# Genome-wide association study reveals loci with sex-specific effects on plasma bile acids

Arianna Landini[1,2], Dariush Ghasemi-Semeskandeh[3,4], Åsa Johansson[5], Shahzad Ahmad[6], Gerhard Liebisch[7], Carsten Gnewuch[7], Regeneron Genetics Center[8], Gannie Tzoneva[8], Alan R. Shuldiner[8], Andrew A. Hicks[3], Peter Pramstaller[3], Cristian Pattaro[3], Harry Campbell[2], Ozren Polašek[9,10], Nicola Pirastu[11], Caroline Hayward[1], Mohsen Ghanbari[6], Ulf Gyllensten[5], Christian Fuchsberger[3], James F. Wilson*[1,2] & Lucija Klarić*[1]

1 MRC Human Genetics Unit, Institute for Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom
2 Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
3 Institute for Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck, Bolzano, Italy
4 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
5 Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
6 Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
7 Institute of Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Regensburg, Germany
8 Regeneron Genetics Center, Tarrytown, NY, USA
9 Department of Public Health, School of Medicine, University of Split, Split, Croatia
10 Algebra University College, Zagreb, Croatia
11 Genomics Research Centre, Human Technopole, Milan, Italy

* Authors contributed equally.
Correspondence to: J.F.W (jim.wilson@ed.ac.uk) or L.K. (lucija.klaric@ed.ac.uk)

## Abstract

Bile acids are essential for food digestion and nutrient absorption, but also act as signalling molecules involved in hepatobiliary diseases, gastrointestinal disorders and carcinogenesis. While many studies have focused on the genetic determinants of blood metabolites, research focusing specifically on genetic regulation of bile acids in the general population is currently lacking. Here we investigate the genetic architecture of primary and secondary bile acids in blood plasma, reporting associations with both common and rare variants. By performing genome-wide association analysis (GWAS) of plasma blood levels of 18 bile acids (N = 4923) we identify two significantly associated loci, a common variant mapping to *SLCO1B1* (encoding a liver bilirubin and drug transporter) and a rare variant in *PRKG1* (encoding soluble cyclic GMP-dependent protein kinase). For these loci, in the sex-stratified GWAS (N♂ = 820, N♀ = 1088), we observe sex-specific effects (*SLCO1B1* β ♂ = -0.51, $P = 2.30\text{x}10^{-13}$, β♀ = -0.3, $P = 9.90\text{x}10^{-07}$; *PRKG1* β ♂ = -0.18, $P = 1.80\text{x}10^{-01}$, β ♀ = -0.79, $P = 8.30\text{x}10^{-11}$), corroborating the contribution of sex to bile acid variability. Using gene-based aggregate tests and whole exome sequencing, we identify rare pLoF and missense variants potentially associated with bile acid levels in 3 genes (*OR1G1*, *SART1* and *SORCS2*), some of which have been linked with liver diseases.

## Introduction

Bile acids (BAs) are synthesised from cholesterol in the liver and subsequently stored in the gallbladder. After ingestion of food, BAs are secreted into the small intestine, where they contribute to the digestion of lipid-soluble nutrients[1]. Approximately 95% of BAs are then re-absorbed by the intestinal epithelium and transported back to the liver via the portal vein - a process termed "enterohepatic circulation"[2]. Primary bile acids in humans consists of cholic acid (CA), chenodeoxycholic acid (CDCA), and their taurine- or glycine-bound derivatives (TCA and TCDCA, GCA and GCDCA). Once secreted in the lower gastrointestinal tract, primary BAs are heavily modified by the gut microbiota to produce a broad range of secondary BAs, with deoxycholic acid (DCA), a CA derivative, and lithocholic acid (LCA), a CDCA derivative, being the most prevalent[2]. Bile acids also act as hormone-like signalling molecules, serving as ligands to nuclear (hormone) receptors. Through activation of these diverse signalling pathways, BAs control not only their own transport and metabolism, but also lipid and glucose metabolism, and innate and adaptive immunity[3]. Bile acids are thus involved in regulating several physiological systems, such as fat digestion, cholesterol metabolism, vitamin absorption, and liver function[4]. In addition, given their role in coordinating bile homeostasis, biliary physiology and gastrointestinal functions, impaired signalling of BAs is associated with development of hepatobiliary diseases, such as cholestatic liver disorders, cholesterol gallstone disease and other gallbladder-related conditions[5], and of inflammatory bowel disease[6]. Further, bile acids have been implicated in carcinogenesis - specifically oesophageal, gastric, hepatocellular, pancreatic, colorectal, breast, prostate and ovarian cancer - both as pro-carcinogenic agents and tumour suppressors[7]. Thanks to their role as signalling molecules, BAs have been considered as possible targets for the treatment of metabolic syndrome and various metabolic diseases[8]. Further, BAs are able to facilitate and promote drug permeation through biological membranes, making them of general interest for drug formulation and delivery[9].

