

30 Abstract

31

32 Bile acids are essential for food digestion and nutrient absorption, but also act as signalling
33 molecules involved in hepatobiliary diseases, gastrointestinal disorders and carcinogenesis.
34 While many studies have focused on the genetic determinants of blood metabolites, research
35 focusing specifically on genetic regulation of bile acids in the general population is currently
36 lacking. Here we investigate the genetic architecture of primary and secondary bile acids in
37 blood plasma, reporting associations with both common and rare variants. By performing
38 genome-wide association analysis (GWAS) of plasma blood levels of 18 bile acids ($N = 4923$)
39 we identify two significantly associated loci, a common variant mapping to *SLCO1B1*
40 (encoding a liver bilirubin and drug transporter) and a rare variant in *PRKG1* (encoding soluble
41 cyclic GMP-dependent protein kinase). For these loci, in the sex-stratified GWAS ($N_{\text{♂}} = 820$,
42 $N_{\text{♀}} = 1088$), we observe sex-specific effects (*SLCO1B1* $\beta_{\text{♂}} = -0.51$, $P = 2.30 \times 10^{-13}$, $\beta_{\text{♀}} = -$
43 0.3 , $P = 9.90 \times 10^{-07}$; *PRKG1* $\beta_{\text{♂}} = -0.18$, $P = 1.80 \times 10^{-01}$, $\beta_{\text{♀}} = -0.79$, $P = 8.30 \times 10^{-11}$),
44 corroborating the contribution of sex to bile acid variability. Using gene-based aggregate tests
45 and whole exome sequencing, we identify rare pLoF and missense variants potentially
46 associated with bile acid levels in 3 genes (*ORIG1*, *SART1* and *SORCS2*), some of which have
47 been linked with liver diseases.

48 **Introduction**

49 Bile acids (BAs) are synthesised from cholesterol in the liver and subsequently stored in the
50 gallbladder. After ingestion of food, BAs are secreted into the small intestine, where they
51 contribute to the digestion of lipid-soluble nutrients¹. Approximately 95% of BAs are then re-
52 absorbed by the intestinal epithelium and transported back to the liver via the portal vein - a
53 process termed “enterohepatic circulation”². Primary bile acids in humans consists of cholic
54 acid (CA), chenodeoxycholic acid (CDCA), and their taurine- or glycine-bound derivatives
55 (TCA and TCDCA, GCA and GCDCA). Once secreted in the lower gastrointestinal tract,
56 primary BAs are heavily modified by the gut microbiota to produce a broad range of secondary
57 BAs, with deoxycholic acid (DCA), a CA derivative, and lithocholic acid (LCA), a CDCA
58 derivative, being the most prevalent². Bile acids also act as hormone-like signalling molecules,
59 serving as ligands to nuclear (hormone) receptors. Through activation of these diverse
60 signalling pathways, BAs control not only their own transport and metabolism, but also lipid
61 and glucose metabolism, and innate and adaptive immunity³. Bile acids are thus involved in
62 regulating several physiological systems, such as fat digestion, cholesterol metabolism, vitamin
63 absorption, and liver function⁴. In addition, given their role in coordinating bile homeostasis,
64 biliary physiology and gastrointestinal functions, impaired signalling of BAs is associated with
65 development of hepatobiliary diseases, such as cholestatic liver disorders, cholesterol gallstone
66 disease and other gallbladder-related conditions⁵, and of inflammatory bowel disease⁶. Further,
67 bile acids have been implicated in carcinogenesis - specifically oesophageal, gastric,
68 hepatocellular, pancreatic, colorectal, breast, prostate and ovarian cancer - both as pro-
69 carcinogenic agents and tumour suppressors⁷. Thanks to their role as signalling molecules, BAs
70 have been considered as possible targets for the treatment of metabolic syndrome and various
71 metabolic diseases⁸. Further, BAs are able to facilitate and promote drug permeation through
72 biological membranes, making them of general interest for drug formulation and delivery⁹.

73 While many studies have focused on the genetic determinants of blood metabolites¹⁰⁻¹⁵,
74 research focusing specifically on bile acids in a large sample from the general population is
75 currently lacking. Here we investigate the genetic architecture of primary and secondary BAs,
76 reporting associations with both common and low-frequency/rare variants. First, we performed
77 a genome-wide association meta-analysis (GWAMA) of plasma blood levels of 18 BA traits
78 (N=4923). For a subset of this sample (female N=1088, male N=820), we perform sex-stratified
79 GWAMA, to describe sex-specific genetic contributions to BA variability. We then explore
80 whether complex traits or diseases have a role in influencing BA variability by using Mendelian
81 Randomisation. We finally employ multiple gene-based aggregation tests to investigate rare
82 (MAF < 5%) predicted loss of function (pLoF) and missense variants from whole exome
83 sequencing affecting the 18 BA traits in a subset of our cohorts (N=1006).

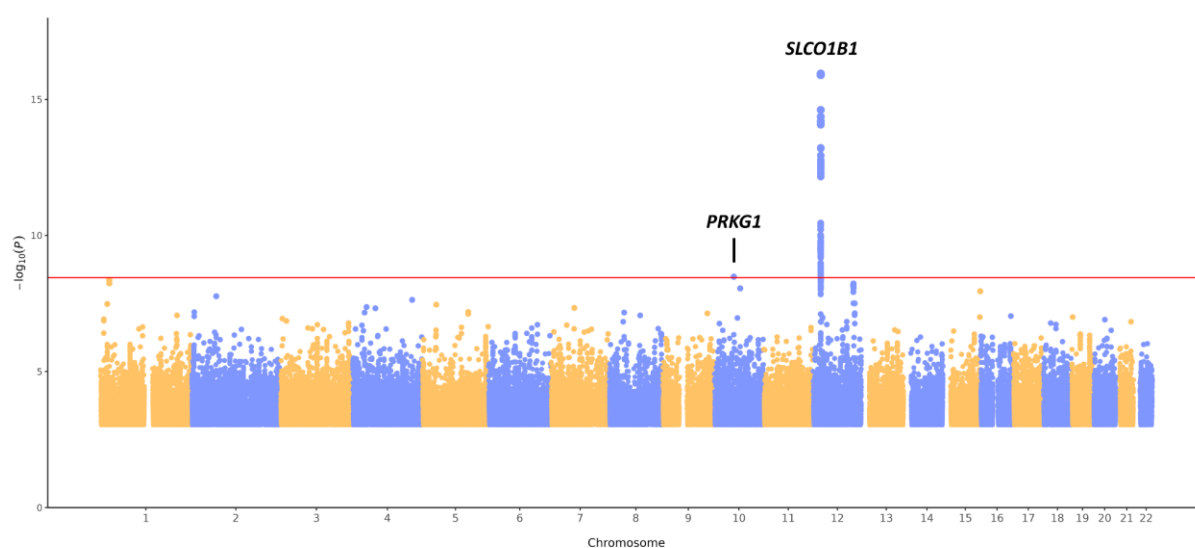
84 **Results**

85

86 **Loci associated with serum levels of bile acids**

87

88 To investigate the genetic control of bile acids, we performed a GWAS meta-analysis on five
89 cohorts of European descent (N = 4923), studying the associations of blood plasma levels of
90 18 primary and secondary bile acid traits with HRC-imputed genotypes/whole exome sequence
91 data. Based on the number of below limit-of-detection (LOD) measurements, BAs were
92 analysed either as quantitative or binary traits (Supplementary Table 1). In addition, two
93 analysis approaches were carried out in parallel for quantitative traits: in one case, <LOD
94 values were considered as missing, in the other case, they were imputed (Methods). An additive
95 linear model was assumed for each bile acid trait, followed by fixed-effect inverse-variance
96 meta-analysis. Overall, we identified 2 loci that passed the significance threshold (p-
97 value < 3.57×10^{-9} , Bonferroni adjusted for the number of independent bile acid traits)
98 (Figure1), near the *SLCO1B1* and *PRKG1* genes. The most strongly associated locus (p= 1.14
99 $\times 10^{-16}$), on chromosome 12 near *SLCO1B1*, showed consistent directionality across 4 of the 5
100 populations (Table 1), with the effect allele T of the sentinel SNP, rs4149056, associated with
101 decreased serum levels of GDCA (quantitative). In the same locus, we found GLCA and the
102 imputed GDCA trait to be significantly associated with the rs73079476 variant (Supplementary
103 Table 2), in high linkage disequilibrium with the sentinel SNP, rs4149056 ($r^2 = 0.97$). On the
104 other hand, rs146800892, the sentinel SNP on chromosome 10 near *PRKG1*, has a minor allele
105 frequency (MAF) lower than 1% in any cohort but CROATIA-Vis and might thus represent a
106 cohort-specific association with GCA (Supplementary Table 2).



