Characterization of gut microbial structural variations as determinants of human bile acid metabolism

Wang, Daoming; Doestzada, Marwah; Chen, Lianmin; Andreu-Sánchez, Sergio; van den Munckhof, Inge C L; Augustijn, Hannah E; Koehorst, Martijn; Ruiz-Moreno, Angel J; Bloks, Vincent W; Riksen, Niels P

Published in:
Cell Host & Microbe

DOI:
10.1016/j.chom.2021.11.003

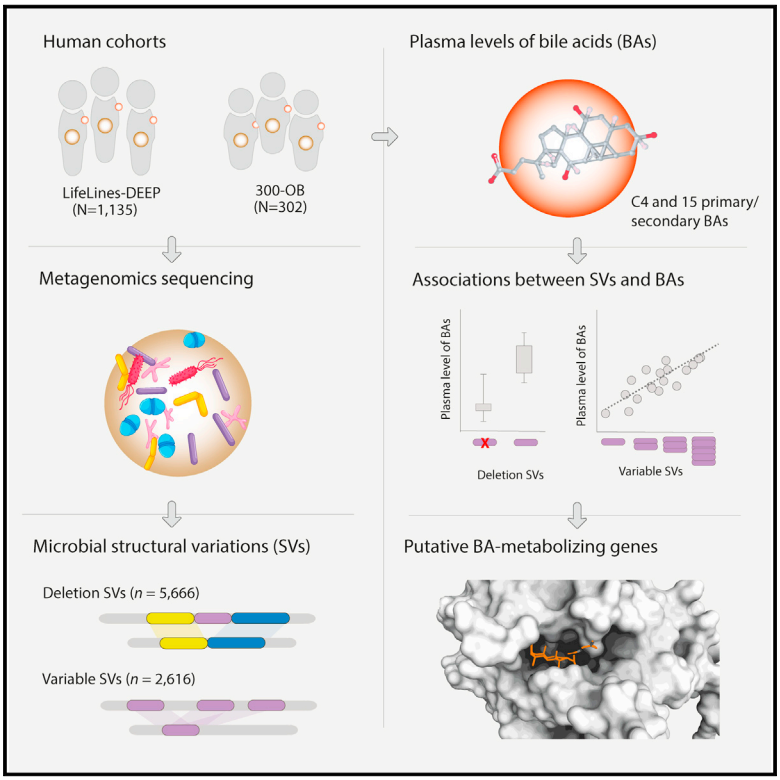
Publication date:
2021

Resource

Cell Host & Microbe

Characterization of gut microbial structural variations as determinants of human bile acid metabolism

Graphical abstract



Highlights

- Structural variations (SVs) underpin the bacterial populational genetic structure
- Numerous microbial SVs are associated with plasma bile acid composition
- SV-based metagenome analysis enables discovery of bile acid transformation genes
- Lifestyle impacts human bile acid metabolism via modulation of microbial genetics

Authors

Daoming Wang, Marwah Doestzada, Lianmin Chen, ..., Alexandra Zhernakova, Folkert Kuipers, Jingyuan Fu

Correspondence

j.fu@umcg.nl

In brief

Wang et al. identify and characterize structural variants in human gut bacterial genomes involved in human bile acid metabolism that may act as mediators that regulate the impact of lifestyle factors on bile acid metabolism.

Wang et al., 2021, Cell Host & Microbe *29*, 1802–1814
December 8, 2021
https://doi.org/10.1016/j.chom.2021.11.003

CellPress

CellPress

Cell Host & Microbe

Resource

Characterization of gut microbial structural variations as determinants of human bile acid metabolism

Daoming Wang,[1,2] Marwah Doestzada,[1,2,8] Lianmin Chen,[1,2,8] Sergio Andreu-Sánchez,[1,2] Inge C.L. van den Munckhof,[3] Hannah E. Augustijn,[1,2] Martijn Koehorst,[2,4] Angel J. Ruiz-Moreno,[1,2] Vincent W. Bloks,[2] Niels P. Riksen,[3] Joost H.W. Rutten,[3] Leo A.B. Joosten,[3,6] Mihai G. Netea,[3,5,7] Cisca Wijmenga,[1] Alexandra Zhernakova,[1] Folkert Kuipers,[2,4] and Jingyuan Fu[1,2,9,*]

[1]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen 9713AV, the Netherlands
[2]University of Groningen, University Medical Center Groningen, Department of Pediatrics, Groningen 9713AV, the Netherlands
[3]Department of Internal Medicine and Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen 6500HB, the Netherlands
[4]University of Groningen, University Medical Center Groningen, Department of Laboratory Medicine, Groningen 9713AV, the Netherlands
[5]Department for Genomics & Immunoregulation, Life and Medical Sciences Institute, University of Bonn, Bonn 53113, Germany
[6]Department of Medical Genetics, Iuliu Haţieganu University of Medicine and Pharmacy, Cluj-Napoca 400000, Romania
[7]Human Genomics Laboratory, Craiova University of Medicine and Pharmacy, Craiova 200349, Romania
[8]These authors contributed equally
[9]Lead contact
*Correspondence: j.fu@umcg.nl
https://doi.org/10.1016/j.chom.2021.11.003

SUMMARY

Bile acids (BAs) facilitate intestinal fat absorption and act as important signaling molecules in host-gut microbiota crosstalk. BA-metabolizing pathways in the microbial community have been identified, but it remains largely unknown how the highly variable genomes of gut bacteria interact with host BA metabolism. We characterized 8,282 structural variants (SVs) of 55 bacterial species in the gut microbiomes of 1,437 individuals from two cohorts and performed a systematic association study with 39 plasma BA parameters. Both variations in SV-based continuous genetic makeup and discrete clusters showed correlations with BA metabolism. Metagenome-wide association analysis identified 809 replicable associations between bacterial SVs and BAs and SV regulators that mediate the effects of lifestyle factors on BA metabolism. This is the largest microbial genetic association analysis to demonstrate the impact of bacterial SVs on human BA composition, and it highlights the potential of targeting gut microbiota to regulate BA metabolism through lifestyle intervention.

INTRODUCTION

Bile acids (BAs) represent an important class of biologically active metabolites that act at the interface between the host and gut microbiota. BAs are amphiphilic steroids synthesized from cholesterol in the liver and are well known for their roles in facilitating intestinal fat absorption, promoting hepatic bile formation, and maintaining whole-body cholesterol balance. In addition, BAs exert hormone-like functions by signaling via membrane-bound and nuclear receptors involved in the control of lipid, glucose, and energy metabolism (Kuipers et al., 2014). Altered BA metabolism has been associated with several metabolic diseases, including cardiometabolic diseases (Chávez-Talavera et al., 2017; Chen et al., 2020; Steiner et al., 2011), colorectal cancer (Dermadi et al., 2017), and hepatocellular carcinoma (Gao et al., 2019), as well as aging (Sato et al., 2021). Therefore, BAs and their signaling pathways have emerged as attractive therapeutic targets in the treatment of metabolic diseases (Đanić et al., 2018).

Gut bacteria are essential players in human BA metabolism: bacterial bile salt hydrolases (BSH) convert the glycine- and taurine-conjugated primary BAs produced by the liver (cholic [CA] and chenodeoxycholic [CDCA] acids) into unconjugated primary BAs that can subsequently be dehydroxylated to form secondary BAs (deoxycholic [DCA] and lithocholic [LCA] acids) (Jia et al., 2018). BAs are efficiently absorbed in the ileum, and to a lesser extent, in the colon, and return to the liver via the portal venous system for re-secretion into the bile (Kuipers et al., 2014). Consequently, the BA pool consists of a mixture of primary and secondary BAs that travel between the liver and the intestine within the enterohepatic circulation. In turn, BAs can themselves also influence gut microbiome composition through their antimicrobial activities and via indirect signaling pathways (Jia et al., 2018). Enthusiasm for identifying BA-related microbial species and genes has been rising since it may allow the design