While many studies have focused on the genetic determinants of blood metabolites[10–15], research focusing specifically on bile acids in a large sample from the general population is currently lacking. Here we investigate the genetic architecture of primary and secondary BAs, reporting associations with both common and low-frequency/rare variants. First, we performed a genome-wide association meta-analysis (GWAMA) of plasma blood levels of 18 BA traits (N=4923). For a subset of this sample (female N=1088, male N=820), we perform sex-stratified GWAMA, to describe sex-specific genetic contributions to BA variability. We then explore whether complex traits or diseases have a role in influencing BA variability by using Mendelian Randomisation. We finally employ multiple gene-based aggregation tests to investigate rare (MAF < 5%) predicted loss of function (pLoF) and missense variants from whole exome sequencing affecting the 18 BA traits in a subset of our cohorts (N=1006).

## Results

### Loci associated with serum levels of bile acids

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

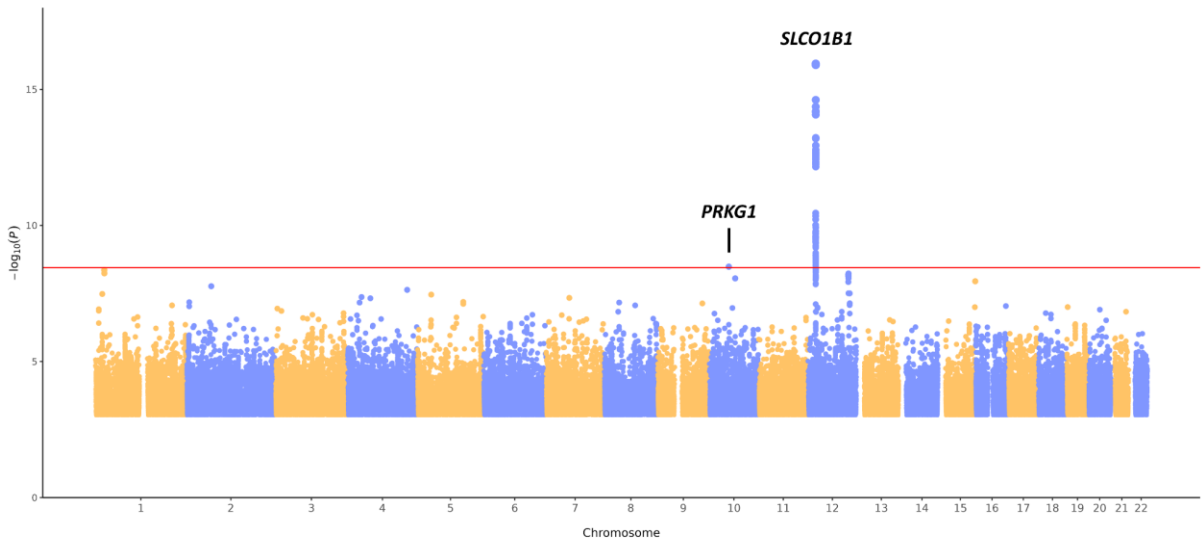



**Figure 1. Summary Manhattan plot pooling together meta-analysis results obtained across 18 bile acid traits.** The pooling was performed by selecting the lowest *p* value (y-axis) from the 18 bile acids for every genomic position (x-axis). The Bonferroni-corrected genome-