107

108 **Figure 1. Summary Manhattan plot pooling together meta-analysis results obtained**
109 **across 18 bile acid traits.** The pooling was performed by selecting the lowest *p* value (y-axis)
110 from the 18 bile acids for every genomic position (x-axis). The Bonferroni-corrected genome-

111 wide significance threshold (horizontal red line) corresponds to 3.57×10^{-9} . For simplicity,
112 SNPs with p value $> 1 \times 10^{-3}$ are not plotted. P values are derived from the two-sided Wald test
113 with one degree of freedom.

114 **Table 1: Loci genome-wide significantly associated with at least one of the 18 bile acid traits in all samples and sex-stratified GWAMA**

All samples										
Locus	Gene	SNP	EA	OA	EAF	Beta	P	SE	N	Lead BA
12:20994540-21463812	<i>SLCO1B1</i>	rs4149056	T	C	0.839	-0.25	1.10x10 ⁻¹⁶	0.03	4547	GDCA
10:53832549-53832549	<i>PRKGI</i>	rs146800892	T	C	0.988	-0.96	3.30x10 ⁻⁰⁹	0.16	900	GCA
Sex-stratified										
Locus	Gene	SNP	EA	OA	Beta M	Beta F	P M	P F	N (M/F)	Lead BA
12:20994540-21463812	<i>SLCO1B1</i>	rs73079476	A	C	-0.51	-0.31	2.30x10 ⁻¹³	9.90x10 ⁻⁰⁷	820/1088	GDCA
10:53832549-53832549	<i>PRKGI</i>	rs117834398	T	G	-0.18	-0.79	1.80x10 ⁻⁰¹	8.30x10 ⁻¹¹	820/1088	GCA

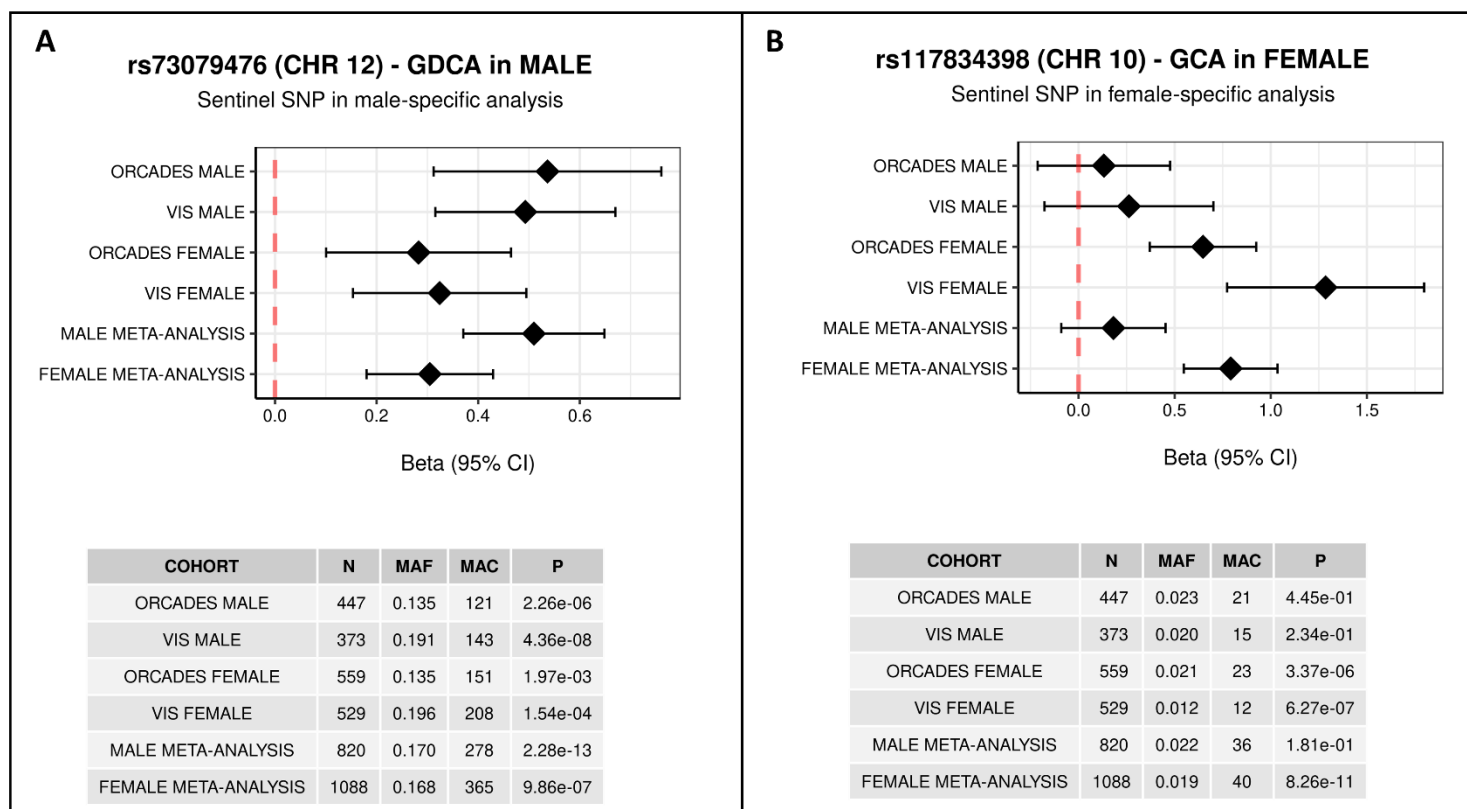
Each locus is represented by the SNP with the strongest association in the region, according to the p-value rejecting the null hypothesis of no association with at least one of 18 bile acid traits. In all samples analysis, an association was considered significant if the p-value was lower than 3.57×10^{-9} , the genome-wide significance threshold Bonferroni-corrected for the number of independent bile acid traits. In sex-stratified analysis, an association was considered significant if the p-value was lower than 5×10^{-9} , the genome-wide significance threshold Bonferroni-corrected for the number of independent bile acid quantitative traits. The two SNPs in the *SLCO1B1* locus are in high LD (LD $r^2 = 0.97$), while the two SNPs in the *PRKGI* locus represent two distinct signals (LD $r^2 < 0.001$).

Locus - coded as 'chromosome: locus start–locus end' (GRCh37 human genome build); Gene - suggested candidate gene; SNP - variant with the strongest association in the locus; EA - SNP allele for which the effect estimate is reported; OA - other allele; EAF - frequency of the effect allele; Beta - effect estimate for the SNP and bile acid with the strongest association in the locus; SE - standard error of the effect estimate, P - p-value of the effect estimate (two-sided Wald test with one degree of freedom); N - sample size; Lead BA - bile acid with the strongest association to the reported SNP; M - male specific analysis; F - female specific analysis.

115

116 Sex-specific associations of bile acid serum levels

117 To investigate whether the genetic component influencing bile acid variation may differ
118 between men and women, we performed sex-specific GWAS meta-analysis of the 14
119 quantitative (imputed) bile acid traits for ORCADES and CROATIA-Vis cohorts (female
120 N=1088, male N=820) and discovered two sex-specific associations. The association of GDCA
121 with rs73079476 from the *SLCO1B1* locus was significant in male-only GWAS (beta = -0.51,
122 p-value = 2.28×10^{-13}) (Table 1, Figure 2A). The signal for the same locus in female-only
123 GWAS, while consistent in terms of directionality, has a smaller effect size than in male-only
124 analysis (beta = -0.31) and does not reach the significance threshold (p-value = 9.86×10^{-7}),
125 despite the slightly higher sample size (Figure 2). This suggests that the genetic effect of
126 *SLCO1B1* locus on the serum levels of GDCA is larger in men than women. We also identified
127 a sex-specific association of GCA at the *PRKGI* locus. In contrast to *SLCO1B1*, the sentinel
128 SNP in *PRKGI*, rs117834398, has a larger effect in females than in males (female beta = -0.79,
129 male beta = -0.18), and passed the significant threshold only in the female-specific analysis
130 (female p-value = 8.26×10^{-11} , male p-value = 1.81×10^{-1}) (Table 1, Figure 2B). Interestingly,
131 the sentinel SNPs at the *PRKGI* locus for the full meta-analysis and for the female-specific
132 analysis are in linkage equilibrium ($r^2 < 0.01$) and represent two independent associations in that
133 locus. Overall, none of the significant association identified in one sex was replicated in the
134 other, suggesting that the genetic contribution to serum BA levels is likely to be different in
135 males and females. We have identified 13 additional associations (p-value $< 5 \times 10^{-9}$,
136 Bonferroni adjusted for the number of independent quantitative bile acid traits) that might have
137 sex-specific effects (Supplementary Table 3, Supplementary Figure 1). However, given the low
138 allele frequencies and allele counts in the two analysed cohorts, further analyses are required
139 to replicate these associations.



