases. As a chronic autoimmune
326 disease, PSC is a progressive disease mainly associated with the hepatic system and
327 inflammation of the bile ducts, consistent with putative causal effects of leukotriene D4, an
328 inflammatory mediator known to be released by basophils⁴⁸. We would speculate that other
329 bioactive metabolites amongst the 15 affecting CHD and PSC risk are also mediators of
330 inflammation, a broad but recognized causal process underlying both diseases.

331

332 Notably, for bioactive metabolites associated with CHD, we found directional inconsistency
333 between baseline risk prediction models and risk conferred by lifetime exposure (derived
334 from MR). This adds to previous evidence for significant but directionally opposed
335 biomolecular associations with cardiovascular diseases in MR and observational analyses. A
336 recent proteome analysis found that genetic predisposition to higher plasma MMP-12 levels
337 was predicted to reduce risk of coronary disease and atherosclerotic stroke, despite
338 observational studies finding a positive association with cardiovascular disease risk⁴⁹. A
339 separate study found levels of PON1, a major anti-atherosclerotic component of high-density
340 lipoprotein (HDL), to be negatively associated with CHD in observational analyses but with
341 MR results predicting higher PON1 to increase risk of CHD⁵⁰. Authors hypothesized that the
342 CHD condition may be linked to a downregulation of PON1 resulting in lower plasma
343 proteins.

344

345 These findings appear to be robust yet puzzling. We hypothesize several scenarios which
346 may explain these data. First, the causal effect estimated using MR is thought to indicate a
347 lifetime average effect^{51,52}; however, biological pathways are highly dynamic and
348 biomolecules may have age or time-varying effects. Second, sub-clinical atherosclerosis or
349 other relevant disease states may rewire biochemical and metabolic pathways such that
350 disease risk is increased but normal biological functions (as estimated by MR) are
351 compromised. Third, MR approaches assume linearity for the examined causal effect and for
352 the genetic effects on the exposure and the outcome⁵³. However, biological risk factors may
353 have non-linear effects on the outcome, For example, high body mass index (BMI) is a risk

354 factor for type 2 diabetes (T2D) in adults; however, high BMI can be protective in infants⁵⁴
355 and early childhood famine is associated with higher T2D risk in later life⁵⁵.

356

357 In conclusion, this study shows the efficiency of coupling untargeted metabolomics, GWAS
358 and MR to prioritize the bioactive molecular features with causal effects on disease. In doing
359 so, it highlights loci harboring potentially key enzymes, uncovers new insights into bioactive
360 molecular pathways, and raises salient questions of observational effects and MR-based
361 lifetime effects of molecular exposures which may have important therapeutic implications.

362 **Materials and Methods**

363

364 **Study cohorts**

365

366 FINRISK is a nation-wide population-based study of Finland periodically recruiting
367 participants every 5 years since 1972⁵⁶. Participants included in this study were randomly
368 recruited from six defined geographical areas across Finland in 2002⁵⁷. Blood samples were
369 taken during their visits. The follow-up data until 2018 were extracted from Finnish national
370 hospital discharge registries, drug reimbursement registries and causes-of-death registries. In
371 summary, there were 8,738 individuals, 4,688 (54%) females and 4,050 (46%) males, with
372 baseline age between 24 to 75 years.

373

374 The Framingham Heart Study (FHS) is a multi-generational population-based study. The data
375 used in this study was from the FHS offspring exam 8 participants⁵⁸. This subset of the FHS
376 cohort included 2,886 individuals, with average age of 66.4 (SD=9.0) years old, average BMI
377 of 28.3 (SD=5.4) kg/m², and 54% were female.

378

379 The FINRISK 2002 survey has been approved by the Ethical Committee on Epidemiology
380 and Public Health of the Helsinki and Uusimaa Hospital District (decision number 87/2001)
381 and the participants have provided an informed consent. The study is conducted according to
382 the World Medical Association's Declaration of Helsinki on ethical principles. The FHS
383 study data was accessed via dbGaP (approved study 2014.2023).

384

385 **Metabolomic profiling with liquid chromatography mass spectrometry**

386

387 Plasma samples were randomly assigned to 90 plates and measured by LC-MS (Thermo
388 Vanquish UPLC coupled to a Thermo Q Exactive Orbitrap mass spectrometer) in a
389 randomized order. Batch effects and machinery/chemical variance were assessed by 19 spike-
390 in internal standards in each measured sample and three pooled plasma control samples per
391 96-well plate. In-house R scripts were used to process the data including initial bulk feature
392 alignment, MS1-MS3 data parsing, pseudo DIA-to-DDA MS2 deconvolution, and CSV-to-
393 MGF file generation. Subsequently, mzMine 2.21⁵⁹ was employed for feature extraction,
394 secondary alignment and compound identification.

395

396 Metabolite features were filtered if they met any of the conditions: (1) less than 5
397 observations per plate, (2) a difference in observation missingness of 50% or greater from the
398 median plate missingness, (3) a plate standard deviation that differed from the median plate
399 standard deviation by a factor of 2.5 or greater, (4) a relative median offset of 4 or greater,
400 and (5) overall missingness >50%. This filtering resulted in an intensity profile of 11,067
401 measured metabolite features across 8,291 samples.