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

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

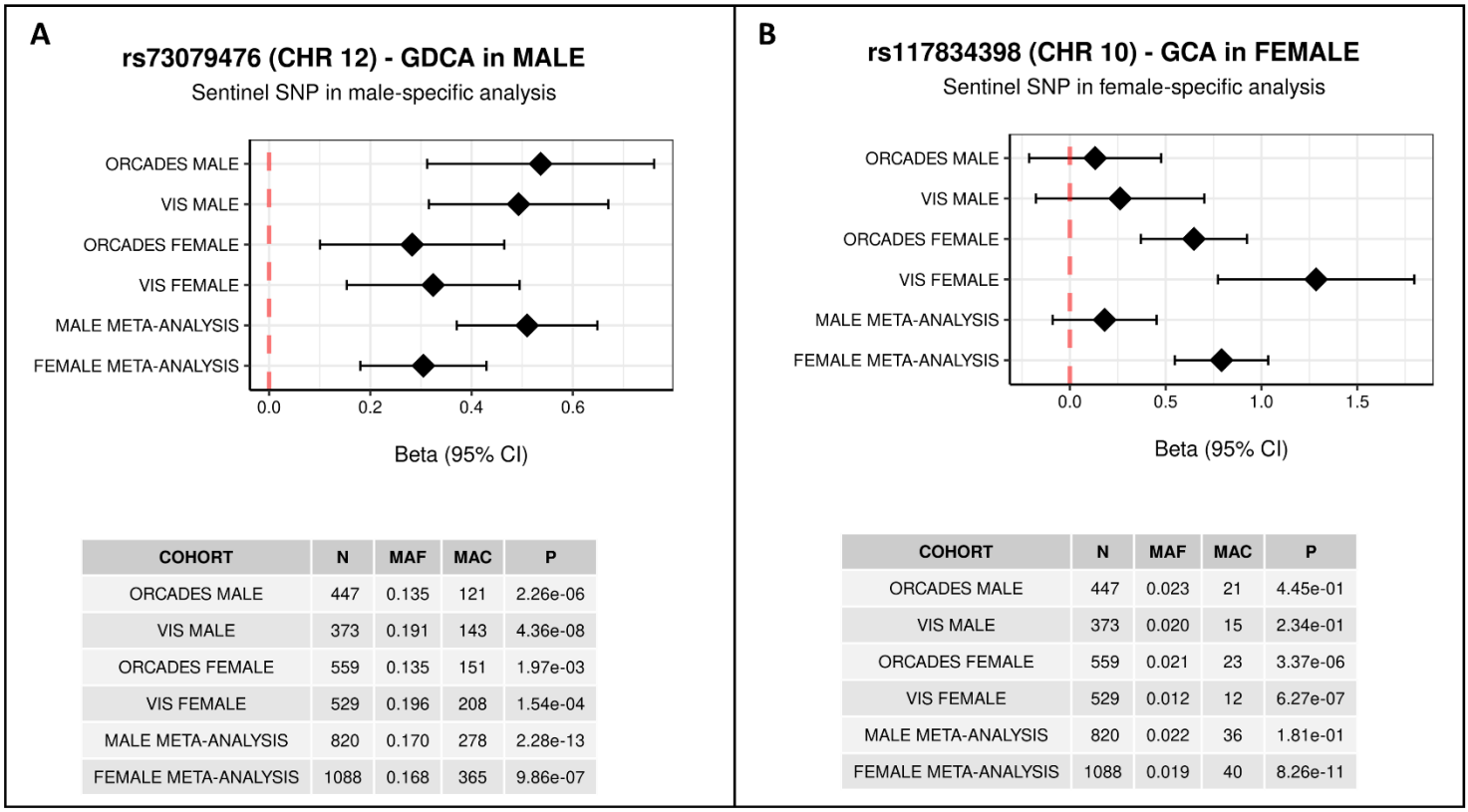
| **All samples** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Locus** | **Gene** | **SNP** | **EA** | **OA** | **EAF** | **Beta** | **P** | **SE** | **N** | **Lead BA** |
| 12:20994540-21463812 | *SLCO1B1* | rs4149056 | T | C | 0.839 | -0.25 | $1.10 \times 10^{-16}$ | 0.03 | 4547 | GDCA |
| 10:53832549-53832549 | *PRKG1* | rs146800892 | T | C | 0.988 | -0.96 | $3.30 \times 10^{-09}$ | 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 | *SLCO1B1* | rs73079476 | A | C | -0.51 | -0.31 | $2.30 \times 10^{-13}$ | $9.90 \times 10^{-07}$ | 820/1088 | GDCA |
| 10:53832549-53832549 | *PRKG1* | rs117834398 | T | G | -0.18 | -0.79 | $1.80 \times 10^{-01}$ | $8.30 \times 10^{-11}$ | 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 $3.57 \times 10^{-9}$, 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 $5 \times 10^{-9}$, 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 *PRKG1* 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.

## Sex-specific associations of bile acid serum levels

To investigate whether the genetic component influencing bile acid variation may differ between men and women, we performed sex-specific GWAS meta-analysis of the 14 quantitative (imputed) bile acid traits for ORCADES and CROATIA-Vis cohorts (female N=1088, male N=820) and discovered two sex-specific associations. The association of GDCA with rs73079476 from the *SLCO1B1* locus was significant in male-only GWAS (beta = -0.51, p-value = $2.28 \times 10^{-13}$) (Table 1, Figure 2A). The signal for the same locus in female-only GWAS, while consistent in terms of directionality, has a smaller effect size than in male-only analysis (beta = -0.31) and does not reach the significance threshold (p-value = $9.86 \times 10^{-7}$), despite the slightly higher sample size (Figure 2). This suggests that the genetic effect of *SLCO1B1* locus on the serum levels of GDCA is larger in men than women. We also identified a sex-specific association of GCA at the *PRKG1* locus. In contrast to *SLCO1B1*, the sentinel SNP in *PRKG1*, rs117834398, has a larger effect in females than in males (female beta = -0.79, male beta = -0.18), and passed the significant threshold only in the female-specific analysis (female p-value = $8.26 \times 10^{-11}$, male p-value = $1.81 \times 10^{-1}$) (Table 1, Figure 2B). Interestingly, the sentinel SNPs at the *PRKG1* locus for the full meta-analysis and for the female-specific analysis are in linkage equilibrium ($r^2 < 0.01$) and represent two independent associations in that locus. Overall, none of the significant association identified in one sex was replicated in the other, suggesting that the genetic contribution to serum BA levels is likely to be different in males and females. We have identified 13 additional associations (p-value $< 5 \times 10^{-9}$, Bonferroni adjusted for the number of independent quantitative bile acid traits) that might have sex-specific effects (Supplementary Table 3, Supplementary Figure 1). However, given the low allele frequencies and allele counts in the two analysed cohorts, further analyses are required to replicate these associations.