140 **Figure 2. Sex-specific associations.** The effect of rs73079476 on chromosome 12 on GDCA
 141 bile acid is almost as twice strong in males compared to the effect in females (Panel A). The
 142 effect of rs117834398 on GCA bile acid is stronger in females than in males (Panel B). N –
 143 sample size, MAF – minor allele frequency, MAC – minor allele count, CI – confidence
 144 interval.

145

146 **Link with complex traits and diseases**

147 Next, we assessed whether variants associated with BA levels have been previously associated
 148 with any other biochemical traits and diseases. Using Phenoscanner^{16,17} we found that
 149 rs4149056, sentinel SNP in *SLCO1B1* locus, and its proxies ($r^2 > 0.8$), were also associated
 150 with concentration of bilirubin, non-bile acid metabolites, mean corpuscular haemoglobin, sex
 151 hormone binding globulin and estrone conjugates, and various responses to drugs (i.e., statin-
 152 induced myopathy, LDL-cholesterol response to simvastatin and methotrexate clearance in
 153 acute lymphoblastic leukaemia) (Supplementary Table 4). To obtain deeper insight into the
 154 causal relationship between BAs and diseases, we conducted bi-directional Mendelian
 155 Randomisation (MR) analysis. Using the sentinel SNPs associated with GLCA, GDCA and
 156 GCA (Table 1) as instrumental variables we tested whether genetically increased levels of BA
 157 influence levels or risk for 548 biochemical traits and diseases available in the IEU Open
 158 GWAS database¹⁸ (Supplementary Table 5). Levels of GLCA and GDCA were significantly

159 (p-value $< 0.05/(548 \times 3) = 3.04 \times 10^{-5}$) associated with different biochemical measurements,
 160 such as levels of sex hormone-binding globulin, testosterone, triglycerides, vitamin D, alanine
 161 transaminase and galectin-3; with blood traits, such as mean corpuscular haemoglobin and
 162 mean corpuscular volume; and with diseases and their risk factors, such as daytime dozing and
 163 stroke (Supplementary Table 6). These MR tests were performed using the Wald ratio test
 164 utilising only a single instrument, thus the results of causal relationship between BAs and
 165 traits/diseases should be interpreted with caution. Yet our results suggest a possible overlap in
 166 genetic regulation, involving the *SLCO1B1* locus. Next, to assess whether complex traits and
 167 disease could have an effect on bile acid levels, we performed reverse MR using 548
 168 traits/diseases as exposure and bile acids as outcomes. We found no significant associations,
 169 suggesting that none of the tested diseases or complex traits have an effect on BA levels
 170 (Supplementary Table 7).

171

172 Exome-wide rare variant analysis of bile acids

173 To assess the contribution of low frequency and rare variants to the bile acid genetic
 174 architecture, we performed exome-wide gene-based tests across 18 bile acid traits in the
 175 ORCADES cohort (N = 1006) by testing the aggregated effect of rare (MAF $< 5\%$) predicted
 176 loss-of-function (pLoF) and non-synonymous missense variants. We identified significant
 177 association (p-value $< 1.79 \times 10^{-7}$) of rare variants from 3 genes with 2 bile acid traits
 178 (quantitative CA and binary THDCA). For these associations, a significant p-value was
 179 reported by at least 2 of the 4 aggregation tests used. Rare variants significantly associated with
 180 quantitative bile acid trait CA are located in the *ORIG1* gene, while those associated with
 181 binary bile acid trait THDCA are located in *SART1* and *SORCS2* genes (Table 2,
 182 Supplementary Table 8). We further identified significant association of rare variants from
 183 *EPS8L1* gene with quantitative bile acid trait DCA and from *EEF2K* with binary bile acid trait
 184 THDCA (Supplementary Table 8). However, a significant p-value was reported by only one of
 185 the 4 aggregation tests used. Due to the lack of replication across aggregations tests, we
 186 considered these associations as not robust.

187

188 **Table 2. Gene-based aggregation analysis results for bile acid traits in ORCADES cohort**

BA	Trait type	Gene	MAF	Functional consequence	N variants	Aggregation test	P	AC
CA	Quantitative	<i>ORIG1</i>	< 0.01	Missense variants	2	SKAT-O	1.67×10^{-8}	17
THDCA	Binary	<i>SORCS2</i>	< 0.05	pLoF and missense variants	10	SMMAT-E	1.44×10^{-8}	174
THDCA	Binary	<i>SART1</i>	< 0.01	pLoF and missense variants	4	SKAT	1.19×10^{-7}	25

BA- bile acid trait tested for rare variants association; Trait type – whether BA was analysed as a quantitative or binary trait; Gene - gene for which variants were aggregated; MAF - upper bound for minor allele frequency of tested variants; Functional consequence - predicted functional consequence for aggregated variants; N variants - number of variants in the mask; Aggregation test - rare-variants aggregation test reporting the lowest p-value out of 4 aggregation tests; P - p-value for the aggregation test; AC - cumulative allele count of all the variants in a mask. Bonferroni-corrected discovery p-value threshold was set to 1.79×10^{-7} ($0.05/20,000$ estimate of number of genes in the human genome/14 number of independent bile acids).

189

190 Discussion

191 Bile acids (BAs) are synthesised from cholesterol in the liver and then secreted into the small
192 intestine to emulsify and promote absorption of lipid-soluble nutrients. BAs also act as
193 hormone-like signalling molecules and have been linked to regulation of lipid and glucose
194 metabolism, immunity, vitamin absorption, hepatobiliary diseases, inflammatory bowel
195 disease and cancer. Despite the crucial role of BAs on whole-body physiology, their genetic
196 architecture has not been extensively investigated in a large sample from the general
197 population. In this study, we performed both pooled and sex-stratified genome-wide
198 association meta-analysis of plasma levels of 18 bile acid compounds, including both primary
199 and secondary forms, in 4923 European individuals.

200 We identified two secondary bile acids (GDCA and GLCA) significantly associated with a
201 locus encompassing the *SLCO1B1* gene. The encoded protein, OATP1B1 (organic anion
202 transporting polypeptide 1B1), is a well-known human hepatocyte transporter mediating the
203 uptake of various endogenous compounds such as bile salts, bilirubin glucuronides, thyroid
204 hormones and steroid hormone metabolites, and also clinically frequently used drugs like
205 statins, HIV protease inhibitors, and the anti-cancer agents irinotecan or methotrexate^{19–23}. The
206 sentinel SNP of the *SLCO1B1* locus, rs4149056, is a missense variant (p.Val174Ala) which
207 has been linked by previous GWA studies to blood concentration of several metabolites,
208 including vitamin D²⁴, triglycerides²⁵ and bilirubin²⁶, a compound resulting from the
209 breakdown of haem catabolism and excreted as a major component of bile. This same variant
210 has also been associated with levels of sex hormone-binding globulin and testosterone²⁷. The
211 knock-out of the gene in mice results in abnormal liver physiology and abnormal xenobiotic
212 pharmacokinetic phenotypes (Open Targets²⁸). A rare variant from the *PRKGI* locus was
213 significantly associated with levels of glycocholic acid (GCA). *PRKGI* encodes a Protein
214 Kinase CGMP-Dependent 1, a protein involved in signal transduction and a key mediator of
215 the nitric oxide/cGMP. The sentinel variant in the region, rs146800892, only passes the MAF
216 threshold (MAF > 0.01) in the CROATIA-Vis cohort, which is therefore the only cohort
217 contributing to this association. Due to its demographic history and geographic position,
218 CROATIA-Vis is a genetic isolate²⁹ so it is possible that this variant has increased in frequency
219 compared to a general population³⁰. The mechanism of how the variation within this gene could
220 relate to bile acid levels is unclear and would need to be further investigated.

221 In the sex-stratified GWAS meta-analysis, we observed sex-specific associations for the two
222 identified loci. Levels of glycodeoxycholic acid (GDCA) are more strongly associated with the
223 variant in *SLCO1B1* in men than in women, while female levels of GCA are more strongly

224 affected by the variant in *PRKGI* than male levels. Later, our Mendelian randomization
225 analysis did not provide evidence that testosterone, oestradiol, sex hormone-binding globulin
226 or other sex-related traits have causal effects on plasma BA levels. While this could be due to
227 a lack of statistical power of our BA meta-analysis, we currently have no evidence to suggest
228 an effect of sex-related hormones on BA levels mediated by genetics. We also detected
229 associations with variants from the same gene, *PRKGI*, in the main, non-stratified analysis.
230 However, the two associations (sex-specific and pooled) appear to be independent (LD r^2
231 <0.001). While the association from the pooled analysis might be either false positive or
232 population-specific, the independent association from the sex-stratified analysis replicates well
233 between two analysed cohorts (CROATIA-Vis and ORCADES).