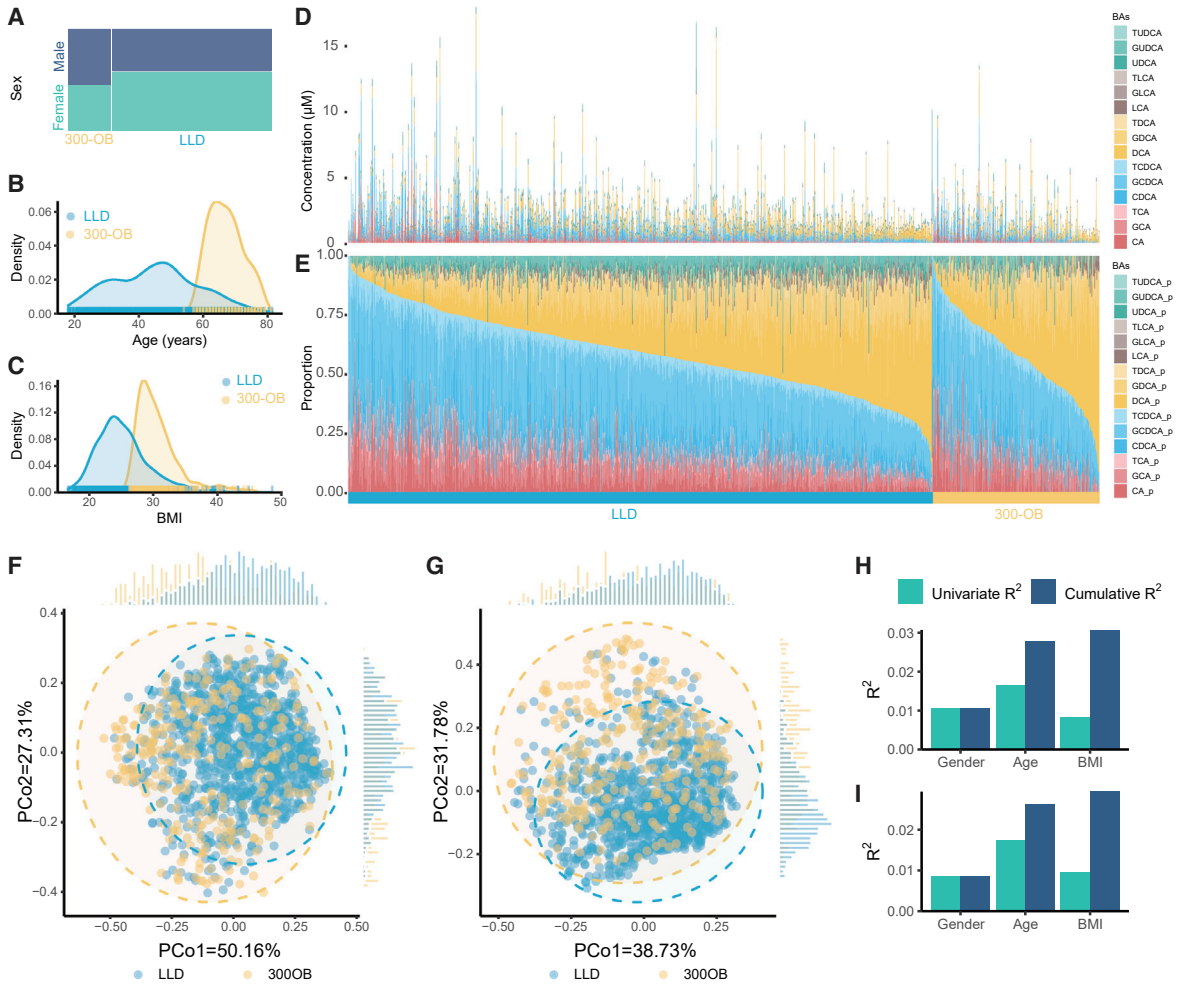


Figure 1. High variability in human fasting plasma bile acid concentration and composition

(A) Sex proportions of LLD and 300-OB.

(B) Age distribution in LLD and 300-OB.

(C) BMI distribution in LLD and 300-OB.

(D) Concentrations of 15 bile acids (BAs) in fasting plasma across all samples of LLD and 300-OB.

(E) Proportions of 15 BAs in plasma across all samples of LLD and 300-OB. Samples were sorted by the proportion of the primary BAs within each cohort. The order of samples is identical in (D) and (E).

(F and G) Principal coordinates analysis (PCoA) plot of the differences between all samples based on BA concentration profile and BA proportion profile.

(H and I) Explained variance proportions (R^2) of BA concentration (H) and proportion (I) profiles by sex, age, and BMI. Blue bars indicate the cumulative explained BA variance proportion in multivariate models. Green bars indicate individually explained BA variance proportion by each factor in univariate models. See also Figure S1 and Table S1.

of BA-targeted therapeutic approaches. For instance, the BA pool has been associated with gut microbial composition in human cohorts (Chen et al., 2020; Gu et al., 2017). Bacterial genes involved in BA biotransformation have been identified through experimental and homologue-based bioinformatic approaches (Heinken et al., 2019; Song et al., 2019). However, a considerable proportion of BA-related genes are still unknown (Heintz-Buschart and Wilmes, 2018).

Microbial structural variants (SVs) are highly variable segments of bacterial genomes that have been defined in recent years based on metagenomic sequencing data (Zeevi et al., 2019). Microbial SV regions potentially contain functional genes involved in host-microbe interactions; thus, they could provide information on the sub-genome resolution of bacterial functionality. A variety of associations have been found between SVs and metabolite levels in human blood (Zeevi et al., 2019). Recently, a longitudinal study comparing subjects with irritable bowel syndrome with healthy individuals was the first to report associations between BAs and microbial SVs (Mars et al., 2020). In this study, fecal levels of two unconjugated primary BA species, CDCA and CA, were found to correlate with variable genomic segments of *Blautia wexlerae*. This finding provided the initial clue that previously unknown bacterial genes are involved in the modification of primary BAs or indirectly associated with host BA metabolism (Mars et al., 2020). However, in view of the limited sample size and number of individual BA species

analyzed in this study and the unknown reproducibility of the associations between SVs and BAs across different cohorts, a systematic analysis in large-scale, population-based cohorts is required. Moreover, although the BA-associated SVs were interpreted as potential BA-metabolizing genomic segments (Mars et al., 2020), the existence of a causal relationship between BAs and microbial variants remains to be established because BAs can also act as regulators of the gut microbiome.

Therefore, we aimed to systematically evaluate the relationships between structural variation of the gut microbiome and human BA metabolism. This study involved 1,437 individuals from two independent Dutch cohorts: the population-based LifeLines-DEEP cohort (LLD, n = 1,135) (Tigchelaar et al., 2015) and the 300-Obesity cohort (300-OB, n = 302) (Horst et al., 2020) (Figures 1A–1C; Table S1). We performed a systematic microbial genetic association analysis of BAs between 39 BA parameters and 8,282 SVs. We further integrated several lifestyle factors, including diet, drug usage, and smoking, and constructed tripartite networks of *in silico*-inferred causal relationships that included exposures, microbial genetics, and host plasma BA composition. This identified potential microbial genetic regulators that mediate the effect of lifestyle on BA metabolism, which supports the potential of targeting the gut microbiome to alter human BA metabolism.

RESULTS

High variability of plasma BA composition between individuals and cohorts

In two independent Dutch cohorts (Figures 1A–1C; Table S1), we assessed the concentrations and proportions of 15 BA species (6 primary and 9 secondary BAs) in fasting plasma: CA, CDCA, LCA, DCA, ursodeoxycholic acid (UDCA), and their glycine- or taurine-conjugated forms (Table S1). We also computed 8 ratios that reflect hepatic and bacterial enzymatic activities and quantified the plasma level of C4, a biomarker of hepatic BA biosynthesis (Table S1; STAR Methods) (Chiang, 2017). In total, we obtained 39 plasma BA parameters in this study.

Both the concentrations and proportions of the 15 BA species showed considerable inter-individual variation in both cohorts (Figures 1D and 1E). Plasma BA composition showed a significant difference between the two cohorts (permutational multivariate analysis of variance [PERMANOVA], $p < 0.001$; Figures 1F and 1G), with 34 out of 39 BA parameters showing significantly different abundance (Wilcoxon rank-sum test, false discovery rate [FDR] <0.05; Figure S1A; Table S1). However, age, sex, and body mass index (BMI) collectively explained only 3.07% of the variance in BA concentration and 2.94% in BA proportion, respectively (Figures 1H and 1I). This indicates that a large proportion of BA variation remains unexplained and may be attributed to other factors, such as lifestyle factors, host genetic background, and gut microbial factors.

Bacterial SV profiling

We detected a total of 8,282 SVs in 55 reference species genomes, including 2,616 variable SVs (vSVs) and 5,666 deletion SVs (dSVs) (STAR Methods), with 32–374 SVs per species (Figures 2A and 2B; Table S2). These 55 species together accounted, on average, for 82.52% of the total microbial composition, ranging from 43.73%–94.71% (Figure S2A). The average number of samples with enough coverage to call microbial SV of the 55 species was 432 (Figure S2B; Table S2). The bacterial species with the most SVs included *B. wexlerae*, *Eubacterium rectale*, *Eubacterium hallii*, and *Ruminococcus* sp. SR1/5.