| COHORT | N | MAF | MAC | P |
|---|---|---|---|---|
| ORCADES MALE | 447 | 0.135 | 121 | 2.26e-06 |
| VIS MALE | 373 | 0.191 | 143 | 4.36e-08 |
| ORCADES FEMALE | 559 | 0.135 | 151 | 1.97e-03 |
| VIS FEMALE | 529 | 0.196 | 208 | 1.54e-04 |
| MALE META-ANALYSIS | 820 | 0.170 | 278 | 2.28e-13 |
| FEMALE META-ANALYSIS | 1088 | 0.168 | 365 | 9.86e-07 |

| COHORT | N | MAF | MAC | P |
|---|---|---|---|---|
| ORCADES MALE | 447 | 0.023 | 21 | 4.45e-01 |
| VIS MALE | 373 | 0.020 | 15 | 2.34e-01 |
| ORCADES FEMALE | 559 | 0.021 | 23 | 3.37e-06 |
| VIS FEMALE | 529 | 0.012 | 12 | 6.27e-07 |
| MALE META-ANALYSIS | 820 | 0.022 | 36 | 1.81e-01 |
| FEMALE META-ANALYSIS | 1088 | 0.019 | 40 | 8.26e-11 |

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

## Link with complex traits and diseases

Next, we assessed whether variants associated with BA levels have been previously associated with any other biochemical traits and diseases. Using Phenoscanner[16,17] we found that rs4149056, sentinel SNP in *SLCO1B1* locus, and its proxies ($r^2 > 0.8$), were also associated with concentration of bilirubin, non-bile acid metabolites, mean corpuscular haemoglobin, sex hormone binding globulin and estrone conjugates, and various responses to drugs (i.e., statin-induced myopathy, LDL-cholesterol response to simvastatin and methotrexate clearance in acute lymphoblastic leukaemia) (Supplementary Table 4). To obtain deeper insight into the causal relationship between BAs and diseases, we conducted bi-directional Mendelian Randomisation (MR) analysis. Using the sentinel SNPs associated with GLCA, GDCA and GCA (Table 1) as instrumental variables we tested whether genetically increased levels of BA influence levels or risk for 548 biochemical traits and diseases available in the IEU Open GWAS database[18] (Supplementary Table 5). Levels of GLCA and GDCA were significantly

(p-value < $0.05/(548x3) = 3.04\times10^{-5}$) associated with different biochemical measurements, such as levels of sex hormone-binding globulin, testosterone, triglycerides, vitamin D, alanine transaminase and galectin-3; with blood traits, such as mean corpuscular haemoglobin and mean corpuscular volume; and with diseases and their risk factors, such as daytime dozing and stroke (Supplementary Table 6). These MR tests were performed using the Wald ratio test utilising only a single instrument, thus the results of causal relationship between BAs and traits/diseases should be interpreted with caution. Yet our results suggest a possible overlap in genetic regulation, involving the *SLCO1B1* locus. Next, to assess whether complex traits and disease could have an effect on bile acid levels, we performed reverse MR using 548 traits/diseases as exposure and bile acids as outcomes. We found no significant associations, suggesting that none of the tested diseases or complex traits have an effect on BA levels (Supplementary Table 7).

## Exome-wide rare variant analysis of bile acids

To assess the contribution of low frequency and rare variants to the bile acid genetic architecture, we performed exome-wide gene-based tests across 18 bile acid traits in the ORCADES cohort (N = 1006) by testing the aggregated effect of rare (MAF <5%) predicted loss-of-function (pLoF) and non-synonymous missense variants. We identified significant association (p-value <$1.79 \times 10^{-7}$) of rare variants from 3 genes with 2 bile acid traits (quantitative CA and binary THDCA). For these associations, a significant p-value was reported by at least 2 of the 4 aggregation tests used. Rare variants significantly associated with quantitative bile acid trait CA are located in the *OR1G1* gene, while those associated with binary bile acid trait THDCA are located in *SART1* and *SORCS2* genes (Table 2, Supplementary Table 8). We further identified significant association of rare variants from *EPS8L1* gene with quantitative bile acid trait DCA and from *EEF2K* with binary bile acid trait THDCA (Supplementary Table 8). However, a significant p-value was reported by only one of the 4 aggregation tests used. Due to the lack of replication across aggregations tests, we considered these associations as not robust.

**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 | *OR1G1* | <0.01 | Missense variants | 2 | SKAT-O | $1.67\text{x}10^{-8}$ | 17 |
| THDCA | Binary | *SORCS2* | <0.05 | pLoF and missense variants | 10 | SMMAT-E | $1.44\text{x}10^{-8}$ | 174 |
| THDCA | Binary | *SART1* | <0.01 | pLoF and missense variants | 4 | SKAT | $1.19\text{x}10^{-7}$ | 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 $1.79 \times 10^{-7}$ (0.05/20,000 estimate of number of genes in the human genome/14 number of independent bile acids).