234 After assaying common variants through GWAS, we performed exome-wide gene-based
235 association tests in a subset of our samples (N = 1006), to investigate the genetic contribution
236 of rare and low frequency (MAF $<5\%$) coding variants (pLoF and missense) to bile acid levels.
237 Overall, we identified associations with rare variants from 3 genes, *ORIG1*, *SART1* and
238 *SORCS2*. *ORIG1* is an olfactory receptor gene, whose coded protein receptor interacts with
239 odorant molecules in the nose to initiate a neuronal response triggering the perception of
240 smell^{31,32}. In addition to the nasal level, the olfactory receptor coded by *ORIG1* is expressed
241 also by enterochromaffin cells, specialised enteroendocrine cells of the gastrointestinal tract.
242 Braun *et al*³³, determined that certain olfactory cues from spices and odorants, such as thymol,
243 present in the luminal environment of the gut may stimulate serotonin release via olfactory
244 receptors present in enterochromaffin cells. Between 90% and 95% of total body serotonin is
245 in fact synthesised by enterochromaffin cells³⁴: serotonin controls gut motility and secretion
246 and is implicated in pathologic conditions such as vomiting, diarrhoea, and irritable bowel
247 syndrome³³. In mice, gut serotonin was shown to stimulate bile acid synthesis and secretion by
248 the liver and gallbladder. Thus, release of serotonin in response to odorant cues increases bile
249 acid turnover³⁵. The hypoxia-associated factor (HAF), encoded by *SART1* gene and also known
250 as SART1(800), is involved in proliferation and hypoxia-related signalling. The protein
251 encoded by *SORCS2* is a receptor for the precursor of nerve growth factor, up-regulation of
252 which has been reported for several liver pathologies, such as hepatotoxin- induced fibrosis³⁶,
253 ischemia-reperfusion injury³⁷, oxidative injury³⁸, cholestatic injury³⁹ and hepatocellular
254 carcinoma^{36,40,41}. However, due to unavailability of exome sequencing data in other cohorts
255 these associations were not replicated.

256 Recently, Chen *et al.*⁴² have performed an association analysis on plasma and faecal levels of
257 bile acids in 297 obese individuals. Their study revealed 27 associated loci, including genes
258 involved in transport of GDP-fucose and zinc/manganese and zinc-finger-protein-related
259 genes, mostly associated with bile acid levels in stool. In our study we analysed blood plasma
260 in a much larger sample from a general population and discovered only two associated loci.
261 Neither of genes identified in our study were reported in Chen *et al*, suggesting that genetic
262 regulation of bile acids between stool and blood plasma or between obese and general
263 populations might differ significantly.

264 We acknowledge several limitations in the present study. We found only a small percentage of
265 BA variability to be affected by genetics, suggesting that a larger sample size is required to
266 further describe BA genetic architecture. BAs are known to be largely influenced by
267 environmental factors, such as sex and gut microbiota. Female sex and oestrogens are
268 considered relevant regulators of BA production and composition^{43,44}. In pregnant women, high
269 levels of circulating oestrogen are associated with development of cholestasis, characterised by
270 increased serum bile acids, likely via oestrogen reducing the expression of BA receptor and
271 transport proteins⁴⁵. Similarly, age-related differences in hormone levels influence the
272 differential production of BAs in women⁴⁶. The relevant impact of sex on plasma BA levels
273 was confirmed by the sex-stratified analysis, where the two significantly associated loci
274 showed to be sex-specific. Similarly, species-composition of gut microbiota has a great impact
275 on BAs levels, especially for secondary BAs that are a direct result of microbiome activity. A
276 recent study describing the effect of gut microbiota on the human plasma metabolome reported
277 that both primary BA cholic acid (CA) and secondary BA deoxycholic acid (DCA) show a high
278 percentage of variance explained by the microbiota ($R^2 = 30\%$ and 36% , respectively),
279 indicating a strong impact on BAs of the variation in microbiota composition⁴⁷. It is important
280 to interpret our findings in the context of the tissue in which BA levels were measured, blood
281 plasma. Bile acids are synthesised in the liver and secreted into the intestine, to be then
282 reabsorbed into portal circulation and returned to the liver: plasma BA levels thus reflect the
283 amount of BAs escaping extraction from the portal blood. Therefore, levels of BAs in plasma
284 are likely to be influenced by genes other than those encoding the particular anabolic and
285 catabolic enzymes, including those involved in hepatic function and dysfunction. In line with
286 this, the major genetic contributor to blood BA levels in our study are variants from the
287 *SLCO1B1* gene, encoding the hepatocyte transporter OATP1B1 and important for flux of bile
288 salts, bilirubin glucuronides and various hormone metabolites, rather than genes encoding key
289 enzymes of primary BA synthesis, such as *CYP7A1* and *CYP7B1*⁴⁸. Similarly, some of the
290 genes with rare variants associations have been linked to liver diseases, such as liver cancer⁴⁹,
291 and intrahepatic cholestasis of pregnancy⁵⁰.

292 In conclusion, we explored the genetic architecture of plasma bile acid levels, including both
293 common and rare variants. By performing GWAS meta-analysis ($N = 4923$), we identified 2
294 significantly associated loci, mapping to the *SLCO1B1* and *PRKGI* genes. In the sex-specific
295 GWAS meta-analysis we observed that variants in these genes have different impact on bile
296 acid levels in men and women. To assess relationships between genetically increased levels of
297 bile acids and risk for diseases we performed Mendelian randomisation, but did not find any
298 bile acids affecting disease risk, nor the reverse, which however might be affected by the lack
299 of statistical power. Using the gene-based aggregated tests and whole exome sequencing, we
300 further identified rare pLoF and missense variants in 3 genes associated with BAs, *ORIG1*,
301 *SART1* and *SORCS2*, some of which are known to be involved in liver disease. Additional
302 studies with larger sample sizes and of more diverse ancestry will be necessary to validate our
303 findings, further unravel the genetic architecture of bile acid levels, and to understand their
304 relationship with human diseases and complex traits.

305 **Materials and methods**

306

307 **Phenotypic data**

308 **Bile acids quantification**

309 Bile acid (BA) analysis was performed from plasma or serum (MICROS cohort) samples by
310 liquid chromatography-tandem mass spectrometry (LC-MS/MS) as previously described⁵¹.
311 The HPLC equipment consisted of a 1200 series binary pump (G1312B), a 1200 series isocratic
312 pump (G1310A) and a degasser (G1379B) (Agilent, Waldbronn, Germany) connected to an
313 HTC Pal autosampler (CTC Analytics, Zwingen, CH). A hybrid triple quadrupole linear ion
314 trap mass spectrometer API 4000 Q-Trap equipped with a Turbo V source ion spray operating
315 in negative ESI mode was used for detection (Applied Biosystems, Darmstadt, Germany). High
316 purity nitrogen was produced by a nitrogen generator NGM 22-LC/MS (cmc Instruments,
317 Eschborn, Germany). Gradient chromatographic separation of BAs was performed on a 50 mm
318 × 2.1 mm (i.d.) Macherey-Nagel NUCLEODUR C18 Gravity HPLC column, packed with 1.8
319 µm particles equipped with a 0.5 µm pre-filter (Upchurch Scientific, Oak Harbor, WA, USA).
320 The injection volume was 5 µL and the column oven temperature was set to 50 °C. Mobile
321 phase A was methanol/water (1/1, v/v), mobile phase B was 100% methanol, both containing
322 0.1% ammonium hydroxide (25%) and 10 mmol/L ammonium acetate (pH 9). A gradient
323 elution was performed with 100% A for 0.5 min, a linear increase to 50% A until 4.5 min,
324 followed by 0% A from 4.6 until 5.5 min and re-equilibration from 5.6 to 6.5 min with 100%
325 A. The flow rate was set to 500 µL/min. To minimize contamination of the mass spectrometer,
326 the column flow was directed only from 1.0 to 5.0 min into the mass spectrometer using a
327 diverter valve. Otherwise, methanol with a flowrate of 250 µL/min was delivered into the mass
328 spectrometer. The turbo ion spray source was operated in the negative ion mode using the
329 following settings: Ion spray voltage = -4500 V, ion source heater temperature = 450 °C,
330 source gas 1 = 40 psi, source gas 2 = 35 psi and curtain gas setting = 20 psi. Analytes were
331 monitored in the multiple reaction monitoring (MRM). Quadrupoles Q1 and Q3 were working
332 at unit resolution. Calibration was achieved by the addition of BAs to EDTA-plasma/serum. A
333 combined BA standard solution containing the indicated amounts (0.5 - 70.5 µmol/L) was
334 placed in a 1.5 ml tube and excess solvent was evaporated under reduced pressure before adding
335 EDTA-plasma/serum. Calibration curves were calculated by linear regression without
336 weighting. Data analysis was performed with Analyst Software 1.4.2. (Applied Biosystems,
337 Darmstadt, Germany). The data were exported to Excel spreadsheets and further processed by
338 self-programmed Excel macros which sort the results, calculate the analyte/internal standard
339 peak area ratios, generate calibration lines and calculate sample concentrations. For the
340 calculation we selected the internal standard with analogous fragmentation and closest
341 retention time to the respective BA species.