We further assessed the Canberra distance of bacterial SV profiles between all samples (Figure 2C). Principal coordinates (PCo) 1 and 2 together explained 20.70% of the total SV-based genetic variance (Figure 2C), which showed significant differences between LLD and 300-OB (Wilcoxon rank-sum test, $p = 9.83 \times 10^{-4}$ for PCo1 and $p = 2.62 \times 10^{-11}$ for PCo_2). Microbial abundance could explain 6.45% of the observed genetic variance (PERMANOVA, $p < 0.001$; Figure S2C; STAR Methods). After correcting for microbial abundance, the cohort itself still significantly contributed to genetic differences (PERMANOVA, $p < 0.001$; Figure S2C) and the genetic PCo1 and PCo_2 were significantly different between the cohorts (Wilcoxon rank-sum test, $p = 1.36 \times 10^{-2}$ for PCo1 and $p = 1.83 \times 10^{-5}$ for PCo_2), indicating a divergence of microbial genetics between the two cohorts that was independent from differences in their microbial abundances. Interestingly, age, gender, BMI, and read counts collectively explained only 1.79% of the variance of the metagenome-wide SV profile (Figure S2C).

Bacterial genetic associations to BAs are independent from taxonomic abundance

We first investigated associations between BA levels and the abundance of species (Figure S3A) and identified 407 significant associations that involved 50 bacterial species and 36 BA parameters ($FDR_{meta} < 0.05$; Figure 3A; Table S3). The most significant abundance association was found between *Clostridium saudiense* and the CA proportion in plasma ($Beta_{meta} = 0.36$, $FDR_{meta} = 1.94 \times 10^{-46}$; Table S3). Our result confirmed many previous findings, such as the negative association of the butyrate-producing species *Faecalibacterium prausnitzii* with the CA dehydroxylation/deconjugation ratio ($Beta_{meta} = -0.21$, $FDR_{meta} = 1.15 \times 10^{-12}$; Table S3) (Chen et al., 2020). The positive association between *E. hallii*, another butyrate-producing species, and C4 concentration ($Beta_{meta} = 0.11$, $FDR_{meta} = 5.63 \times 10^{-5}$; Table S3) is consistent with the previous finding that *E. hallii* could modify BA metabolism in mice (Udayappan et al., 2016).

In addition to species abundance, the genetic makeup of species may also be relevant to BA metabolism. Therefore, we constructed a SV-based populational structure of the genetic makeup for each species and identified 245 significant associations between the genetic makeup of 37 bacterial species and 35 BA parameters (PERMANOVA, $FDR_{meta} < 0.05$; Figures 3A and S3B; Table S3), after correcting for age, sex, BMI, read counts, and corresponding species abundance. Interestingly, of the 245 BA associations with species-specific genetic makeup, only 81 were also detected at the species-abundance level (Figure S4C), which highlights that microbial genetic variation represents an extra layer of information about bacterial functionality.

The species with the highest number of genetic associations was *B. wexlerae*. The inter-individual genetic differences of *B. wexlerae* were significantly associated with 28 BA parameters (PERMANOVA, $FDR_{meta} < 0.05$; Table S3), with the strongest association with plasma CA proportion ($P_{meta} = 8.70 \times 10^{-6}$; Figure 3B; Table S3), whereas only 12 BA parameters correlated

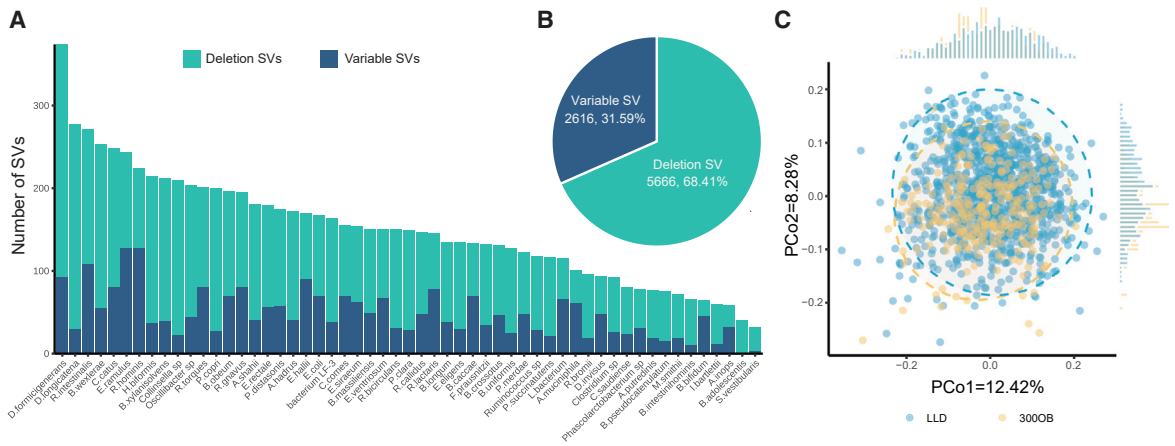


Figure 2. Overview of structural variation profile in LLD and 300-OB

(A) Number of structural variants (SV) of each species.

(B) Total SV numbers.

(C) Population structure of SV-based genetic makeup. See also Figure S2 and Table S2.

with the relative abundance of *B. wexlerae* (Linear regression, $FDR_{meta} < 0.05$; Table S3). Another species, *F. prausnitzii*, contributes to 12-dehydro-CA production, and the depletion of *F. prausnitzii* was inferred to lower the unconjugated CA and CDCA levels in the feces of IBD patients (Heinken et al., 2019). In addition to associations at the species-abundance level, genetic differences in *F. prausnitzii* were also associated with 12 BA parameters (Table S3). For instance, genetic differences in *F. prausnitzii* correlated with the proportion of glycoursodeoxycholic acid (GUDCA) in plasma (PERMANOVA, $FDR_{meta} < 0.05$; Figure 3C; Table S3), but no significant association was found at the abundance level. Together, we observed that species-specific genetic makeup correlates with BA composition independent of their relative abundances.

Discrete populational genetic clusters correlate with human BA metabolism

Based on observed bacterial genetic distance, we stratified the population genetic structure for each species (STAR Methods) and detected two or more distinct clusters for 29 of the 55 species (Figure S4; Table S4). Interestingly, different clusters from *Prevotella copri*, *Streptococcus vestibularis*, and *Parabacteroides merdae* showed different enrichments in LLD and 300-OB (chi-square test, FDR < 0.05, Figure S5; Table S4). We also detected 41 significant associations between species clusters and BAs (Permutational Kruskal-Wallis rank-sum test, FDR < 0.05; Figure 4; Table S4). *E. rectale* showed two distinct clusters that had the most associations (10 associations), with the top association being with C4 concentration (Permutational Kruskal-Wallis rank-sum test, $FDR = 2.26 \times 10^{-5}$). We compared the SV profiles of two clusters of *E. rectale* and found that 55 of the 56 vSVs and 72 of the 124 dSVs were enriched differently between the two clusters (Wilcoxon rank-sum test for vSV and χ^2 test for dSV, FDR < 0.05).

Metagenome-wide SV-based associations point to known and putative BA genes

To identify SVs that potentially harbor genes related to BA metabolism, we performed a metagenome-wide microbial SV-based association analysis. Considering the cohort heterogeneity, we performed the association per cohort, followed by meta- and heterogeneity analysis (random effect model). In addition to age, sex, BMI, and total read counts, we also included the corresponding species abundances as a covariate to correct for the impact of species abundance (*model 1*). Additionally, to disentangle the individual SV effect from the genetic lineage effect, i.e., to correct for strong bacterial population structure and linkage disequilibrium among variants, we further included the top genetic principal components (PCs) of each species as covariates in the linear model to correct for the lineage effect (*model 2*; STAR Methods; Table S5).

In total, we identified 809 significant and consistent associations ($FDR_{meta} < 0.05$) between 321 SVs from 37 species and 34 BA parameters (Figure 5A), including 755 associations identified by model 1 and 177 associations identified by *model 2* (Figure S6A; Table S5). *Coprococcus comes* showed the highest number of associations (Figure 5B), followed by *E. rectale*, *Blautia obeum*, and *B. wexlerae* (Figure 5C). The effect sizes and directions of all 809 associations were highly consistent between cohorts (Figures S6B–S9E). These results indicate that the SV associations we identified were robust and replicable between the two cohorts despite the large differences in their profiles of gut microbial genetic makeup and plasma BA composition.