## Discussion

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

We identified two secondary bile acids (GDCA and GLCA) significantly associated with a locus encompassing the *SLCO1B1* gene. The encoded protein, OATP1B1 (organic anion transporting polypeptide 1B1), is a well-known human hepatocyte transporter mediating the uptake of various endogenous compounds such as bile salts, bilirubin glucuronides, thyroid hormones and steroid hormone metabolites, and also clinically frequently used drugs like statins, HIV protease inhibitors, and the anti-cancer agents irinotecan or methotrexate[19–23]. The sentinel SNP of the *SLCO1B1* locus, rs4149056, is a missense variant (p.Val174Ala) which has been linked by previous GWA studies to blood concentration of several metabolites, including vitamin D[24], triglycerides[25] and bilirubin[26], a compound resulting from the breakdown of haem catabolism and excreted as a major component of bile. This same variant has also been associated with levels of sex hormone-binding globulin and testosterone[27]. The knock-out of the gene in mice results in abnormal liver physiology and abnormal xenobiotic pharmacokinetic phenotypes (Open Targets[28]). A rare variant from the *PRKG1* locus was significantly associated with levels of glycocholic acid (GCA). *PRKG1* encodes a Protein Kinase CGMP-Dependent 1, a protein involved in signal transduction and a key mediator of the nitric oxide/cGMP. The sentinel variant in the region, rs146800892, only passes the MAF threshold (MAF > 0.01) in the CROATIA-Vis cohort, which is therefore the only cohort contributing to this association. Due to its demographic history and geographic position, CROATIA-Vis is a genetic isolate[29] so it is possible that this variant has increased in frequency compared to a general population[30].The mechanism of how the variation within this gene could relate to bile acid levels is unclear and would need to be further investigated.

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

affected by the variant in *PRKG1* than male levels. Later, our Mendelian randomization analysis did not provide evidence that testosterone, oestradiol, sex hormone-binding globulin or other sex-related traits have causal effects on plasma BA levels. While this could be due to a lack of statistical power of our BA meta-analysis, we currently have no evidence to suggest an effect of sex-related hormones on BA levels mediated by genetics. We also detected associations with variants from the same gene, *PRKG1*, in the main, non-stratified analysis. However, the two associations (sex-specific and pooled) appear to be independent (LD $r^2$ <0.001). While the association from the pooled analysis might be either false positive or population-specific, the independent association from the sex-stratified analysis replicates well between two analysed cohorts (CROATIA-Vis and ORCADES).

After assaying common variants through GWAS, we performed exome-wide gene-based association tests in a subset of our samples (N = 1006), to investigate the genetic contribution of rare and low frequency (MAF <5%) coding variants (pLoF and missense) to bile acid levels. Overall, we identified associations with rare variants from 3 genes, *OR1G1*, *SART1* and *SORCS2*. *OR1G1* is an olfactory receptor gene, whose coded protein receptor interacts with odorant molecules in the nose to initiate a neuronal response triggering the perception of smell[31,32]. In addition to the nasal level, the olfactory receptor coded by *OR1G1* is expressed also by enterochromaffin cells, specialised enteroendocrine cells of the gastrointestinal tract. Braun *et al*[33]. determined that certain olfactory cues from spices and odorants, such as thymol, present in the luminal environment of the gut may stimulate serotonin release via olfactory receptors present in enterochromaffin cells. Between 90% and 95% of total body serotonin is in fact synthesised by enterochromaffin cells[34]: serotonin controls gut motility and secretion and is implicated in pathologic conditions such as vomiting, diarrhoea, and irritable bowel syndrome[33]. In mice, gut serotonin was shown to stimulate bile acid synthesis and secretion by the liver and gallbladder. Thus, release of serotonin in response to odorant cues increases bile acid turnover[35]. The hypoxia-associated factor (HAF), encoded by *SART1* gene and also known as SART1(800), is involved in proliferation and hypoxia-related signalling. The protein encoded by *SORCS2* is a receptor for the precursor of nerve growth factor, up-regulation of which has been reported for several liver pathologies, such as hepatotoxin- induced fibrosis[36], ischemia-reperfusion injury[37], oxidative injury[38], cholestatic injury[39] and hepatocellular carcinoma[36,40,41]. However, due to unavailability of exome sequencing data in other cohorts these associations were not replicated.