342

343 **Pre-processing of bile acid traits**

344 Prior to genetic analysis, bile acid traits were grouped into three groups based on the percentage
345 of samples with below the limit of detection (<LOD) measurements: high <LOD group (>
346 ~30% of all samples below LOD) and low <LOD group (< ~7% of all samples below LOD)
347 (Supplementary Table 1). Accordingly, different phenotypic pre-processing and different
348 analysis strategies were applied to the groups. Bile acids within a high <LOD were considered
349 as binary traits: individuals were categorised based on whether their bile acid levels were
350 effectively measured (category 1) or were below the LOD (category 0). Bile acid traits
351 belonging to this group were THDCA, TUDCA, TCA and GHDCa. All other bile acids were
352 considered as quantitative traits and were \log_{10} -transformed. However, to increase the sample
353 size, in addition to a complete-case analysis (considering as missing all samples with <LOD),
354 we also imputed <LOD measurements. For each bile acid, imputation of <LOD measurements
355 was performed by fitting a truncated normal distribution, with mean and standard deviation of
356 the effectively measured raw data, truncated (as an upper bound) to the lowest measured value
357 for the given bile acid. To do so, we used the “rtnorm” function from the MCMCglmm R
358 package⁵². After imputation, measurements were \log_{10} -transformed.

359

360 **Genome-wide association analysis**

361 Genome-wide association studies (GWAS) were performed in 5 cohorts of European descent,
362 CROATIA-Vis (N=971), ORCADES (Orkney Complex Disease Study) (N=1019), NSPHS
363 (Northern Sweden Population Health Study) (N=718), MICROS (Micro-Isolates in South
364 Tyrol) (N=1336) and ERF (Erasmus Rucphen Family Study) (N=879), for a combined sample
365 size of 4923. Specific sample size for each bile acid molecule, in both meta-analysis and single
366 cohort GWAS, can be found in Supplementary Table 10. Bile acid traits were adjusted for age,
367 sex, batch, population structure/cryptic relatedness by including population principal
368 components or applying linear mixed models and using a kinship matrix estimated from
369 genotyped data. Within each cohort, residuals of covariate and population structure/relatedness
370 correction were tested for association with Haplotype Reference Consortium (HRC)⁵³ imputed
371 SNP dosages or SNP genotypes from whole genome sequencing, applying an additive genetic
372 model of association. Details of cohorts, individual-level pre-imputation QC, GWAS software
373 and parameters specific for each cohort can be seen in Supplementary Table 11 Single-cohort
374 summary statistics were filtered for minor allele frequency (MAF) > 0.01. The genomic control
375 inflation factor (λ_{GC}) was calculated for each bile acid trait. Cohort-level λ_{GC} overall ranged
376 from 0.9 to 1.1 for quantitative bile acid traits, both imputed and not, suggesting little residual
377 influence of population stratification and family structure (Supplementary Table 12). In a few
378 cases, ERF cohort reported somewhat deflated λ_{GC} (GCDCA at 0.884 and GLCA at 0.899). On
379 the other hand, there was considerable inflation for binary bile acid in the case of NSPHS
380 (Supplementary Table 12), with values of λ_{GC} above 1.1, suggesting that population
381 structure/cryptic relatedness was not fully controlled for these traits in the NSPHS cohort.

382

383 **Meta-analysis**

384 Prior to meta-analysis, cohort-level GWAS were quality controlled using the EasyQC software
385 package, following the protocol described in Winkler *et al.*⁵⁴ Cohort-level results were
386 corrected for the genomic control inflation factor, then pooled and analysed with METAL
387 v2011-03-25 software⁵⁵, applying the fixed-effect inverse-variance method. The mean genomic
388 control inflation factor after the meta-analysis was 0.991 (range 0.938 – 1.009), suggesting that
389 the confounding effects of the family structure were correctly accounted for (Supplementary
390 Table 12). The standard genome-wide significance threshold was Bonferroni corrected for the
391 number of independent bile acid traits, calculated as $14 (5 \times 10^{-8} / 14 = 3.57 \times 10^{-9})$. The number
392 of independent bile acid traits was estimated as the sum of the number of binary traits (4) and
393 the number of principal components that jointly explained 99% of the total variance of log₁₀-
394 transformed quantitative traits in each cohort (10) (Supplementary Table 13).

395

396 **Sex-stratified GWAS meta-analysis**

397 To identify possible differences in the genetic contribution to bile acid variability between men
398 and women, we performed sex-specific GWAS of the 14 quantitative bile acid traits for
399 ORCADES and CROATIA-Vis cohorts. Given that for the sex-stratified GWAS we implicitly
400 halve our sample size, we performed these analyses only on the imputed bile acid traits. The
401 same analysis steps and procedures already described for the full meta-analysis were applied.
402 Bile acid traits were adjusted for age, sex and batch as fixed effects, and relatedness (estimated
403 as the kinship matrix calculated from genotyped data) as a random effect in a linear mixed
404 model, calculated using the ‘polygenic’ function from the GenABEL R package⁵⁶. Residuals
405 of covariate and relatedness correction were tested for association with HRC-imputed⁵³ SNP
406 dosages using the RegScan v0.5 software⁵⁷, applying an additive genetic model of association.
407 Prior to meta-analysis, SNPs having a difference in allele frequency between the two cohorts
408 higher than ± 0.3 or a minor allele count (MAC) lower or equal to 6 were filtered out. Cohort-
409 level GWAS were corrected for genomic control inflation factor and then meta-analysed
410 ($N=820$ for male and $N=1088$ for female individuals) using METAL v2011-03-25 software⁵⁵,
411 applying the fixed-effect inverse-variance method. The mean λ_{GC} was 0.993 (range 0.978–
412 1.011) for male-specific meta-analysis and 0.996 (range 0.984–1.003) for female-specific
413 meta-analysis. The Bonferroni-corrected significance threshold applied is 5×10^{-9} .

414

415 **Phenoscaner and Mendelian Randomisation**

416 To assess link between bile acids and diseases we explored the overlap of SNPs associated with
417 BAs with complex human traits by using PhenoScanner v1.1 database^{16,17}, taking into account
418 significant genetic association ($p < 5 \times 10^{-9}$) at the same or strongly ($LD r^2 > 0.8$) linked SNPs
419 in populations of European ancestry. We then performed bi-directional Mendelian
420 Randomisation (MR) to investigate the effect of 548 complex traits and diseases available in
421 the IEU Open GWAS database¹⁸ (manually curated list of studies from identifiers ebi-a, ieu-a,
422 ieu-b and ukb-a; the complete list reported in the Supplementary Table 5) on BA levels, and

423 vice-versa. The set of genome-wide significant, LD clumped SNPs used as instruments for
424 complex traits/diseases was extracted from the selected studies by using the
425 “extract_instruments” function from the TwoSampleMR 0.5.6 R package⁵⁸. Similarly, sentinel
426 SNPs from BAs meta-analysis (Supplementary Table 2) were selected as instruments. MR tests
427 were performed by using fixed effects inverse variance-weighted (IVW) in case of multiple
428 instruments or Wald Ratio method in case of a single instrument, as implemented in the
429 TwoSampleMR 0.5.6 R package⁵⁸. Multiple testing correction was controlled for using either
430 the Bonferroni correction or false discovery rate (FDR).