402

403 For FINRISK02, a two-step normalization was applied to the intensity data to control for
404 technical variation. The first step was to remove variation observed in external bracket
405 pooled plasma with RUV⁶⁰, a tool implemented in R package *MetNorm* to remove unwanted
406 variation of metabolomics data. The second step was to remove variation observed in internal
407 standards with RUV. There were two components for each RUV process, RUV-random
408 (function *NormalizeRUVRand*) and RUV-Kmeans (function *NormalizeRUVRandClust*).
409 RUV-random was to remove k technical factors estimated from external pooled plasma or
410 internal standards, where k was set to the number of principal components explaining >97.5%
411 of variation. RUV-Kmeans was used to refine the k technical factors to better represent the

412 biological variation. Following RUV, the plate median was subtracted from the resulted
413 matrix to get normalized matrix. Samples which exhibited excessive missing data of >10%
414 were removed (N=449), and K-nearest neighbors (R package *impute*)⁶¹ was used to impute
415 missing data into the remaining samples. Post-QC normalized metabolite feature data was
416 standardized to mean 0 and standard deviation of 1. The final intensity matrix consisted of
417 11,067 metabolite features and 7,842 samples.

418

419 For FHS, normalization of the metabolite intensity values included a four-step process: (1)
420 shift to positive values, (2) cap outliers, (3) log transformation, and (4) scale to a normal
421 distribution. Metabolite features presented in >20% individuals were kept for the analysis.

422

423 **Genotyping, imputation and quality control**

424

425 For FINRISK02, genotyping was performed on Illumina genome-wide SNP arrays (the
426 HumanCoreExome BeadChip, the Human610-Quad BeadChip and the HumanOmniExpress)
427 and has been described previously^{62,63}. Stringent criteria were applied to remove samples and
428 variants of low quality. Samples with call rate <95%, sex discrepancies, excess
429 heterozygosity and non-European ancestry were excluded. Variants with call rate <98%,
430 deviation from Hardy-Weinberg Equilibrium ($p < 1 \times 10^{-6}$), and minor allele count < 3 were
431 filtered. Data was pre-phased by using Eagle2 v2.3⁶⁴. Imputation was performed using
432 IMPUTE2 v2.3.0⁶⁵ with two Finnish-population-specific reference panels: 2,690 high-
433 coverage whole-genome sequencing and 5,092 whole-exome sequencing samples. To
434 evaluate the imputation quality, we compared the sample allele frequencies with reference
435 populations and examined imputation quality (INFO scores) distributions. Imputed SNPs
436 with INFO >0.7 were kept for analysis.

437

438 Post imputation quality control was carried out by using plink v2.0⁶⁶. Samples with >10%
439 missing rate were removed. Individuals with extreme height or BMI values were further
440 excluded (31 individuals with height < 1.47m; 5 with BMI > 50 were removed). We also
441 removed 42 pregnant women since pregnancy is known to have dramatic changes to body
442 metabolism. Both genotyped and imputed SNPs were kept for analysis if they met the
443 following criteria: call rate >90%, no significant deviation from Hardy-Weinberg Equilibrium
444 ($p > 1.0 \times 10^{-6}$), and minor allele frequency >1%. The post-QC dataset comprised 7,013
445 individuals and 7,980,477 SNPs. SNPs with ambiguous A/T or C/G alleles were removed in
446 meta-analysis.

447

448 FHS genotyping was performed using "Affymetrix Nsp, Sty and 50K gene centric" arrays. To
449 impute genotypes for SNPs, short insertions and deletions (indels), and larger deletions that
450 were not genotyped directly but are available from the 1000 Genomes Project, imputation of
451 8,493,311 genetic variants was performed with Minimac3 using the 1000 Genomes Project
452 Phase I Integrated Release Version 3 Haplotypes (2010-11 data freeze, 2012-03-14
453 haplotypes).

454

455 **Genome-wide association analysis**

456

457 Univariate testing was performed using BOLT-LMM (v2.3.2)⁶⁷, a Bayesian mixed model.
458 The kinship matrix was constructed using 106,201 SNPs selected by pruning the hard-called
459 SNPs with $r^2 < 0.1$ (plink2 command *--indep-pairwise 1000 80 0.1*). Genetic principal
460 components were calculated using FlashPCA2⁶⁸ on the pruned SNPs. Leave-one-
461 chromosome-out (LOCO) analysis within BOLT-LMM was used to avoid proximal

462 contamination. The linear mixed model included the following covariates for FINRISK02:
463 genotyping batch, age, gender and top 10 genetic principal components; for FHS, the
464 covariates were age and sex. The average genomic inflation factor was 1.0067 (range 0.9776
465 to 1.0389) and a positive correlation (Pearson correlation coefficient was 0.57) was observed
466 between the genomic inflation factor and SNP-heritability (**Figure S1**). For those SNPs
467 reaching genome-wide significance ($p < 5 \times 10^{-8}$), to obtain conditionally independent SNP-
468 metabolite associations GCTA-COJO⁶⁹ was used to conduct step-wise conditional and joint
469 analysis on individual genotype data. Given the large number of metabolite features, the
470 number of effective tests was based on eigenvalue variance and estimated using
471 matSpDLite^{70,71}. For all 11,067 metabolite features, the number of effective tests was 3,450
472 thus the study-wide significant threshold for SNP-metabolite associations was $p < 1.45 \times 10^{-11}$.