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

We acknowledge several limitations in the present study. We found only a small percentage of BA variability to be affected by genetics, suggesting that a larger sample size is required to further describe BA genetic architecture. BAs are known to be largely influenced by environmental factors, such as sex and gut microbiota. Female sex and oestrogens are considered relevant regulators of BA production and composition[43,44]. In pregnant women, high levels of circulating oestrogen are associated with development of cholestasis, characterised by increased serum bile acids, likely via oestrogen reducing the expression of BA receptor and transport proteins[45]. Similarly, age-related differences in hormone levels influence the differential production of BAs in women[46]. The relevant impact of sex on plasma BA levels was confirmed by the sex-stratified analysis, where the two significantly associated loci showed to be sex-specific. Similarly, species-composition of gut microbiota has a great impact on BAs levels, especially for secondary BAs that are a direct result of microbiome activity. A recent study describing the effect of gut microbiota on the human plasma metabolome reported that both primary BA cholic acid (CA) and secondary BA deoxycholic acid (DCA) show a high percentage of variance explained by the microbiota ($R^2$ = 30% and 36%, respectively), indicating a strong impact on BAs of the variation in microbiota composition[47]. It is important to interpret our findings in the context of the tissue in which BA levels were measured, blood plasma. Bile acids are synthesised in the liver and secreted into the intestine, to be then reabsorbed into portal circulation and returned to the liver: plasma BA levels thus reflect the amount of BAs escaping extraction from the portal blood. Therefore, levels of BAs in plasma are likely to be influenced by genes other than those encoding the particular anabolic and catabolic enzymes, including those involved in hepatic function and dysfunction. In line with this, the major genetic contributor to blood BA levels in our study are variants from the *SLCO1B1* gene, encoding the hepatocyte transporter OATP1B1 and important for flux of bile salts, bilirubin glucuronides and various hormone metabolites, rather than genes encoding key enzymes of primary BA synthesis, such as *CYP7A1* and *CYP7B1*[48]. Similarly, some of the genes with rare variants associations have been linked to liver diseases, such as liver cancer[49], and intrahepatic cholestasis of pregnancy[50].

In conclusion, we explored the genetic architecture of plasma bile acid levels, including both common and rare variants. By performing GWAS meta-analysis (N = 4923), we identified 2 significantly associated loci, mapping to the *SLCO1B1* and *PRKG1* genes. In the sex-specific GWAS meta-analysis we observed that variants in these genes have different impact on bile acid levels in men and women. To assess relationships between genetically increased levels of bile acids and risk for diseases we performed Mendelian randomisation, but did not find any bile acids affecting disease risk, nor the reverse, which however might be affected by the lack of statistical power. Using the gene-based aggregated tests and whole exome sequencing, we further identified rare pLoF and missense variants in 3 genes associated with BAs, *OR1G1*, *SART1* and *SORCS2*, some of which are known to be involved in liver disease. Additional studies with larger sample sizes and of more diverse ancestry will be necessary to validate our findings, further unravel the genetic architecture of bile acid levels, and to understand their relationship with human diseases and complex traits.

## Materials and methods

### Phenotypic data

#### Bile acids quantification

Bile acid (BA) analysis was performed from plasma or serum (MICROS cohort) samples by liquid chromatography-tandem mass spectrometry (LC-MS/MS) as previously described[51]. The HPLC equipment consisted of a 1200 series binary pump (G1312B), a 1200 series isocratic pump (G1310A) and a degasser (G1379B) (Agilent, Waldbronn, Germany) connected to an HTC Pal autosampler (CTC Analytics, Zwingen, CH). A hybrid triple quadrupole linear ion trap mass spectrometer API 4000 Q-Trap equipped with a Turbo V source ion spray operating in negative ESI mode was used for detection (Applied Biosystems, Darmstadt, Germany). High purity nitrogen was produced by a nitrogen generator NGM 22-LC/MS (cmc Instruments, Eschborn, Germany). Gradient chromatographic separation of BAs was performed on a 50 mm × 2.1 mm (i.d.) Macherey-Nagel NUCLEODUR C18 Gravity HPLC column, packed with 1.8 µm particles equipped with a 0.5 µm pre-filter (Upchurch Scientific, Oak Harbor, WA, USA). The injection volume was 5 µL and the column oven temperature was set to 50 °C. Mobile phase A was methanol/water (1/1, v/v), mobile phase B was 100% methanol, both containing 0.1% ammonium hydroxide (25%) and 10 mmol/L ammonium acetate (pH 9). A gradient elution was performed with 100% A for 0.5 min, a linear increase to 50% A until 4.5 min, followed by 0% A from 4.6 until 5.5 min and re-equilibration from 5.6 to 6.5 min with 100% A. The flow rate was set to 500 µL/min. To minimize contamination of the mass spectrometer, the column flow was directed only from 1.0 to 5.0 min into the mass spectrometer using a diverter valve. Otherwise, methanol with a flowrate of 250 µL/min was delivered into the mass spectrometer. The turbo ion spray source was operated in the negative ion mode using the following settings: Ion spray voltage = −4500 V, ion source heater temperature = 450 °C, source gas 1 = 40 psi, source gas 2 = 35 psi and curtain gas setting = 20 psi. Analytes were monitored in the multiple reaction monitoring (MRM). Quadrupoles Q1 and Q3 were working at unit resolution. Calibration was achieved by the addition of BAs to EDTA-plasma/serum. A combined BA standard solution containing the indicated amounts (0.5 - 70.5 µmol/L) was placed in a 1.5 ml tube and excess solvent was evaporated under reduced pressure before adding EDTA-plasma/serum. Calibration curves were calculated by linear regression without weighting. Data analysis was performed with Analyst Software 1.4.2. (Applied Biosystems, Darmstadt, Germany). The data were exported to Excel spreadsheets and further processed by self-programmed Excel macros which sort the results, calculate the analyte/internal standard peak area ratios, generate calibration lines and calculate sample concentrations. For the calculation we selected the internal standard with analogous fragmentation and closest retention time to the respective BA species.