431

432 **Whole-exome sequencing data**

433 Exome sequencing

434 The “Goldilocks” exome sequence data for ORCADES cohort was prepared at the Regeneron
435 Genetics Center, following the protocol detailed in Van Hout *et al.*⁵⁹ for the UK Biobank
436 whole-exome sequencing project. In summary, sequencing was performed using S2 flow cells
437 on the Illumina NovaSeq 6000 platform with multiplexed samples. DNAnexus platform⁶⁰ was
438 used for processing raw sequencing data. The files were converted to FASTQ format and
439 aligned using the BWA-mem⁶¹ to GRCh38 genome reference. The Picard tool⁶² was used for
440 identifying and flagging duplicated reads, followed by calling the genotypes for each individual
441 sample using the WeCall variant caller⁶³. During quality control, 33 samples genetically
442 identified as duplicates, 3 samples showing disagreement between genetically determined and
443 reported sex, 4 samples with high rates of heterozygosity or contamination, 2 samples having
444 low sequence coverage (less than 80% of targeted bases achieving 20X coverage) and 1 being
445 discordant with genotyping chip were excluded. Finally, the “Goldilocks” dataset was
446 generated by (i) filtering out genotypes with read depth lower than 7 reads, (ii) keeping variants
447 having at least one heterozygous variant genotype with allele balance ratio greater than or equal
448 to 15% ($AB \geq 0.15$) or at least one homozygous variant genotype, and (iii) filtering out variants
449 with more than 10% of missingness and HWE $p < 10^{-6}$. Overall, a total of 2,090 ORCADES
450 (820 male and 1,270 female) participants passed all exome sequence and genotype quality
451 control thresholds. A pVCF file containing all samples passing quality control was then created
452 using the GLnexus joint genotyping tool⁶⁴.

453

454 Variant annotation

455 Exome sequencing variants were annotated as described in Van Hout, *et al.*⁵⁹ Briefly, they were
456 annotated with the most severe consequence across all protein-coding transcripts using
457 SnpEff⁶⁵. Gene regions were defined based on Ensembl release 85⁶⁶. Predicted loss-of function
458 (pLoF) variants were defined as variants annotated as start lost, stop gained/lost, splice
459 donor/acceptor and frameshift. The deleteriousness of missense variants was based on dbNSFP
460 3.2^{67,68} and assessed using the following algorithms: (1) SIFT⁶⁹: “D” (Damaging), (2)

461 Polyphen2_HDIV: “D” (Damaging) or “P” (Possibly damaging), (3) Polyphen2_HVAR⁷⁰: “D”
462 (Damaging) or “P” (Possibly damaging), (4) LRT⁷¹: “D” (Deleterious) and (5)
463 MutationTaster⁷²: “A” (Disease causing automatic) or “D” (Disease causing). If not predicted
464 as deleterious by any of the algorithms the missense variants were considered “likely benign”,
465 “possibly deleterious” if predicted as deleterious by at least one of the algorithms and “likely
466 deleterious” if predicted as deleterious by all five algorithms.

467

468 **Exome-wide gene-based aggregation analysis of rare variants**

469 Generation of gene masks

470 For each gene, the variants were grouped into four categories (masks), based on severity of
471 their functional consequence. The first mask (mask 1) included only pLoF variants. Masks 3
472 and 4 included both pLoF and variants predicted to be deleterious, by 5/5 algorithms (mask 3)
473 or by at least one algorithm (mask 4). The most permissive mask (mask 2) included pLoF and
474 all missense variants. These masks were then further split by the frequencies of the minor allele
475 (MAF \leq 5%, e.g. mask1_maf5; and MAF \leq 1%, e.g. mask1_maf1), resulting in up to 8 burden
476 tests for each gene (Supplementary Table 9).

477

478 ORCADES gene-based aggregation analysis

479 We performed variant Set Mixed Model Association Tests (SMMAT)⁷³ on the 18 bile acid
480 traits from ORCADES cohort, quantified and pre-processed as previously described, fitting a
481 GLMM adjusting for age, sex, batch, and familial or cryptic relatedness by kinship matrix. The
482 kinship matrix was estimated from the genotyped data using the ‘ibs’ function from GenABEL
483 R package⁵⁶. The SMMAT framework includes 4 variant aggregate tests: burden test, sequence
484 kernel association test (SKAT), SKAT-O and SMMAT-E, a hybrid test combining the burden
485 test and SKAT. The 4 variant aggregate tests were performed on 8 different pools of genetic
486 variants, called “masks”, each one including a different set of variants based on both MAF and
487 predicted consequence of variants (e.g., loss of function and missense) (Supplementary Table
488 9), as described above. Discovery significance threshold was Bonferroni corrected for the
489 rough estimate of the number of genes in the human genome, 20,000, and the number of
490 independent bile acid traits, 14, calculated as previously described ($0.05/20000/14 = 1.79 \times 10^{-7}$
491 ⁷). A gene association was considered significant if it passed the above reported Bonferroni
492 corrected significance threshold in at least two of the 4 performed variant aggregate tests and
493 if the cumulative allele count of the variants included in the gene was equal or higher than 10.

494 **Code availability**

495 We used publicly available software tools for all analyses. These software tools are listed in
496 the main text and in the Methods.

497

498 **Data availability**

499 The full summary statistics from GWAS meta-analysis of bile acids will be uploaded to the
500 University of Edinburgh Datashare repository and to GWAS catalog upon manuscript
501 acceptance. There is neither Research Ethics Committee approval, nor consent from individual
502 participants, to permit open release of the individual level research data underlying this study.
503 The datasets analysed during the current study are therefore not publicly available. Instead, the
504 research data and/or DNA samples for the ORCADES study are available from
505 accessQTL@ed.ac.uk on reasonable request, following approval by the QTL Data Access
506 Committee and in line with the consent given by participants. Each approved project is subject
507 to a data or materials transfer agreement (D/MTA) or commercial contract. The summary
508 statistics for complex traits and diseases (full list reported in Supplementary Table 5) are
509 available in the IEU Open GWAS database <https://gwas.mrcieu.ac.uk/>.

510 References

- 511 1. Lorbek, G., Lewinska, M. & Rozman, D. Cytochrome P450s in the synthesis of
512 cholesterol and bile acids – from mouse models to human diseases. *FEBS J.* **279**,
513 1516–1533 (2012).
- 514 2. Chiang, J. Y. L. Bile Acid Metabolism and Signaling. *Compr. Physiol.* **3**, 1191–1212
515 (2013).
- 516 3. Thomas, C., Pellicciari, R., Pruzanski, M., Auwerx, J. & Schoonjans, K. Targeting
517 bile-acid signalling for metabolic diseases. *Nature Reviews Drug Discovery* **7**, 678–
518 693 (2008).
- 519 4. de Aguiar Vallim, T. Q., Tarling, E. J. & Edwards, P. A. Pleiotropic roles of bile acids
520 in metabolism. *Cell Metab* **17**, 657–669 (2013).
- 521 5. Perino, A., Demagny, H., Velazquez-Villegas, L. & Schoonjans, K. Molecular
522 Physiology of Bile Acid Signaling in Health, Disease, and Aging. *Physiol Rev* **101**,
523 683–731 (2021).
- 524 6. Fiorucci, S. *et al.* Bile Acid Signaling in Inflammatory Bowel Diseases. *Dig Dis Sci*
525 **66**, 674–693 (2021).
- 526 7. Rezen, T. *et al.* The role of bile acids in carcinogenesis. *Cell Mol Life Sci* **79**, 243
527 (2022).
- 528 8. Danic, M. *et al.* Pharmacological Applications of Bile Acids and Their Derivatives in
529 the Treatment of Metabolic Syndrome. *Front Pharmacol* **9**, 1382 (2018).
- 530 9. Stojančević, M., Pavlović, N., Goločorbin-Kon, S. & Mikov, M. Application of bile
531 acids in drug formulation and delivery. *Front. Life Sci.* **7**, 112–122
- 532 10. Surendran, P. *et al.* Rare and common genetic determinants of metabolic individuality
533 and their effects on human health. *Nat Med* **28**, 2321–2332 (2022).
- 534 11. Bomba, L. *et al.* Whole-exome sequencing identifies rare genetic variants associated
535 with human plasma metabolites. *Am. J. Hum. Genet.* **109**, 1038–1054 (2022).
- 536 12. Demirhan, A. *et al.* Insight in genome-wide association of metabolite quantitative traits
537 by exome sequence analyses. *PLoS Genet* **11**, e1004835 (2015).
- 538 13. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and
539 reveals novel systemic effects of LPA. *Nat. Commun.* **7**, 1–9 (2016).
- 540 14. Lotta, L. A. *et al.* A cross-platform approach identifies genetic regulators of human
541 metabolism and health. *Nat. Genet.* **53**, 54–64 (2021).
- 542 15. Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat.*
543 *Genet.* **46**, 543–550 (2014).
- 544 16. Staley, J. R. *et al.* PhenoScanner: A database of human genotype-phenotype
545 associations. *Bioinformatics* **32**, 3207–3209 (2016).
- 546 17. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human
547 genotype-phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).
- 548 18. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv*
549 2020.08.10.244293 (2020). doi:10.1101/2020.08.10.244293
- 550 19. Hagenbuch, B. & Meier, P. J. Organic anion transporting polypeptides of the OATP/
551 SLC21 family: phylogenetic classification as OATP/SLCO superfamily, new
552 nomenclature and molecular/functional properties. *Pflugers Arch* **447**, 653–665
553 (2004).
- 554 20. Ho, R. H. & Kim, R. B. Transporters and drug therapy: implications for drug
555 disposition and disease. *Clin Pharmacol Ther* **78**, 260–277 (2005).
- 556 21. International Transporter, C. *et al.* Membrane transporters in drug development. *Nat*
557 *Rev Drug Discov* **9**, 215–236 (2010).
- 558 22. Niemi, M., Pasanen, M. K. & Neuvonen, P. J. Organic anion transporting polypeptide