473
474 Associations passing genome-wide significance in FINRISK02 were taken forward for
475 validation and meta-analysis in FHS, with the mapping of metabolite features between the
476 two cohorts being based on MS1 and MS2 spectra alignment of metabolites with similar
477 mass charge ratio and retention time in principle⁷². Of the metabolite features, 76%
478 (4474/5874) were matched to at least one feature present in >20% people in the FHS cohort.
479 For metabolites with multiple matches, the strongest associations (lowest p values) were kept.
480 Meta-analysis was performed using the inverse variance weighted method for fixed effects (R
481 package *meta*). Only SNP-metabolite associations reached $p < 5 \times 10^{-8}$ in discovery cohort and
482 $p < 0.01$ in replication cohort were meta-analyzed.

483
484 ANNOVAR⁷³ was used to annotate significant SNPs. SNPs were assigned to genetic loci
485 using the 200kb region flanking the top SNPs (i.e. lowest p value)⁷⁴. This aggregation process
486 started from the overall top SNP, followed by the second top SNP of the remaining SNPs and
487 so on, until there was no SNP left. Loci names were determined from the nearest genes to the
488 associated top SNPs. No merging was performed for neighboring loci, a particular SNP could
489 be assigned to two different loci if those were both located within 200kb of the SNP position.
490 As such, 41 SNPs in total were assigned to more than one locus; in these cases, we reported
491 both loci in **Table S4** but do not count them twice when summarizing results. A locus was
492 defined as novel if it was not located within 200kb of any previously reported variants
493 associated with blood metabolites in the GWAS Catalog repository (release 2020-01-27), or
494 in the latest table from Kastenmüller et al.⁷⁵ ([http://www.metabolomix.com/list-of-all-](http://www.metabolomix.com/list-of-all-published-gwas-with-metabolomics)
495 [published-gwas-with-metabolomics](http://www.metabolomix.com/list-of-all-published-gwas-with-metabolomics); last accessed 02/2020).

496 497 **Mendelian randomization**

498
499 For genetic instruments, we utilized SNPs reaching genome-wide significance in meta-
500 analysis. Summary statistics of SNP-disease associations were extracted from MR-base
501 database using R package *TwoSampleMR*⁷⁶. If an index SNP was not present, the strongest
502 proxy SNP was used or set to missing if $r^2 < 0.8$). GWAS summary statistics were required to
503 include European ancestries and be based on at least 1,000 individuals (**Table S7**). For each
504 metabolite-disease analysis, SNPs were clumped using a linkage-disequilibrium threshold of
505 $r^2 < 0.05$ in a 500kb window to minimize the impact of correlated SNPs on causal estimates.
506 As MR analysis with multiple instruments is more reliable, five or more genetic instruments
507 were required for a metabolite to be taken forward for MR analysis. The effect allele was
508 taken to be the effect-increasing allele of metabolite in FINRISK02. We estimated causal
509 effects using an ensemble of five widely utilized methods: inverse variance weighted
510 (IVW)⁷⁷, simple mode⁷⁸, weighted mode⁷⁸, weighted median⁷⁹, and MR-Egger⁸⁰. As these
511 methods have different assumptions, agreement among multiple methods would indicate a

512 robust estimate of causal effects⁸¹. We defined a significant causal effect as $p < 0.05$ in three of
513 the five selected methods. For significant causal estimates, details of genetic instruments are
514 provided in **Table S8**. As a second step, for putative causal associations passing three of the
515 five methods in the ensemble, we then used MR-PRESSO⁴⁵ to detect and correct for
516 horizontal pleiotropic outliers.

517

518 **Cox proportional hazards models**

519

520 To test the association between metabolite feature and incident CHD in both FINRISK02 (16
521 years follow-up; 541 incident events) and FHS (6 years follow-up; 97 incident events), Cox
522 proportional hazards regression (*coxph* function in the *survival* R package) was utilized to
523 predict the CHD event for metabolite feature. Metabolite levels were at log₁₀ scale.
524 Covariates included age, sex, and log-transformed BMI. Participants with prevalent CHD at
525 baseline were excluded. Cox models were sex-stratified with time-on-study as the time scale.
526 Fixed effect meta-analysis was conducted to combine the summary statistics from both
527 cohorts. Sensitivity analysis was conducted to ensure associations were robust to LDL-
528 cholesterol, smoking status, hypertension, diabetes as well as medications for lipid-lowering,
529 anti-hypertension and diabetes.

530

531

532

533 **Acknowledgements**

534

535 The FINRISK02 cohort was mainly funded by budgetary funds of the Finnish Institute for
536 Health and Welfare. Important additional funding has been obtained from the Finnish
537 Academy and domestic non-profit foundations. SCR was funded by the UK National Institute
538 for Health Research (NIHR) (Cambridge Biomedical Research Centre at the Cambridge
539 University Hospitals NHS Foundation Trust). LL is supported by the Academy of Finland
540 (grant n:o 295741). TN is supported by the Academy of Finland (grant n:o 321351), the
541 Finnish Medical Foundation, the Paavo Nurmi Foundation, and the Emil Aaltonen
542 Foundation. VS has been supported by the Finnish Foundation for Cardiovascular Research.
543 MI was supported by the Munz Chair of Cardiovascular Prediction and Prevention. This
544 study was supported by the Victorian Government's Operational Infrastructure Support (OIS)
545 program.