Notably, 123 associations were significantly detected in both models (Figure S6A). The strongest association was observed between the secondary/primary BA ratio and a 5-kbp vSV (2,932–2,935 and 2,935–2,937 kbp) of *C. comes* (*model 2*, $P_{meta} = 1.98 \times 10^{-28}$; Figures 5D and 5E; Table S5). Strikingly, a BSH gene is found to be close to this SV region (Figure 5D). This BSH gene encodes the enzyme that catalyzes the deconjugation of glycine- and taurine-conjugated BAs and was also found in associated SV regions of other species, such as at the genomic segment (2,081–2,082 kbp) of *B. wexlerae* (Figure 5F) that was significantly associated with 5 BA parameters (Figure 5C), including a negative association with the DCA proportion (Figure 5G; Table S5).

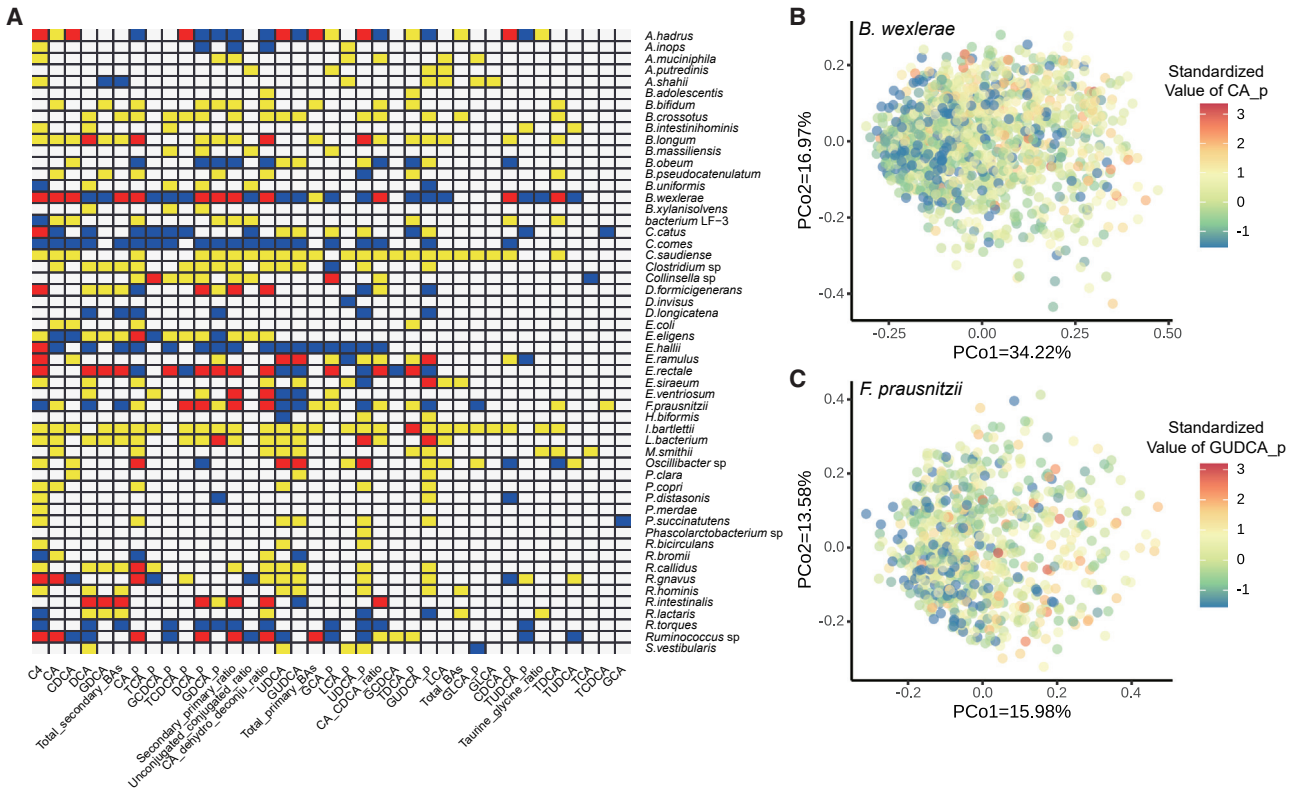


Figure 3. Species-level associations of gut microbiome with human bile acid parameters

(A) Heatmap of species-level associations with BA parameters. Blue indicates purely genetics-based associations. Yellow indicates purely relative abundance-based associations. Red indicates associations based on both genetics and relative abundance. Black indicates genetics-based associations where relative abundance is not available for the corresponding species. White indicates no association.

(B) Genetic association of *B. wexlerae* with CA proportion in plasma. The color scale from red to blue represents the increase in standardized value of CA proportion.

(C) Genetic association of *F. prausnitzii* with GUDCA proportion in plasma. The color scale from red to blue represents the increase in the standardized value of GUDCA proportion. See also Figure S3 and Table S3.

The analysis model correcting for lineage effects (*model 2*) detected fewer associations than the model without the correction (*model 1*), suggesting that spurious associations induced by lineage effects are likely to have been removed. However, this model may also have lost the power to identify bacterial SVs that contribute to both lineage effects and BA metabolism (Earle et al., 2016). For instance, near a BSH gene (Figure 5H) in the genome of *Eubacterium ventriosum*, 4 variable SVs were significantly associated with 9 BA parameters identified by *model 1* ($FDR_{meta} < 0.05$) (Table S11), with the most significant association being between the vSV region 1,512–1,517 kbp and the secondary/primary BA ratio (*model 1*, $Beta_{meta} = 0.43$; $P_{meta} = 3.39 \times 10^{-8}$; Figure 5I). However, these associations were not captured by *model 2*. On the other hand, *model 2* also reported 54 associations that were not significant in *model 1* (Figure S6A), suggesting that some SV effects could be masked by the lineage effect.

Besides the BSH genes, we found several other bacterial BA biotransformation genes in BA-associated regions (STAR Methods). For instance, 7β-hydroxysteroid dehydrogenase (7βHSDH) catalyzes the biotransformation between 7-dehydro-CDCA and UDCA (Heinken et al., 2019), and we found a 7βHSDH gene located in a vSV of *Ruminococcus torques* (1,671–1,673 and 1,673–1,677 kbp; Figure 6A) that correlated with UDCA proportion (*model 1*, $Beta_{meta} = -0.14$, $P_{meta} = 2.70 \times 10^{-4}$; Table S5). We also identified some bacterial genes with putative hydroxysteroid dehydrogenase (HSDH) functionality in BA-associated regions, including CK1_08630 (UniProt: D4LGZ3) in *Ruminococcus* sp. SR1/5 (Figure 6B), RUMOBE_03494 (UniProt: A5ZWV0) in *B. obeum* (Figure 6C), and EUBHAL_00727 (UniProt: C0ETJ6) in *E. hallii* (Figure 6D). CK1_08630 and RUMOBE_03494 encode the same amino acid sequences (248 amino acids) and are annotated as oxidoreductases of the short-chain dehydrogenase/reductase family, whereas EUBHAL_00727 has only 82 amino acids and is functionally uncharacterized. We used AlphaFold2 (Jumper et al., 2021) to predict high-quality 3D structures of these proteins and then compared them with the entire protein databank (PDB). This led to the identification of the homotetrameric form of the 7α-hydroxysteroid dehydrogenase (7αHSDH) of *Escherichia coli* complexed with reduced nicotinamide adenine dinucleotide (NADH) and 7-oxo glycochenodeoxycholic acid (GCDCA) (PDB: 1FMC; Figure 6E). The structural alignment of the CK1_08630/RUMOBE_03494 and EUBHAL_00727 models against the reference 7αHSDH crystals showed that