- 559 1B1: a genetically polymorphic transporter of major importance for hepatic drug
560 uptake. *Pharmacol Rev* **63**, 157–181 (2011).
- 561 23. Nies, A. T., Schwab, M. & Keppler, D. Interplay of conjugating enzymes with OATP
562 uptake transporters and ABCC/MRP efflux pumps in the elimination of drugs. *Expert*
563 *Opin Drug Metab Toxicol* **4**, 545–568 (2008).
- 564 24. Revez, J. A. *et al.* Genome-wide association study identifies 143 loci associated with
565 25 hydroxyvitamin D concentration. *Nat Commun* **11**, 1647 (2020).
- 566 25. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat.*
567 *Genet.* **45**, 1274–1285 (2013).
- 568 26. Johnson, A. D. *et al.* Genome-wide association meta-analysis for total serum bilirubin
569 levels. *Hum. Mol. Genet.* **18**, 2700–2710 (2009).
- 570 27. Ruth, K. S. *et al.* Using human genetics to understand the disease impacts of
571 testosterone in men and women. *Nat Med* **26**, 252–258 (2020).
- 572 28. Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug-target
573 identification and prioritisation. *Nucleic Acids Res* **49**, D1302–D1310 (2021).
- 574 29. Vitart, V. *et al.* 3000 years of solitude: extreme differentiation in the island isolates of
575 Dalmatia, Croatia. *Eur. J. Hum. Genet.* **14**, 478–487 (2006).
- 576 30. Zuk, O. *et al.* Searching for missing heritability : Designing rare variant association
577 studies. (2014). doi:10.1073/pnas.1322563111
- 578 31. Koh, M. Y., Lemos Jr., R., Liu, X. & Powis, G. The hypoxia-associated factor
579 switches cells from HIF-1alpha- to HIF-2alpha-dependent signaling promoting stem
580 cell characteristics, aggressive tumor growth and invasion. *Cancer Res* **71**, 4015–4027
581 (2011).
- 582 32. Semenza, G. L. Hypoxia, clonal selection, and the role of HIF-1 in tumor progression.
583 *Crit Rev Biochem Mol Biol* **35**, 71–103 (2000).
- 584 33. Braun, T., Volland, P., Kunz, L., Prinz, C. & Gratzl, M. Enterochromaffin cells of the
585 human gut: sensors for spices and odorants. *Gastroenterology* **132**, 1890–1901 (2007).
- 586 34. Erspamer, V. Pharmacology of indole-alkylamines. *Pharmacol Rev* **6**, 425–487 (1954).
- 587 35. Watanabe, H. *et al.* Peripheral serotonin enhances lipid metabolism by accelerating
588 bile acid turnover. *Endocrinology* **151**, 4776–4786 (2010).
- 589 36. Oakley, F. *et al.* Hepatocytes express nerve growth factor during liver injury: evidence
590 for paracrine regulation of hepatic stellate cell apoptosis. *Am J Pathol* **163**, 1849–1858
591 (2003).
- 592 37. Ohkubo, T. *et al.* Early induction of nerve growth factor-induced genes after liver
593 resection-reperfusion injury. *J Hepatol* **36**, 210–217 (2002).
- 594 38. Valdovinos-Flores, C. & Gonsbatt, M. E. Nerve growth factor exhibits an antioxidant
595 and an autocrine activity in mouse liver that is modulated by buthionine sulfoximine,
596 arsenic, and acetaminophen. *Free Radic Res* **47**, 404–412 (2013).
- 597 39. Gigliozzi, A. *et al.* Nerve growth factor modulates the proliferative capacity of the
598 intrahepatic biliary epithelium in experimental cholestasis. *Gastroenterology* **127**,
599 1198–1209 (2004).
- 600 40. Rasi, G. *et al.* Nerve growth factor involvement in liver cirrhosis and hepatocellular
601 carcinoma. *World J Gastroenterol* **13**, 4986–4995 (2007).
- 602 41. Tokusashi, Y. *et al.* Expression of NGF in hepatocellular carcinoma cells with its
603 receptors in non-tumor cell components. *Int J Cancer* **114**, 39–45 (2005).
- 604 42. Chen, L. *et al.* Genetic and Microbial Associations to Plasma and Fecal Bile Acids in
605 Obesity Relate to Plasma Lipids and Liver Fat Content. *Cell Rep.* **33**, 108212 (2020).
- 606 43. Phelps, T., Snyder, E., Rodriguez, E., Child, H. & Harvey, P. The influence of
607 biological sex and sex hormones on bile acid synthesis and cholesterol homeostasis.
608 *Biol. Sex Differ.* 2019 101 **10**, 1–12 (2019).

- 609 44. Li-Hawkins, J. *et al.* Cholic acid mediates negative feedback regulation of bile acid
610 synthesis in mice. *J Clin Invest* **110**, 1191–1200 (2002).
- 611 45. Abu-Hayyeh, S. *et al.* Intrahepatic cholestasis of pregnancy levels of sulfated
612 progesterone metabolites inhibit farnesoid X receptor resulting in a cholestatic
613 phenotype. *Hepatology* **57**, 716–726 (2013).
- 614 46. Frommherz, L. *et al.* Age-Related Changes of Plasma Bile Acid Concentrations in
615 Healthy Adults--Results from the Cross-Sectional KarMeN Study. *PLoS One* **11**,
616 e0153959 (2016).
- 617 47. Dekkers, K. F. *et al.* An online atlas of human plasma metabolite signatures of gut
618 microbiome composition. *Nat. Commun. 2022 131* **13**, 1–12 (2022).
- 619 48. Russell, D. W. The Enzymes, Regulation, and Genetics of Bile Acid Synthesis. *Annu.*
620 *Rev. Biochem.* **72**, 137–174 (2003).
- 621 49. Thomas, C. E. *et al.* Association between Pre-Diagnostic Serum Bile Acids and
622 Hepatocellular Carcinoma: The Singapore Chinese Health Study. *Cancers (Basel)* **13**,
623 (2021).
- 624 50. Manzotti, C., Casazza, G., Stimac, T., Nikolova, D. & Gluud, C. Total serum bile acids
625 or serum bile acid profile, or both, for the diagnosis of intrahepatic cholestasis of
626 pregnancy. *Cochrane Database Syst Rev* **7**, CD012546 (2019).
- 627 51. Scherer, M., Gnewuch, C., Schmitz, G. & Liebisch, G. Rapid quantification of bile
628 acids and their conjugates in serum by liquid chromatography-tandem mass
629 spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **877**, 3920–3925
630 (2009).
- 631 52. Hadfield, J. D. MCMC Methods for Multi-Response Generalized Linear Mixed
632 Models: TheMCMCglmmRPackage. *J. Stat. Softw.* **33**, (2010).
- 633 53. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation.
634 *Nat. Genet.* **48**, 1279–1283 (2016).
- 635 54. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-
636 analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
- 637 55. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of
638 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
- 639 56. Karssen, L. C., van Duijn, C. M. & Aulchenko, Y. S. The GenABEL Project for
640 statistical genomics. *F1000Research* **5**, (2016).
- 641 57. Haller, T., Kals, M., Esko, T., Mägi, R. & Fischer, K. RegScan: A GWAS tool for
642 quick estimation of allele effects on continuous traits and their combinations. *Brief.*
643 *Bioinform.* **16**, 39–44 (2013).
- 644 58. Hemani, G. *et al.* The MR-base platform supports systematic causal inference across
645 the human phenome. *Elife* **7**, (2018).
- 646 59. Van Hout, C. V *et al.* Exome sequencing and characterization of 49,960 individuals in
647 the UK Biobank. *Nature* **586**, 749–756 (2020).
- 648 60. Reid, J. G. *et al.* Launching genomics into the cloud: deployment of Mercury, a next
649 generation sequence analysis pipeline. *BMC Bioinformatics* **15**, 30 (2014).
- 650 61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
651 transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 652 62. “Picard Toolkit.” 2019. Broad Institute, GitHub Repository.
653 <https://broadinstitute.github.io/picard/>; Broad Institute.
- 654 63. PLC, G. weCall. (2018).
- 655 64. Lin, M. F. *et al.* GLnexus: joint variant calling for large cohort sequencing. *bioRxiv*
656 (2018). doi:10.1101/343970
- 657 65. Cingolani, P. *et al.* A program for annotating and predicting the effects of single
658 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*