546

547 This work was supported by Health Data Research UK, which is funded by the UK Medical
548 Research Council, Engineering and Physical Sciences Research Council, Economic and
549 Social Research Council, Department of Health and Social Care (England), Chief Scientist
550 Office of the Scottish Government Health and Social Care Directorates, Health and Social
551 Care Research and Development Division (Welsh Government), Public Health Agency
552 (Northern Ireland), British Heart Foundation and Wellcome.

553

554 The funders had no role in study design, data collection and analysis, decision to publish, or
555 preparation of the manuscript. The views expressed in this manuscript are those of the
556 authors and not necessarily those of the NHS, the NIHR or the Department of Health and
557 Social Care.

558

559 **Conflicts of Interest**

560 VS has received honoraria from Novo Nordisk and Sanofi for consultations. He also has
561 ongoing research collaboration with Bayer Ltd (all unrelated to the present study).

562

563

564

565 **References**

566

- 567 1. Newgard, C.B. Metabolomics and Metabolic Diseases: Where Do We Stand? *Cell*
568 *Metab* **25**, 43-56 (2017).
- 569 2. Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of
570 metabolite profiles in human serum. *PLoS genetics* **4**, e1000282-e1000282 (2008).
- 571 3. Illig, T. *et al.* A genome-wide perspective of genetic variation in human metabolism.
572 *Nature genetics* **42**, 137-141 (2010).
- 573 4. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical
574 research. *Nature* **477**, 54-60 (2011).
- 575 5. Inouye, M. *et al.* Novel Loci for Metabolic Networks and Multi-Tissue Expression
576 Studies Reveal Genes for Atherosclerosis. *PLOS Genetics* **8**, e1002907 (2012).
- 577 6. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nature*
578 *genetics* **46**, 543-550 (2014).
- 579 7. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci
580 and reveals novel systemic effects of LPA. *Nature Communications* **7**, 11122 (2016).
- 581 8. Long, T. *et al.* Whole-genome sequencing identifies common-to-rare variants
582 associated with human blood metabolites. *Nature Genetics* **49**, 568-578 (2017).
- 583 9. Yousri, N.A. *et al.* Whole-exome sequencing identifies common and rare variant
584 metabolic QTLs in a Middle Eastern population. *Nature Communications* **9**, 333
585 (2018).
- 586 10. Hong, M.G. *et al.* A genome-wide assessment of variability in human serum
587 metabolism. *Hum Mutat* **34**, 515-24 (2013).
- 588 11. Smith, G.D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology
589 contribute to understanding environmental determinants of disease? *Int J Epidemiol*
590 **32**, 1-22 (2003).
- 591 12. Alonso, A., Marsal, S. & Julià, A. Analytical Methods in Untargeted Metabolomics:
592 State of the Art in 2015. *Frontiers in Bioengineering and Biotechnology* **3**(2015).
- 593 13. Watrous, J.D. *et al.* Directed Non-targeted Mass Spectrometry and Chemical
594 Networking for Discovery of Eicosanoids and Related Oxylipins. *Cell Chem Biol* **26**,
595 433-442.e4 (2019).
- 596 14. Lagerborg, K.A., Watrous, J.D., Cheng, S. & Jain, M. High-Throughput Measure of
597 Bioactive Lipids Using Non-targeted Mass Spectrometry. *Methods Mol Biol* **1862**,
598 17-35 (2019).
- 599 15. Gilroy, D.W. *et al.* Inducible cyclooxygenase may have anti-inflammatory properties.
600 *Nature Medicine* **5**, 698-701 (1999).
- 601 16. Tilley, S.L., Coffman, T.M. & Koller, B.H. Mixed messages: modulation of
602 inflammation and immune responses by prostaglandins and thromboxanes. *The*
603 *Journal of Clinical Investigation* **108**, 15-23 (2001).
- 604 17. Norris, P.C. & Dennis, E.A. A lipidomic perspective on inflammatory macrophage
605 eicosanoid signaling. *Advances in Biological Regulation* **54**, 99-110 (2014).
- 606 18. Barquissau, V. *et al.* Control of adipogenesis by oxylipins, GPCRs and PPARs.
607 *Biochimie* **136**, 3-11 (2017).
- 608 19. Yeung, J., Hawley, M. & Holinstat, M. The expansive role of oxylipins on platelet
609 biology. *Journal of Molecular Medicine* **95**, 575-588 (2017).