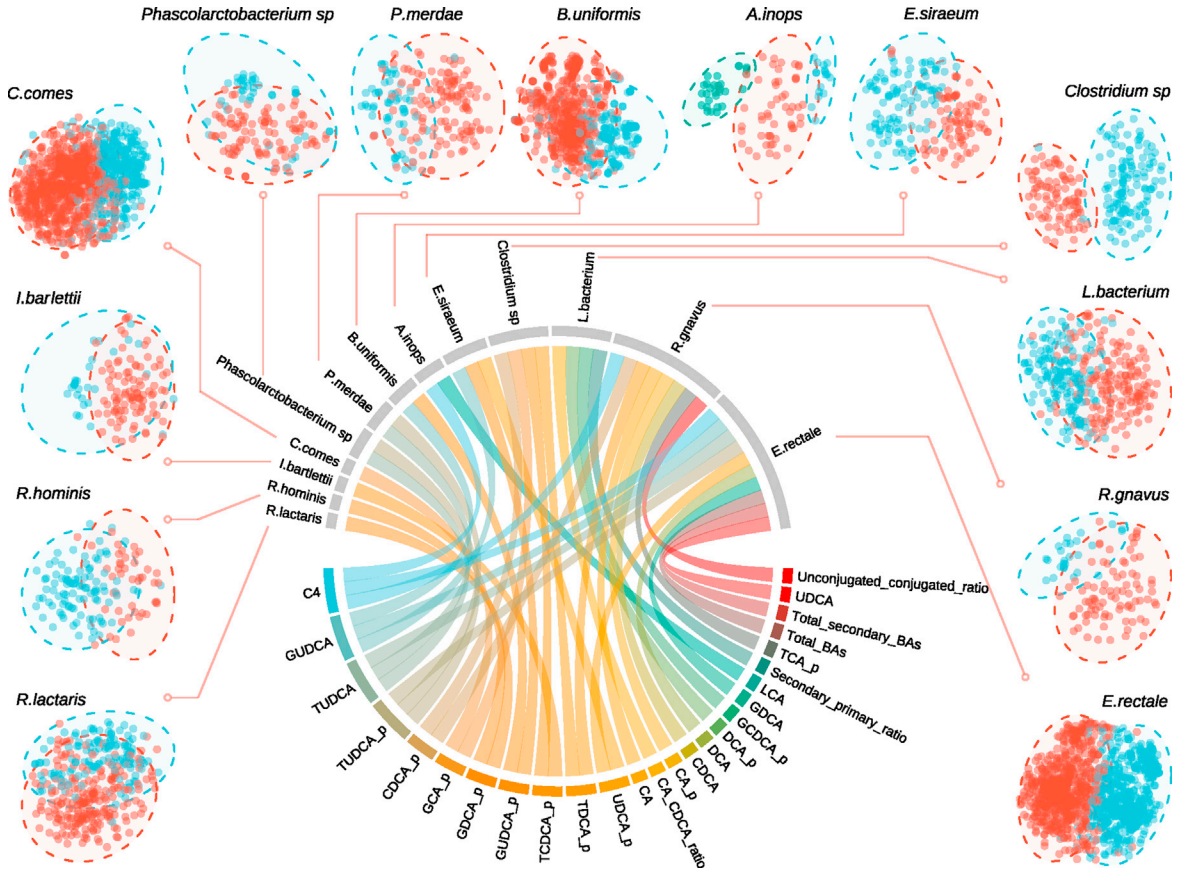


Figure 4. Bile acid parameters correlate with structural-variant-based populational genetic clusters

The populational genetic clusters of 13 species are shown by the t-SNE plots, with distinct clusters shown by different colors. The circos correlation plot shows their associations with BA. Each line indicates an association between clusters of a species and a BA parameter. See also Figures S4 and S5 and Table S4.

the 3D conformations of the proteins are closely related (Figure 6F), whereas EUBHAL_00727 corresponded to a truncated version of the 7αHSDH protein (Figure 6F). Furthermore, we could also identify the well-characterized 7αHSDH catalytic triad (Lou et al., 2016; Tanaka et al., 1996) in the CK1_08630/RUMOBE_03494 proteins (Figure 6G). These findings suggest that CK1_08630, RUMOBE_03494 and EUBHAL_00727 may have putative HSDH functionality.

In addition to the 809 associations that showed consistent effect sizes in both cohorts, we also identified 125 and 24 BA–SV associations with significant heterogeneity between our general population and obesity-based cohorts in *model 1*; *model 2*, respectively ($P_{hetero} < 0.05$, $FDR_{LLD} < 0.05$ and/or $FDR_{300\text{-}OB} < 0.05$; Table S5). In *model 1*, the most significant heterogeneity between the two cohorts was observed for the association between a 3-kbp vSV of *E. coli* (1,062–1,065 kbp) and TCDCA proportion (*model 1*, $Beta_{LLD} = -0.17$, $Beta_{300\text{-}OB} = 0.68$, $P_{hetero} = 1.79 \times 10^{-6}$, $I^2 = 0.96$; Table S5). This variable genomic region contains two genes, *Salmochelin siderophore protein IroE* and *Enterochelin esterase*, which play a role in maintaining iron homeostasis of *E. coli*. A 1-kbp vSV of *C. comes* (966–967 kbp) harboring a BSH gene was associated with three BA parameters (CA/CDCA ratio, secondary/primary BA ratio, and DCA proportion) with significant heterogeneity between LLD and 300-OB (*model 1*, $P_{hetero} < 0.05$; Table S5), and the absolute values of the effect sizes of the BA associations were higher in 300-OB than in LLD. In *model 2*, the most significant heterogeneity was observed for the association between a 1-kbp vSV of *E. hallii* (2,822–2,823 kbp) and plasma DCA concentration (*model 2*, $Beta_{LLD} = 0.007$, $Beta_{300\text{-}OB} = 0.45$, $P_{hetero} = 2.51 \times 10^{-6}$, $I^2 = 0.95$; Table S5).

Considering the physiological impact of BAs on the host's cardiometabolic phenotypes and stool consistency, we further assessed whether the BA-associated SVs can be associated with traits related to cardiometabolic disease risk and with stool characteristics (Table S6). In the LLD cohort, we detected significant SV associations with plasma triglycerides (TG) and stool type at an FDR < 0.05 level (Table S6). For instance, plasma TG was negatively associated with a 7-kbp vSV (1,125–1,130 kbp and 1,132–1,134 kbp) of *E. rectale*, and this SV region also negatively correlated with plasma C4 concentration and UDCA proportion (Table S5). Stool type, defined by Bristol stool scale, was associated with a 13-kbp vSV region (1,623–1,632 kbp and 2 segments) containing a BSH gene of *E. hallii*, and this vSV region was positively associated with plasma C4 level (Table S15). We also detected several suggestive associations at a nominal p < 0.05 level. For instance, a 14-kbp dSV (2,969–2,983 kbp) of *E. hallii* was associated with TG