- 659 strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
- 660 66. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
- 661 67. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of
662 Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site
663 SNVs. *Hum Mutat* **37**, 235–241 (2016).
- 664 68. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human
665 nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894–899
666 (2011).
- 667 69. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense
668 predictions for genomes. *Nat Protoc* **11**, 1–9 (2016).
- 669 70. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense
670 mutations. *Nat. Methods* **7**, 248–249 (2010).
- 671 71. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human
672 genomes. *Genome Res* **19**, 1553–1561 (2009).
- 673 72. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster
674 evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575–576
675 (2010).
- 676 73. Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous
677 and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *Am. J. Hum.*
678 *Genet.* **104**, 260–274 (2019).
- 679

680 Acknowledgments

681 The Orkney Complex Disease Study (ORCADES) was supported by the Chief Scientist Office
682 of the Scottish Government (CZB/4/276, CZB/4/710), a Royal Society URF to J.F.W., the
683 MRC Human Genetics Unit quinquennial programme “QTL in Health and Disease”, Arthritis
684 Research UK and the European Union framework program 6 EUROSPAN project (contract
685 no. LSHG-CT-2006-018947). DNA extractions were performed at the Edinburgh Clinical
686 Research Facility, University of Edinburgh. We would like to acknowledge the invaluable
687 contributions of the research nurses in Orkney, the administrative team in Edinburgh and the
688 people of Orkney. For the purpose of open access, the author has applied a Creative Commons
689 Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this
690 submission. The CROATIA-VIS study in the Croatian island of Vis was supported through the
691 grants from the Medical Research Council UK and Ministry of Science, Education and Sport
692 of the Republic of Croatia (number 108-1080315-0302). The authors collectively thank a large
693 number of individuals for their individual help in organising, planning and carrying out the
694 field work related to the project and data management: Professor Pavao Rudan and the staff of
695 the Institute for Anthropological Research in Zagreb, Croatia (organisation of the field work,
696 anthropometric and physiological measurements, and DNA extraction); Professor Ariana
697 Vorko-Jovic and the staff and medical students of the Andrija Stampar School of Public Health
698 of the Faculty of Medicine, University of Zagreb, Croatia (questionnaires, genealogical
699 reconstruction and data entry); Dr Branka Salzer from the biochemistry lab “Salzer”, Croatia
700 (measurements of biochemical traits); local general practitioners and nurses (recruitment and
701 communication with the study population); and the employees of several other Croatian
702 institutions who participated in the field work, including but not limited to the University of
703 Rijeka and Split, Croatia; Croatian Institute of Public Health; Institutes of Public Health in Split
704 and Dubrovnik, Croatia. SNP Genotyping of the Vis samples was carried out by the Genetics
705 Core Laboratory at the Wellcome Trust Clinical Research Facility, WGH, Edinburgh. The
706 MICROS (Micro-Isolates in South Tyrol) study is part of the genomic health care program
707 'GenNova' and was carried out in three villages of the Val Venosta on the populations of
708 Stelvio, Vallelunga and Martello. We thank the primary care practitioners Raffaella Stocker,
709 Stefan Waldner, Toni Pizzocco, Josef Plangger, Ugo Marcadent and the personnel of the
710 Hospital of Silandro (Department of Laboratory Medicine) for their participation and
711 collaboration in the research project. In South Tyrol, the study was supported by the Ministry
712 of Health and Department of Educational Assistance, University and Research of the
713 Autonomous Province of Bolzano and the South Tyrolean Sparkasse Foundation. The Northern
714 Swedish Population Health Study (NSPHS) was funded by the Swedish Medical Research
715 Council (project number K2007-66X-20270-01-3), and the Foundation for Strategic Research
716 (SSF). The NSPHS as part of EUROSPAN (European Special Populations Research Network)
717 was also supported by European Commission FP6 STRP grant number 01947 (LSHGCT-2006-
718 01947). This work was also supported by the Swedish Society for Medical Research (ÅJ). The
719 authors are grateful for the contribution of district nurse Svea Hennix for data collection and
720 Inger Jonasson for logistics and coordination of the health survey. Finally, the authors thank
721 all the community participants for their interest and willingness to contribute to the study. The
722 Erasmus Rucphen Family (ERF) study was supported by grants from The Netherlands

723 Organisation for Scientific Research (NWO), Erasmus MC, the Centre for Medical Systems
724 Biology (CMSB) and the European Community's Seventh Framework Programme (FP7/2007-
725 2013), ENGAGE Consortium, grant agreement HEALTH-F4-2007- 201413. We are grateful
726 to all general practitioners for their contributions, Cornelia van Duijn and Ben Oostra for
727 setting-up the ERF study, Petra Veraart for sorting out the genealogy records, Jeannette
728 Vergeer and Peter Snijders for help in retrieving the materials needed to analyse data. We
729 acknowledge support from the European Union's Horizon 2020 research and innovation
730 programme IMforFUTURE (A.L.: H2020-MSCA-ITN/721815); the RCUK Innovation
731 Fellowship from the National Productivity Investment Fund (L.K.: MR/R026408/1) and the
732 MRC Human Genetics Unit programme grant, 'QTL in Health and Disease' (J.F.W. and C.H.:
733 MC_UU_00007/10).

734

735 **Ethics**

736 All studies were approved by local research ethics committees and all participants have given
737 written informed consent. The ORCADES study was approved by the NHS Orkney Research
738 Ethics Committee and the North of Scotland REC. The CROATIA-Vis study was approved by
739 the ethics committee of the medical faculty in Zagreb and the Multi-Centre Research Ethics
740 Committee for Scotland. The Northern Swedish Population Health Study (NSPHS) was
741 approved by the local ethics committee at the University of Uppsala (Regionala
742 Etikprövningsnämnden, Uppsala). The MICROS study was approved by the ethical committee
743 of the Autonomous Province of Bolzano, Italy. The ERF study was approved by the Erasmus
744 institutional medical-ethics committee in Rotterdam, The Netherlands.

745

746 **Author contributions**

747 A.L.: Data analysis and interpretation, visualisation, writing—original draft preparation,
748 writing—review and editing. D.G.-S.: Data analysis, writing—review and editing. Å.J.: Data
749 analysis. S.A.: Data analysis. G.L.: Quantification of bile acids, writing—original draft
750 preparation. C.G.: Quantification of bile acids, writing—original draft preparation. G.T.:
751 preparation, quality control and annotation of whole-exome sequencing data. A.R.S.: Funding,
752 writing—review and editing. A.A.H.: Genomic and demographic data provider for MICROS
753 cohort. P.P.: Funding, genomic and demographic data provider for MICROS cohort. C.P.:
754 genomic and demographic data provider for MICROS cohort. H.C. Funding. O.P.: Genomic
755 and demographic data provider for CROATIA-Vis cohort. C.H.: Funding, genomic and
756 demographic data provider for CROATIA-Vis cohort. N.P.: supervision and data interpretation
757 for bile acid pre-processing and imputation. M.G.: Genomic and demographic data provider
758 for ERF cohort, writing—review and editing. U.G.: Funding, genomic and demographic data
759 provider for NSPHS cohort. C.F.: Genomic and demographic data provider for MICROS
760 cohort. J.F.W.: Funding, conceptualisation, genomic and demographic data provider for
761 ORCADES cohort, supervision, data interpretation, writing—original draft preparation,

762 writing—review and editing. L.K.: Conceptualisation, supervision, data interpretation,
763 writing—original draft preparation, writing—review and editing.

764

765 **Competing interests**

766 G.T. and A.R.S. are full-time employees of Regeneron Genetics Center and receive salary,
767 stock and stock options as compensation. L.K. is an employee of Humanity Inc., a company
768 developing direct-to-consumer measures of biological ageing. All other authors declare no
769 competing interests.