- 610 20. Hofmann, A.F. THE FUNCTION OF BILE SALTS IN FAT ABSORPTION. THE
611 SOLVENT PROPERTIES OF DILUTE MICELLAR SOLUTIONS OF
612 CONJUGATED BILE SALTS. *Biochemical Journal* **89**, 57-68 (1963).
- 613 21. Molinaro, A., Wahlström, A. & Marschall, H.-U. Role of Bile Acids in Metabolic
614 Control. *Trends in Endocrinology & Metabolism* **29**, 31-41 (2018).
- 615 22. Vona-Davis, L. & Rose, D.P. The Obesity-Inflammation-Eicosanoid Axis in Breast
616 Cancer. *Journal of Mammary Gland Biology and Neoplasia* **18**, 291-307 (2013).
- 617 23. Gilroy, D.W. *et al.* CYP450-derived oxylipins mediate inflammatory resolution.
618 *Proceedings of the National Academy of Sciences* **113**, E3240 (2016).
- 619 24. Nayeem, M.A. Role of oxylipins in cardiovascular diseases. *Acta Pharmacologica*
620 *Sinica* **39**, 1142-1154 (2018).
- 621 25. Yee, S.W. *et al.* Unraveling the functional role of the orphan solute carrier,
622 SLC22A24 in the transport of steroid conjugates through metabolomic and genome-
623 wide association studies. *PLOS Genetics* **15**, e1008208 (2019).
- 624 26. Bush, K.T., Wu, W., Lun, C. & Nigam, S.K. The drug transporter OAT3 (SLC22A8)
625 and endogenous metabolite communication via the gut-liver-kidney axis. *J Biol Chem*
626 **292**, 15789-15803 (2017).
- 627 27. Tchernof, A. *et al.* Expression of the androgen metabolizing enzyme UGT2B15 in
628 adipose tissue and relative expression measurement using a competitive RT-PCR
629 method. *Clin Endocrinol (Oxf)* **50**, 637-42 (1999).
- 630 28. Keppler, D. The roles of MRP2, MRP3, OATP1B1, and OATP1B3 in conjugated
631 hyperbilirubinemia. *Drug Metab Dispos* **42**, 561-5 (2014).
- 632 29. Gu, X. & Manautou, J.E. Regulation of hepatic ABCC transporters by xenobiotics
633 and in disease states. *Drug Metabolism Reviews* **42**, 482-538 (2010).
- 634 30. Hirohashi, T., Suzuki, H. & Sugiyama, Y. Characterization of the Transport
635 Properties of Cloned Rat Multidrug Resistance-associated Protein 3 (MRP3). *Journal*
636 *of Biological Chemistry* **274**, 15181-15185 (1999).
- 637 31. Leier, I. *et al.* The MRP gene encodes an ATP-dependent export pump for leukotriene
638 C4 and structurally related conjugates. *Journal of Biological Chemistry* **269**, 27807-
639 10 (1994).
- 640 32. Fernández-Barrena, M.G. *et al.* Lack of *Abcc3* expression impairs bile-
641 acid induced liver growth and delays hepatic regeneration after partial hepatectomy in
642 mice. *Journal of Hepatology* **56**, 367-373 (2012).
- 643 33. Chang, T.H., Hakamada, K., Toyoki, Y., Tsuchida, S. & Sasaki, M. Expression of
644 MRP2 and MRP3 during liver regeneration after 90% partial hepatectomy in rats.
645 *Transplantation* **77**, 22-7 (2004).
- 646 34. Zelcer, N. *et al.* Mice lacking *Mrp3* (*Abcc3*) have normal bile salt transport, but
647 altered hepatic transport of endogenous glucuronides. *J Hepatol* **44**, 768-75 (2006).
- 648 35. Zelcer, N., Saeki, T., Reid, G., Beijnen, J.H. & Borst, P. Characterization of Drug
649 Transport by the Human Multidrug Resistance Protein 3 (ABCC3). *Journal of*
650 *Biological Chemistry* **276**, 46400-46407 (2001).
- 651 36. Ruiz, M.L. *et al.* Ethynylestradiol increases expression and activity of rat liver MRP3.
652 *Drug Metab Dispos* **34**, 1030-4 (2006).
- 653 37. Ouzzine, M., Gulberti, S., Ramalanjaona, N., Magdalou, J. & Fournel-Gigleux, S. The
654 UDP-glucuronosyltransferases of the blood-brain barrier: their role in drug
655 metabolism and detoxication. *Frontiers in cellular neuroscience* **8**, 349-349 (2014).
- 656 38. Girard, C., Barbier, O., Veilleux, G., El-Alfy, M. & Bélanger, A. Human uridine
657 diphosphate-glucuronosyltransferase UGT2B7 conjugates mineralocorticoid and
658 glucocorticoid metabolites. *Endocrinology* **144**, 2659-68 (2003).