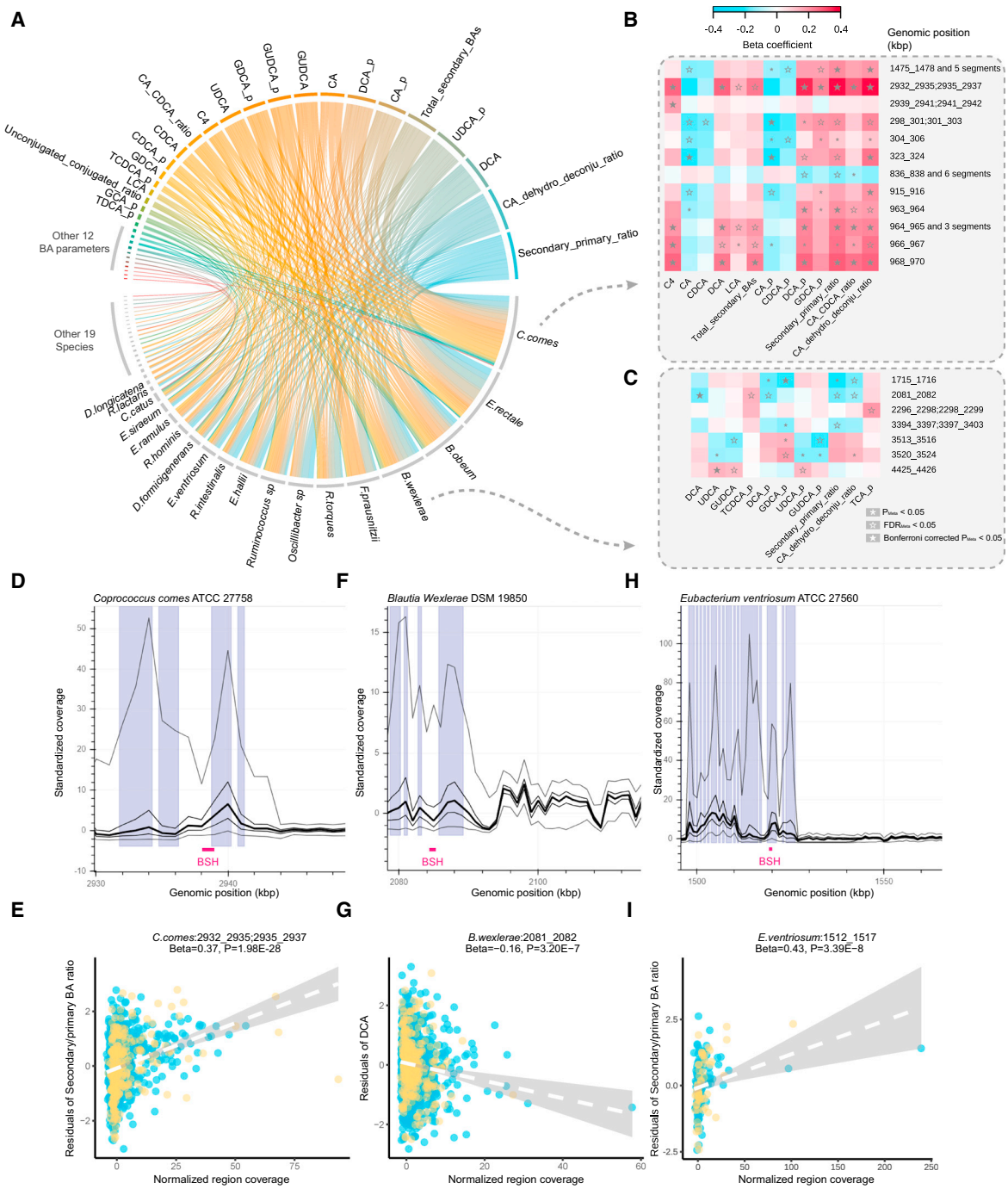


Figure 5. Associations between bile acid parameters and structural variants

(A) Replicable significant associations between BA parameters and SVs (FDR_{meta} < 0.05).

(B and C) Heatmap of associations between BA parameters and SVs of *C. comes* (B) and *B. wexlerae* (C). Associations identified by both *model 1*; *model 2* were selected and colored by beta coefficient from *model 2*.

(D–I) Examples of SV regions close to known BA biotransformation genes (D, F, and H) and associations with BA parameters (E, G, and I). Blue and yellow dots represent LLD and 300-OB samples, respectively. The Beta coefficients and p values of (E), (G), and (I) are from *model 2*, *model 2*, and *model 1,* respectively. See also Figure S6 and Tables S5 and S6.

level and high-density lipoprotein level in LLD and with dyslipidemia in 300-OB (Table S6). This region harbors a series of genes encoding outer membrane proteins and the lipopolysaccharide export system protein, and the strongest BA association of this dSV was found for the dehydroxylation-to-deconjugation ratio of CA (*model 2*, $P_{meta} = 1.32 \times 10^{-12}$; Table S5). In

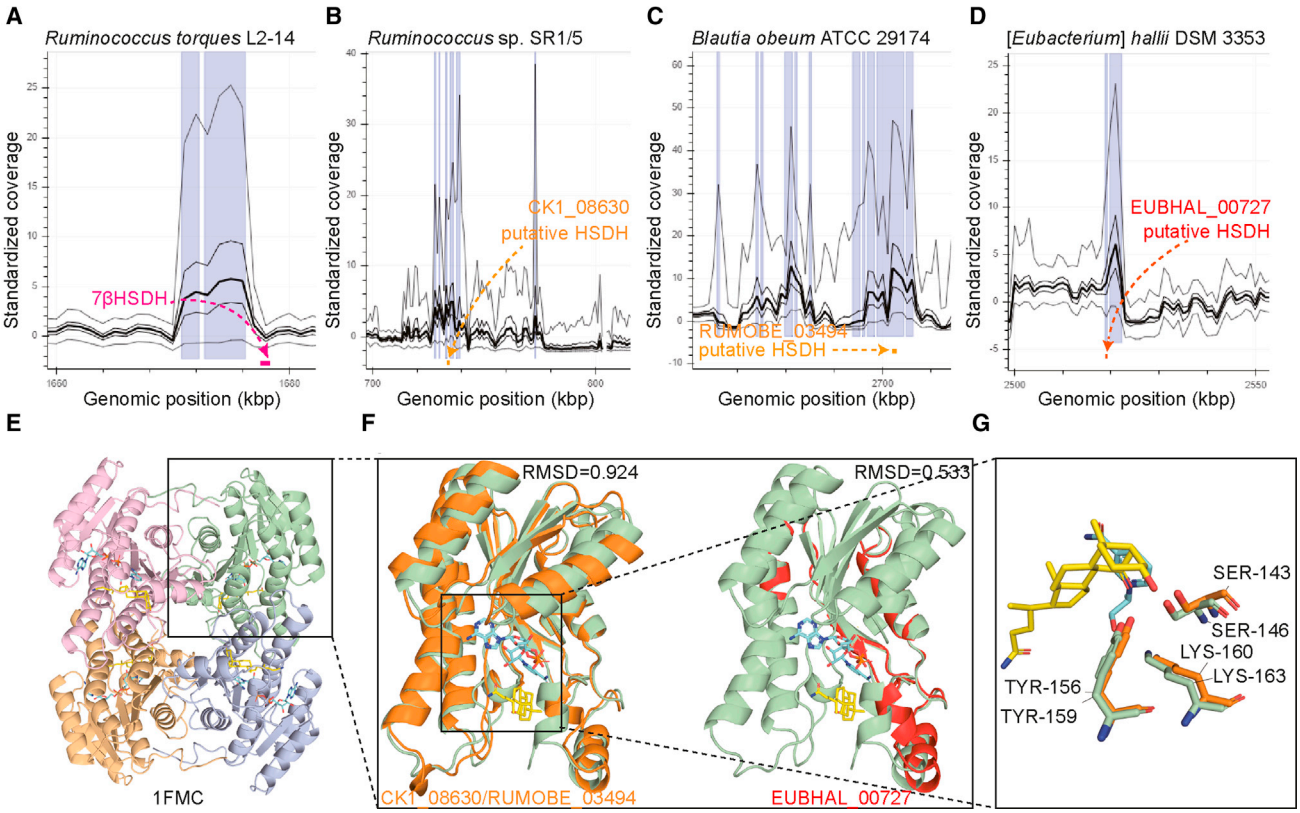


Figure 6. Examples of BA transformation genes and protein structure-based analysis

(A–D) Genomic position of 4 BA transformation genes and the closest SV regions.

(E) Homotetrameric structures of 7α-HSDH from *E. coli* (1FMC), which were the best hit from the structural similarity search of the CK1_08630/RUMOBE_03494 and EUBHAL 00727 models. 7-oxo GCDCA complexed with the proteins are displayed as yellow sticks and NADH as cyan sticks.

(F) Structural alignment and root mean square deviation (RMSD) values of the CK1_08630/RUMOBE_03494 (orange ribbons) and EUBHAL 00727 (red ribbons) models with the 7α-HSDH reference structures 1FMC (green ribbons).

(G) Comparison of the catalytic triad Ser146-Tyr159-Lys163 of 7α-HSDH (green sticks) in the reference 1FMC with CK1_08630/RUMOBE_03494 (Ser143-Tyr156-Lys160, orange sticks).

300-OB, no association was significant at an FDR < 0.05 level, but we found many nominally significant associations with blood lipids, diabetes, number and thickness of plaques, etc. (Table S6). Together, our data support a potential regulatory role for BA-associated microbial SVs in host cardiometabolic health.

Bidirectional causality between bacterial SVs and host BAs

The causality behind most of the BA-SV associations we identified remains unknown, although we did identify several bacterial genes known to be involved in BA biotransformation that were located in BA-associated SV regions. The lifestyle exposure factors collected in the LLD cohort enabled us to infer *in silico* causal relationships between correlated SVs and BAs. We integrated 127 lifestyle factors (78 dietary factors, 44 drug usage factors, and 5 smoking-related factors; Table S7) with SV and BA data. Here, we first identified lifestyle—SV-BA groups in which all the variables correlated with each other and then conducted bidirectional mediation analysis. In the first causal direction, we hypothesized that SVs act as regulators that mediate the effects of lifestyle factors on the composition of the BA pool, i.e., we treated SVs as mediators and BA parameters as outcomes (direction 1). In the second causal direction, we assessed whether BAs could mediate the effects of lifestyle factors on bacterial SVs (direction 2) (Figure 7A). In total, we identified 509 groups of inferred *in silico* causal relationships, including 217 unidirectional causal relationships in direction 1, 51 unidirectional causal relationships in direction 2, and 241 bidirectional causal relationships ($FDR_{mediation} < 0.05$; Figure 7B; Table S7). Most of the unidirectional causal effects were from SVs to BAs, which indicates the important role of microbial genetics in regulating human BA metabolism.