#### Pre-processing of bile acid traits

Prior to genetic analysis, bile acid traits were grouped into three groups based on the percentage of samples with below the limit of detection (<LOD) measurements: high <LOD group (> ~30% of all samples below LOD) and low <LOD group (< ~7% of all samples below LOD) (Supplementary Table 1). Accordingly, different phenotypic pre-processing and different analysis strategies were applied to the groups. Bile acids within a high <LOD were considered as binary traits: individuals were categorised based on whether their bile acid levels were effectively measured (category 1) or were below the LOD (category 0). Bile acid traits belonging to this group were THDCA, TUDCA, TCA and GHDCA. All other bile acids were considered as quantitative traits and were $\log_{10}$-transformed. However, to increase the sample size, in addition to a complete-case analysis (considering as missing all samples with <LOD), we also imputed <LOD measurements. For each bile acid, imputation of <LOD measurements was performed by fitting a truncated normal distribution, with mean and standard deviation of the effectively measured raw data, truncated (as an upper bound) to the lowest measured value for the given bile acid. To do so, we used the "rtnorm" function from the MCMCglmm R package[52]. After imputation, measurements were $\log_{10}$-transformed.

## Genome-wide association analysis

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

## Meta-analysis

Prior to meta-analysis, cohort-level GWAS were quality controlled using the EasyQC software package, following the protocol described in Winkler *et al.*[54] Cohort-level results were corrected for the genomic control inflation factor, then pooled and analysed with METAL v2011-03-25 software[55], applying the fixed-effect inverse-variance method. The mean genomic control inflation factor after the meta-analysis was 0.991 (range 0.938 – 1.009), suggesting that the confounding effects of the family structure were correctly accounted for (Supplementary Table 12). The standard genome-wide significance threshold was Bonferroni corrected for the number of independent bile acid traits, calculated as 14 ($5x10^{-8}/14 = 3.57x10^{-9}$). The number of independent bile acid traits was estimated as the sum of the number of binary traits (4) and the number of principal components that jointly explained 99% of the total variance of $log_{10}$-transformed quantitative traits in each cohort (10) (Supplementary Table 13).

## Sex-stratified GWAS meta-analysis

To identify possible differences in the genetic contribution to bile acid variability between men and women, we performed sex-specific GWAS of the 14 quantitative bile acid traits for ORCADES and CROATIA-Vis cohorts. Given that for the sex-stratified GWAS we implicitly halve our sample size, we performed these analyses only on the imputed bile acid traits. The same analysis steps and procedures already described for the full meta-analysis were applied. Bile acid traits were adjusted for age, sex and batch as fixed effects, and relatedness (estimated as the kinship matrix calculated from genotyped data) as a random effect in a linear mixed model, calculated using the 'polygenic' function from the GenABEL R package[56]. Residuals of covariate and relatedness correction were tested for association with HRC-imputed[53] SNP dosages using the RegScan v0.5 software[57], applying an additive genetic model of association. Prior to meta-analysis, SNPs having a difference in allele frequency between the two cohorts higher than ±0.3 or a minor allele count (MAC) lower or equal to 6 were filtered out. Cohort-level GWAS were corrected for genomic control inflation factor and then meta-analysed (N =820 for male and N =1088 for female individuals) using METAL v2011-03-25 software[55], applying the fixed-effect inverse-variance method. The mean $\lambda_{GC}$ was 0.993 (range 0.978–1.011) for male-specific meta-analysis and 0.996 (range 0.984–1.003) for female-specific meta-analysis. The Bonferroni-corrected significance threshold applied is $5 \times 10^{-9}$.

## Phenoscanner and Mendelian Randomisation

To assess link between bile acids and diseases we explored the overlap of SNPs associated with BAs with complex human traits by using PhenoScanner v1.1 database[16,17], taking into account significant genetic association ($p < 5 \times 10^{-9}$) at the same or strongly (LD $r2 > 0.8$) linked SNPs in populations of European ancestry. We then performed bi-directional Mendelian Randomisation (MR) to investigate the effect of 548 complex traits and diseases available in the IEU Open GWAS database[18] (manually curated list of studies from identifiers ebi-a, ieu-a, ieu-b and ukb-a; the complete list reported in the Supplementary Table 5) on BA levels, and

vice-versa. The set of genome-wide significant, LD clumped SNPs used as instruments for complex traits/diseases was extracted from the selected studies by using the "extract_instruments" function from the TwoSampleMR 0.5.6 R package[58]. Similarly, sentinel SNPs from BAs meta-analysis (Supplementary Table 2) were selected as instruments. MR tests were performed by using fixed effects inverse variance-weighted (IVW) in case of multiple instruments or Wald Ratio method in case of a single instrument, as implemented in the TwoSampleMR 0.5.6 R package[58]. Multiple testing correction was controlled for using either the Bonferroni correction or false discovery rate (FDR).