- 659 39. Rhee, Eugene P. *et al.* A Genome-wide Association Study of the Human Metabolome
660 in a Community-Based Cohort. *Cell Metabolism* **18**, 130-143 (2013).
- 661 40. Link, E. *et al.* SLCO1B1 variants and statin-induced myopathy--a genomewide study.
662 *N Engl J Med* **359**, 789-99 (2008).
- 663 41. Yu, B. *et al.* Loss-of-function variants influence the human serum metabolome. *Sci*
664 *Adv* **2**, e1600800 (2016).
- 665 42. Bellemare, J., Rouleau, M., Harvey, M., Têtu, B. & Guillemette, C. Alternative-
666 splicing forms of the major phase II conjugating UGT1A gene negatively regulate
667 glucuronidation in human carcinoma cell lines. *Pharmacogenomics J* **10**, 431-41
668 (2010).
- 669 43. Girard, H. *et al.* Genetic diversity at the UGT1 locus is amplified by a novel 3'
670 alternative splicing mechanism leading to nine additional UGT1A proteins that act as
671 regulators of glucuronidation activity. *Pharmacogenet Genomics* **17**, 1077-89 (2007).
- 672 44. Nebert, D.W. & Russell, D.W. Clinical importance of the cytochromes P450. *The*
673 *Lancet* **360**, 1155-1162 (2002).
- 674 45. Verbanck, M., Chen, C.Y., Neale, B. & Do, R. Detection of widespread horizontal
675 pleiotropy in causal relationships inferred from Mendelian randomization between
676 complex traits and diseases. *Nat Genet* **50**, 693-698 (2018).
- 677 46. Colazzo, F., Gelosa, P., Tremoli, E., Sironi, L. & Castiglioni, L. Role of the Cysteinyl
678 Leukotrienes in the Pathogenesis and Progression of Cardiovascular Diseases.
679 *Mediators of inflammation* **2017**, 2432958-2432958 (2017).
- 680 47. Nigam, S.K. What do drug transporters really do? *Nat Rev Drug Discov* **14**, 29-44
681 (2015).
- 682 48. Tietz-Bogert, P.S. *et al.* Metabolomic Profiling of Portal Blood and Bile Reveals
683 Metabolic Signatures of Primary Sclerosing Cholangitis. *Int J Mol Sci* **19**(2018).
- 684 49. Sun, B.B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79
685 (2018).
- 686 50. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively
687 causal genes and pathways for cardiovascular disease. *Nature Communications* **9**,
688 3268 (2018).
- 689 51. Holmes, M.V., Ala-Korpela, M. & Smith, G.D. Mendelian randomization in
690 cardiometabolic disease: challenges in evaluating causality. *Nat Rev Cardiol* (2017).
- 691 52. Labrecque, J.A. & Swanson, S.A. Interpretation and Potential Biases of Mendelian
692 Randomization Estimates With Time-Varying Exposures. *Am J Epidemiol* **188**, 231-
693 238 (2019).
- 694 53. Didelez, V. & Sheehan, N. Mendelian randomization as an instrumental variable
695 approach to causal inference. *Statistical Methods in Medical Research* **16**, 309-330
696 (2007).
- 697 54. Warrington, N.M. *et al.* Maternal and fetal genetic effects on birth weight and their
698 relevance to cardio-metabolic risk factors. *Nature Genetics* **51**, 804-814 (2019).
- 699 55. van Abeelen, A.F.M. *et al.* Famine Exposure in the Young and the Risk of Type 2
700 Diabetes in Adulthood. *Diabetes* **61**, 2255 (2012).
- 701 56. Borodulin, K. *et al.* Cohort Profile: The National FINRISK Study. *International*
702 *Journal of Epidemiology* **47**, 696-696i (2017).
- 703 57. Salosensaari, A. *et al.* Taxonomic Signatures of Long-Term Mortality Risk in Human
704 Gut Microbiota. *medRxiv*, 2019.12.30.19015842 (2020).
- 705 58. Tsao, C.W. & Vasan, R.S. Cohort Profile: The Framingham Heart Study (FHS):
706 overview of milestones in cardiovascular epidemiology. *International journal of*
707 *epidemiology* **44**, 1800-1813 (2015).

- 708 59. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular
709 framework for processing, visualizing, and analyzing mass spectrometry-based
710 molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
- 711 60. De Livera, A.M. *et al.* Statistical methods for handling unwanted variation in
712 metabolomics data. *Anal Chem* **87**, 3606-15 (2015).
- 713 61. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays.
714 *Bioinformatics* **17**, 520-5 (2001).
- 715 62. Kiiskinen, T. *et al.* Genomic prediction of alcohol-related morbidity and mortality.
716 *Translational Psychiatry* **10**, 23 (2020).
- 717 63. Mars, N. *et al.* Polygenic and clinical risk scores and their impact on age at onset and
718 prediction of cardiometabolic diseases and common cancers. *Nature Medicine* **26**,
719 549-557 (2020).
- 720 64. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium
721 panel. *Nature genetics* **48**, 1443-1448 (2016).
- 722 65. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and
723 accurate genotype imputation in genome-wide association studies through pre-
724 phasing. *Nat Genet* **44**, 955-9 (2012).
- 725 66. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-
726 based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
- 727 67. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power
728 in large cohorts. *Nat Genet* **47**, 284-90 (2015).
- 729 68. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of
730 Biobank-scale genotype datasets. *Bioinformatics (Oxford, England)* **33**, 2776-2778
731 (2017).
- 732 69. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary
733 statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-
734 75, s1-3 (2012).
- 735 70. Nyholt, D.R. A simple correction for multiple testing for single-nucleotide
736 polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* **74**, 765-9
737 (2004).
- 738 71. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues
739 of a correlation matrix. *Heredity (Edinb)* **95**, 221-7 (2005).
- 740 72. Kantz, E.D., Tiwari, S., Watrous, J.D., Cheng, S. & Jain, M. Deep Neural Networks
741 for Classification of LC-MS Spectral Peaks. *Anal Chem* **91**, 12407-12413 (2019).
- 742 73. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic
743 variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164-
744 e164 (2010).
- 745 74. McVean, G.A. *et al.* The fine-scale structure of recombination rate variation in the
746 human genome. *Science* **304**, 581-4 (2004).
- 747 75. Kastenmüller, G., Raffler, J., Gieger, C. & Suhre, K. Genetics of human metabolism:
748 an update. *Human Molecular Genetics* **24**, R93-R101 (2015).
- 749 76. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across
750 the human phenome. *eLife* **7**, e34408 (2018).
- 751 77. Burgess, S., Butterworth, A. & Thompson, S.G. Mendelian Randomization Analysis
752 With Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology* **37**,
753 658-665 (2013).
- 754 78. Hartwig, F.P., Davey Smith, G. & Bowden, J. Robust inference in summary data
755 Mendelian randomization via the zero modal pleiotropy assumption. *International
756 journal of epidemiology* **46**, 1985-1998 (2017).