The tripartite causal network in direction 1 was composed of 43 lifestyle factors, 80 SVs as mediators, and 22 plasma BA parameters as outcomes ($FDR_{mediation} < 0.05$; Table S7). The 35 regulatory groups with a high mediated proportion (mediated proportion > 25%) are shown in Figure 7C. Notably, 29 of the 80 SVs were from *B. wexlerae*, including the SVs with known BA biotransformation genes. For instance, a 2-kbp vSV (2,081–2,082 kbp) close to a BSH gene in *B. wexlerae* regulated the effect of eating fish on plasma DCA concentration ($FDR_{mediation} < 0.05$; Mediated proportion = 29%; Figure 7D). Another two SVs of *B. wexlerae* in 3,840–3,846 kbp and 1,715–1,716 kbp mediated the effect of

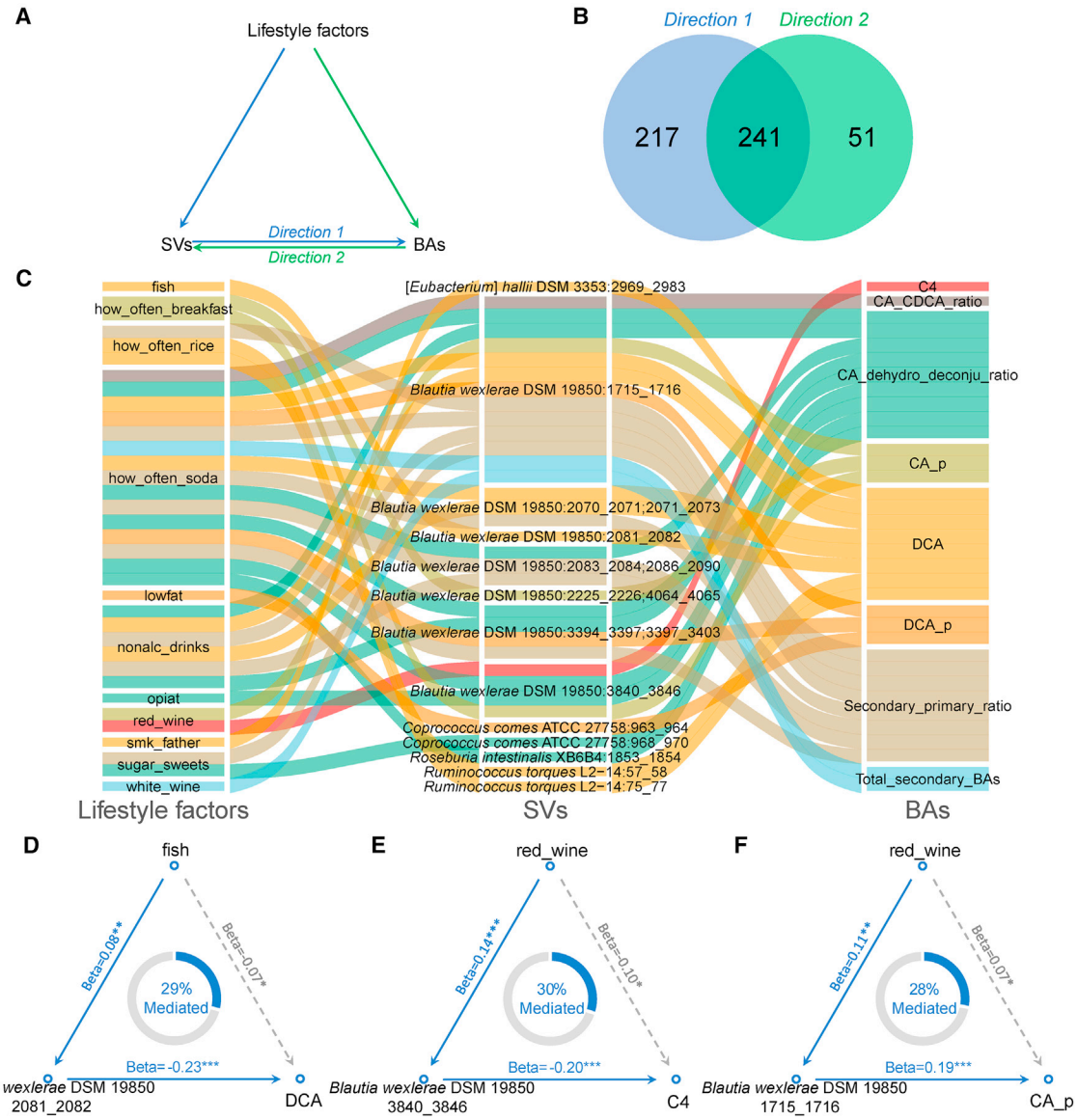


Figure 7. Causal relationship inference using bidirectional mediation analysis

(A) Framework of bidirectional mediation analysis between lifestyle factors, SVs and BAs.

(B) Number of inferred causal relationships for direction 1 (from SV to BA), direction 2 (from BA to SV) and both.

(C) Sankey diagram showing the inferred causal relationship network of direction 1 with mediated proportion >0.25.

(D–F) Examples of causal relationships between lifestyle factors, SVs and BAs inferred by bidirectional mediation analysis. The Beta coefficient and significance are labeled at each edge and the proportions of indirect effect (mediation effect) are labeled at the center of the ring charts. See also Table S7.

drinking red wine on C4 concentration ($FDR_{mediation}$ < 0.05; Mediated proportion = 30%; Figure 7E) and CA proportion in plasma ($FDR_{mediation}$ < 0.05; Mediated proportion = 28%; Figure 7F). Strikingly, we observed that the frequency of soda consumption was involved in 36 (16.6%) of the 217 unidirectional causal relationships in direction 1 and that 15 of the 36 causal relationships had a high mediated proportion ($FDR_{mediation}$ < 0.05; mediated proportion > 25%; Figure 7C), suggesting that soda consumption has a high impact on gut bacteria-related BA metabolism.

In direction 2, we found 51 *in silico* causal relationships in which 18 BA parameters mediated the effects of 20 lifestyle factors on 40 bacterial SVs belonging to 12 bacterial species ($FDR_{mediation}$ < 0.05; Table S7). Among the 40 regulated SVs, 16 were from *B. wexlerae*, followed by 8 from *C. comes*, 3 from *R. torques*, and 2 from *Ruminococcus* sp. SR1/5. The 5 SVs from *Ruminococcus* species can be negatively regulated by C4, UDCA, GUDCA, and total secondary BA levels ($FDR_{mediation}$ < 0.05; Table S7). In mice, the growth of *Ruminococcus* species can be inhibited by DCA and LCA (Tian et al., 2020), whereas UDCA was reported to have an antibacterial effect in an animal experiment (Kim et al., 2020), indicating that enrichment of the circulating BA pool with secondary BAs may exert selective pressure

on *Ruminococcus* species and cause an alteration of their genomic content.

DISCUSSION

We characterized the gut microbial SV and plasma BA profiles of 1,437 Dutch individuals from two independent cohorts and systemically assessed the correlation between gut microbial genetics and host BA metabolism from species genetic makeup level down to single-variant level. The species genetic makeup was found to correlate with BA parameters independent of the relative abundances of these species. We also identified populational genetic clusters of 29 bacterial species using SV-based clustering analysis, revealed the within-species genetic diversity and associated the SV-based genetic clusters with plasma BA parameters. We further performed a metagenome-wide microbial SV association study on 39 BA parameters and identified a total of 809 consistent and 125 heterogeneous associations using meta-analysis with two different models. Some of the BA-associated SVs can also be associated with plasma triglycerides and stool type. Bidirectional mediation analysis between bacterial SVs, BAs, and dietary factors inferred *in silico* regulatory relationships behind the correlations we identified. To the best of our knowledge, this is the largest study so far on the microbial genetic determinants of plasma BA concentrations and composition in humans. In view of the growing awareness of the involvement of specific BAs in the onset and progression of human diseases (Chávez-Talavera et al., 2017; Dermadi et al., 2017; Gao et al., 2019), as well as the current development of pharmacological agents that target BA-signaling pathways for treatment of liver and metabolic diseases (Jia et al., 2018; Krautkramer et al., 2021; Pathak et al., 2018; Sun et al., 2018), this knowledge is of direct clinical relevance.

Our study demonstrates that SV-based metagenome-wide association is a powerful method to bring microbial associations closer to functionality and mechanistic understanding. First, our study shows that the BA associations with microbial SVs are often stronger than those with species relative abundances and can even be independent of species relative abundances. This highlights the value of metagenomic SVs as an extra source of information that describes the functionality of the human gut microbiome. We also assessed the impact of the lineage effect on the SV-based metagenome-wide association study by adding the PCs of population genetic structure into our linear model. On the one hand, the model considering the lineage effect enabled the discovery of lineage-independent associations and was also able to reveal extra associations that were seemingly hidden by the lineage effect. On the other hand, we also observed that the model considering lineage effect caused a loss of power for some associations involving SVs that contribute to both target phenotype and bacterial genetic lineage. Although there are many tools available that consider lineage effects and populational structure in bacterial GWAS analysis (Collins and Didelot, 2018; Earle et al., 2016), these tools were developed for binary genetic variation data and not for the metagenomic scenario. The SV-based metagenome-wide association study still calls for the development of tools that support appropriate adjustment for lineage effects.