## Whole-exome sequencing data

### Exome sequencing

The "Goldilocks" exome sequence data for ORCADES cohort was prepared at the Regeneron Genetics Center, following the protocol detailed in Van Hout *et al.*[59] for the UK Biobank whole-exome sequencing project. In summary, sequencing was performed using S2 flow cells on the Illumina NovaSeq 6000 platform with multiplexed samples. DNAnexus platform[60] was used for processing raw sequencing data. The files were converted to FASTQ format and aligned using the BWA-mem[61] to GRCh38 genome reference. The Picard tool[62] was used for identifying and flagging duplicated reads, followed by calling the genotypes for each individual sample using the WeCall variant caller[63]. During quality control, 33 samples genetically identified as duplicates, 3 samples showing disagreement between genetically determined and reported sex, 4 samples with high rates of heterozygosity or contamination, 2 samples having low sequence coverage (less than 80% of targeted bases achieving 20X coverage) and 1 being discordant with genotyping chip were excluded. Finally, the "Goldilocks" dataset was generated by (i) filtering out genotypes with read depth lower than 7 reads, (ii) keeping variants having at least one heterozygous variant genotype with allele balance ratio greater than or equal to 15% (AB $\geq$ 0.15) or at least one homozygous variant genotype, and (iii) filtering out variants with more than 10% of missingness and HWE $p<10^{-6}$. Overall, a total of 2,090 ORCADES (820 male and 1,270 female) participants passed all exome sequence and genotype quality control thresholds. A pVCF file containing all samples passing quality control was then created using the GLnexus joint genotyping tool[64].

### Variant annotation

Exome sequencing variants were annotated as described in Van Hout, *et al.*[59] Briefly, they were annotated with the most severe consequence across all protein-coding transcripts using SnpEff[65]. Gene regions were defined based on Ensembl release 85[66]. Predicted loss-of function (pLoF) variants were defined as variants annotated as start lost, stop gained/lost, splice donor/acceptor and frameshift. The deleteriousness of missense variants was based on dbNSFP 3.2[67,68] and assessed using the following algorithms: (1) SIFT[69]: "D" (Damaging), (2)

Polyphen2_HDIV: "D" (Damaging) or "P" (Possibly damaging), (3) Polyphen2_HVAR[70]: "D" (Damaging) or "P" (Possibly damaging), (4) LRT[71]: "D" (Deleterious) and (5) MutationTaster[72]: "A" (Disease causing automatic) or "D" (Disease causing). If not predicted as deleterious by any of the algorithms the missense variants were considered "likely benign", "possibly deleterious" if predicted as deleterious by at least one of the algorithms and "likely deleterious" if predicted as deleterious by all five algorithms.

## Exome-wide gene-based aggregation analysis of rare variants

### Generation of gene masks

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

### ORCADES gene-based aggregation analysis

We performed variant Set Mixed Model Association Tests (SMMAT)[73] on the 18 bile acid traits from ORCADES cohort, quantified and pre-processed as previously described, fitting a GLMM adjusting for age, sex, batch, and familial or cryptic relatedness by kinship matrix. The kinship matrix was estimated from the genotyped data using the 'ibs' function from GenABEL R package[56]. The SMMAT framework includes 4 variant aggregate tests: burden test, sequence kernel association test (SKAT), SKAT-O and SMMAT-E, a hybrid test combining the burden test and SKAT. The 4 variant aggregate tests were performed on 8 different pools of genetic variants, called "masks", each one including a different set of variants based on both MAF and predicted consequence of variants (e.g., loss of function and missense) (Supplementary Table 9), as described above. Discovery significance threshold was Bonferroni corrected for the rough estimate of the number of genes in the human genome, 20,000, and the number of independent bile acid traits, 14, calculated as previously described ($0.05/20000/14 = 1.79 \times 10^{-7}$). A gene association was considered significant if it passed the above reported Bonferroni corrected significance threshold in at least two of the 4 performed variant aggregate tests and if the cumulative allele count of the variants included in the gene was equal or higher than 10.

## Code availability

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

## Data availability

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

## References

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

1B1: a genetically polymorphic transporter of major importance for hepatic drug uptake. *Pharmacol Rev* **63**, 157–181 (2011).

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

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

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

## Acknowledgments

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

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

## Ethics

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

## Author contributions

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

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

## Competing interests

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