- 757 79. Bowden, J., Davey Smith, G., Haycock, P.C. & Burgess, S. Consistent Estimation in
758 Mendelian Randomization with Some Invalid Instruments Using a Weighted Median
759 Estimator. *Genetic epidemiology* **40**, 304-314 (2016).
- 760 80. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid
761 instruments: effect estimation and bias detection through Egger regression. *Int J*
762 *Epidemiol* **44**, 512-25 (2015).
- 763 81. Burgess, S. *et al.* Guidelines for performing Mendelian randomization investigations
764 [version 1; peer review: 1 approved]. *Wellcome Open Research* **4**(2019).
- 765
- 766

767 **Figure legends**

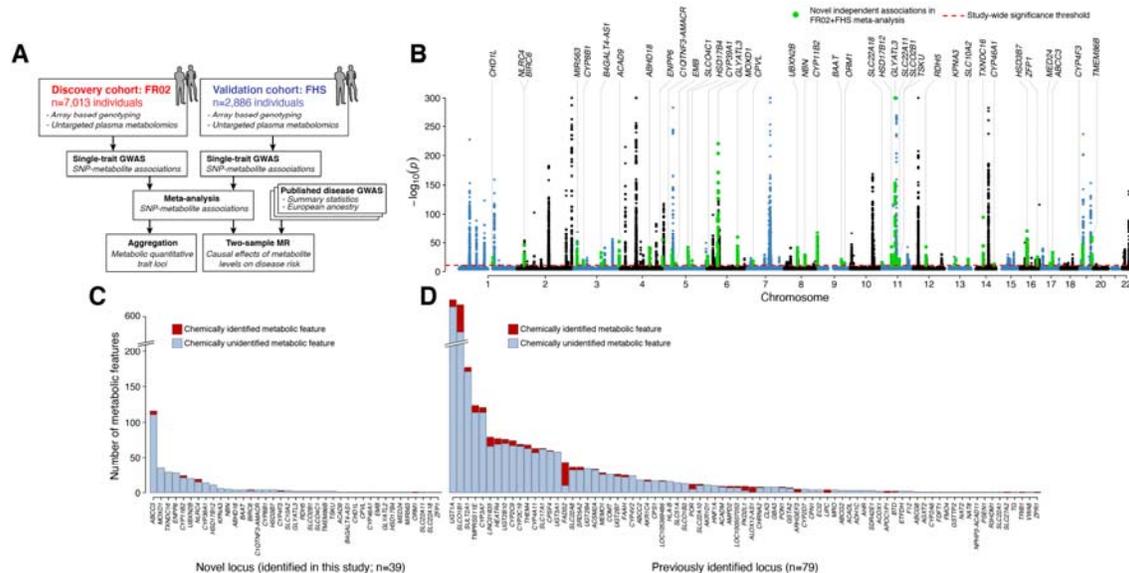
768

769 **Figure 1: Genome-wide association analysis of circulating metabolite features.** (A)
770 Summary of the workflow for this study, (B) Manhattan plot of associations between genetic
771 variants and circulating metabolite features. Novel loci are highlighted in green and annotated
772 with locus name. The red dashed line indicates study-wide significance ($p=1.45\times 10^{-11}$). For
773 SNPs with multiple associations, only the lowest p value is shown. The y-axis is truncated at
774 $-\log_{10}(p)=300$ for improved visualization, and SNPs with $p>10^{-4}$ are omitted. (C) Novel and
775 (D) previously reported loci associated with chemically identified (red bars) and chemically
776 unidentified (blue bars) metabolite features. (FHS: Framingham Heart study; FR02:
777 FINRISK02; GWAS: genome-wide association studies; MR: Mendelian randomization.)
778

779 **Figure 2. Causal effects of circulating metabolite features on common diseases.** (A)
780 Predicted causal (red) and protective (blue) effects reaching statistical significance in three of
781 the five MR methods tested. Causal effect estimates are from the weighted median method.
782 Asterisk (*) indicates causal effect which was not significant in MR-PRESSO outlier-
783 corrected test. (B) Dose-response plots of leukotriene D4 (LTD4) levels on coronary heart
784 disease and primary sclerosing cholangitis. Causal estimates of the five MR methods are
785 shown with confidence intervals indicated by shaded area in corresponding color.
786

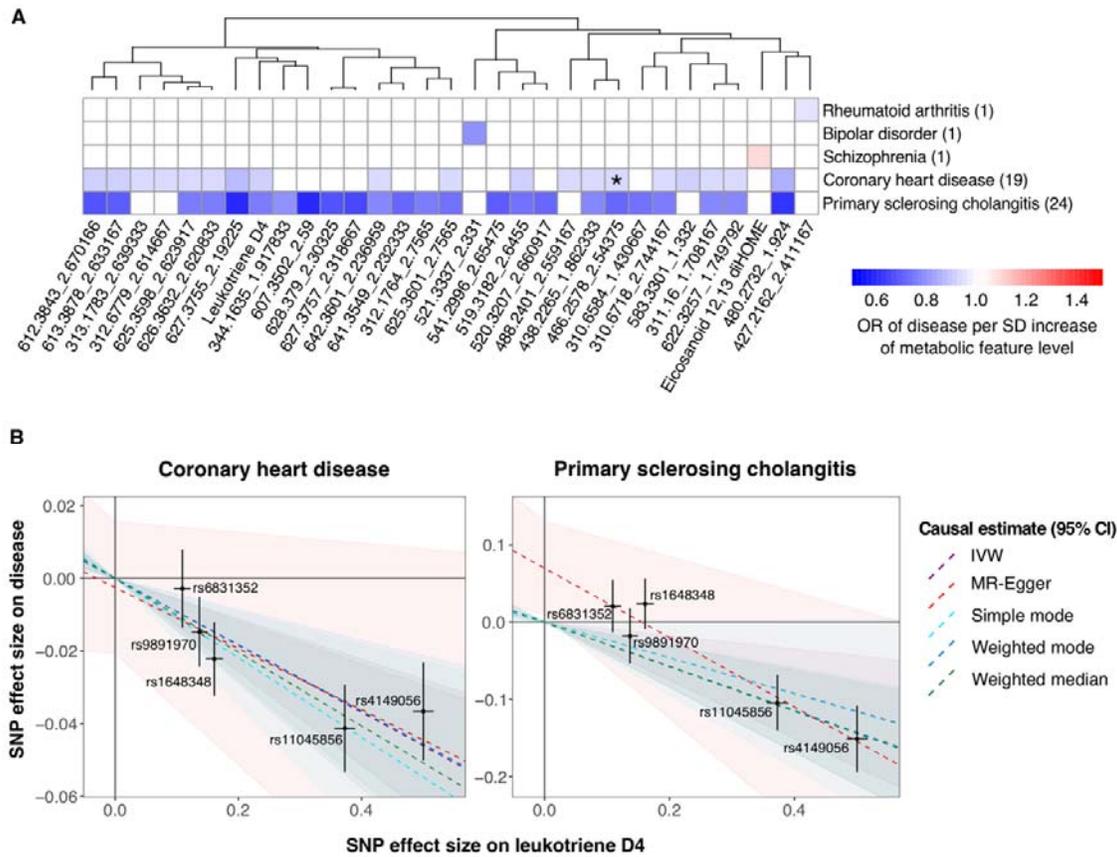
787 **Figure 3. Comparison of MR-based causal effects and association of baseline metabolite**
788 **feature levels with incident coronary heart disease risk.** A forest plot of MR-based effect
789 sizes (weighted median method) and hazard ratios were meta-analysis of Cox regression in
790 FINRISK02 and FHS. Effect sizes are per SD of metabolite feature.
791

792 **Figure 1: Genome-wide associations of circulating metabolite feature levels with known**
 793 **and previously unreported loci**
 794



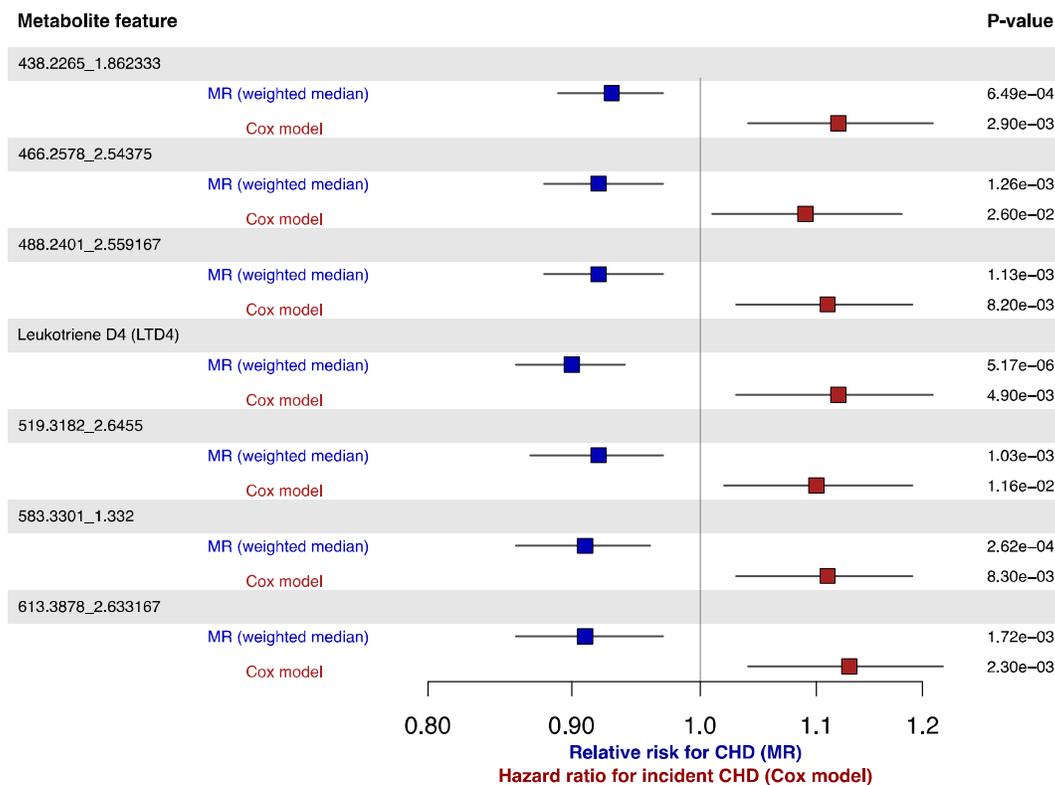
795
 796
 797

798 **Figure 2: Causal effects of circulating metabolite features on common diseases**
 799



800
 801

802 **Figure 3: Comparison of MR-based causal effects and association of baseline metabolic**
 803 **feature levels with incident coronary heart disease risk.**
 804



805
 806
 807