Additionally, we observed that the associated SVs were biased toward highly prevalent and abundant species, highlighting the statistical challenges in studying low prevalence or rare species. Increased sample sizes and deep sequencing are thus required. Nevertheless, our sub-genome-scale analysis pinpointed the location of genomic segments that are associated with the host BA pool, which means that associating microbial SVs from across the whole metagenome with host phenotypes helps to locate microbial genes or genetic elements involved in host-microbe interaction. Our study underscores the contribution of gut microbial genetics to the individuality of host BA metabolism, and the comprehensive association analysis approach we used provides a template for cohort-based microbial genetics studies, demonstrating a paradigm shift from ''micro-ecology'' to ''micro-population genetics.''

Our study further highlights the complex, bidirectional effect between the gut microbiome and BA metabolism. We used lifestyle factors as exogenous predictors to infer *in silico* potential causal relationships between SVs and BAs using bidirectional mediation analysis, and this identified specific lifestyle factors that are involved in the interaction between bacterial genetics and BA metabolism. This highlights the potential of targeting the gut microbiota to regulate BA metabolism through lifestyle intervention. For instance, we found that an SV of *B. wexlerae* mediated the effect of red wine drinking on plasma CA proportion and C4 level, reflecting the hepatic BA biosynthesis. Red wine is rich in polyphenols, a group of molecules with antioxidative properties (Naumann et al., 2020; Queipo-Ortuño et al., 2012) that can increase fecal BA excretion by regulating gut microbiota (Chambers et al., 2019). Previous studies reported that dietary polyphenols from plant-derived foods can affect the composition of fecal BAs in humans by regulating gut microbiota (Chambers et al., 2019; Ozdal et al., 2016; Queipo-Ortuño et al., 2012; Sembries et al., 2006); our result suggests that polyphenols from red wine may impact biosynthesis of BA by regulating gut bacterial genes. Additionally, we also observed that the frequency of soda consumption had a striking impact on BA metabolism through bacterial SVs, especially those from *B. wexlerae*. The consumption of soda or soft drinks has been reported to correlate with all-cause mortality (Mullee et al., 2019) and an increased risk of many diseases, including stroke and coronary artery disease (Mossavar-Rahmani et al., 2019). In LLD, our previous study also showed the negative impact of soda consumption on microbiome diversity (Zhernakova et al., 2016). Considering the high popularity of soda drinks and the importance of a balanced BA pool and gut microecosystem, our current study should raise public awareness of the impact of soda consumption on the gut microbiome and BA metabolism. Altogether, our *in silico* causal inference analysis revealed that bacterial SVs serve as mediators that regulate the effects of dietary factors on BA metabolism. Conversely, our study also provides evidence that BAs, likely via their antibacterial activities as ''intestinal soaps,'' not only affect the growth of intestinal microbes but also pose selective pressure on bacterial genetics.

Experimental validation of microbial SV associations with BAs remains challenging. Bioinformatically, we lack a good approach for prioritizing putative causal genes at the SV

regions, and we propose using the recently developed AlphaFold2 approach to predict 3D protein structures in order to discover putative bacterial proteins that bind with BAs (Jumper et al., 2021). Experimentally, we first need to isolate and culture bacterial species or strains from human fecal samples and then conduct whole-genome sequencing to confirm the presence/absence of SVs at the single-bacterial-isolate level. Verified bacterial isolates with SVs of interest can then be tested for their capacities in BA metabolism *in vitro* by co-culturing them with BAs. Finally, the identified isolates can be colonized in animal models, such as our recently developed Cyp2c70$^{-/-}$ mice with a human-like BA pool (Boer et al., 2020), to verify their effects on host BA metabolism and physiology. However, each of these steps faces technical challenges. For example, although the technology of bacterial culturomics is developing rapidly, gut microbes remain difficult to isolate and culture, particularly when specific selective media are needed to enrich certain types of strains, as is the case for bacteria with a specific SV region.

Limitations of the study

We acknowledge several limitations of our current study. We investigated the association between plasma BA parameters and variable genomic segments of gut bacteria in two independent cohorts, identified substantial consistent associations in both these general population and obese individuals and demonstrated the reliability of BA associations with microbial SVs. However, all the samples included in this study were collected from residents of the Netherlands. Considering the potential heterogeneity of host—microbiome interaction across populations with different genetic and environmental backgrounds, the associations between plasma BA parameters and microbial SVs need to be replicated in other populations with different backgrounds. As this is a cross-sectional study, we inferred the regulatory relationships between BA parameters and microbial SVs using mediation analysis, but whether the shifts of microbial genetic elements causally correlate with host BA metabolism still requires further confirmation in a longitudinal study design and through experimental validation, as we elaborated in the Discussion. Additionally, as plasma BA parameters cannot fully represent the flux of the BA pool in enterohepatic circulation and are only modestly correlated with the fecal BA pool (Chen et al., 2020); thus, further study of the association between microbial genetic variation and BA metabolism in their actual niche—the enterohepatic circulation—is needed. Furthermore, the microbiome in feces does not directly reflect the microbial composition of other colonic segments, in particular the microbiome of the ileum, where most BAs (~95%) are reabsorbed, and that of the colonic regions, where most BA biotransformation occurs. Therefore, investigation of the links between BA pool composition and the microbiome from the various intestinal compartments is of interest, and further efforts should be made to fully elucidate bacteria-BA interactions. Despite these limitations, our study represents a step toward microbiome-targeted interventions to improve host metabolism, in particular through modulation of BA metabolism, which is a major target for the treatment of non-alcoholic fatty liver disease and its metabolic co-morbidities.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - LifeLines-DEEP cohort
  - 300-Obesity cohort
  - Ethical approval
- METHOD DETAILS
  - Bile acid quantification
  - Metagenomic sequencing and quality control
  - Taxonomic abundance
  - Detection of structural variations
  - Functional annotation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Association analysis
  - Mediation analysis
  - Distance calculation
  - PCoA and PERMANOVA
  - Clustering analysis
  - Protein 3D structure prediction and protein structure-based analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2021.11.003.

ACKNOWLEDGMENTS

We thank all the volunteers in the LifeLines-DEEP cohort and 300-Obesity cohort of the Human Functional Genomics Project (HFGP) for their participation and the project staff for their help and management. This study was supported by an IN-CONTROL CVON grant (CVON2012-03, CVON2018-27) to N.P.R., L.A.B.J., M.G.N., A.Z., F.K., and J.F. J.F. is supported by a European Research Council (ERC) Consolidator grant (101001678) and an NWO VICI grant (VI.C.202.022). F.K. is supported by the Noaber Foundation, Lunteren, the Netherlands. A.Z. is supported by an ERC Starting grant 715772, NWO-VIDI grant 016.178.056, the NWO Gravitation grant exposome-NL (024.004.017), and ZonMw Memorabel grant (733050814). C.W. is supported by the NWO Spinoza Prize SPI 92-266, a FP7/2007-2013/ERC advanced grant 2012-322698 and the University of Groningen Investment Agenda Grant Personalized Health. Moreover, J.F. and C.W. are supported by the Netherlands Organ-on-Chip Initiative, an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the Government of the Netherlands. D.W. is supported by China Scholarship Council (CSC201904910478). L.C. holds a joint fellowship from the University Medical Center Groningen and China Scholarship Council (CSC201708320268) and a Foundation de Cock-Hadders grant (20:20-13). M.D. holds a MD-PhD fellowship from the University Medical Center Groningen. We also thank the Genomics Coordination Center for providing data infrastructure and access to high-performance computing clusters and Kate Mc Intyre for critical reading and editing.

AUTHOR CONTRIBUTIONS

J.F., F.K., A.Z., and C.W. conceptualized and managed the study. D.W., M.D., L.C., I.C.L.v.d.M., M.K., N.P.R., J.H.W.R., and L.A.B.J. contributed to sample

collection and data generation. D.W. analyzed the data. A.J.R.M. performed protein structure-based analysis. D.W., J.F., and F.K. drafted the manuscript. D.W., M.D., L.C., S.A.S., I.C.L.v.d.M., H.E.A., M.K., A.J.R.M, V.W.B., N.P.R., J.H.W.R., M.G.N., C.W., A.Z., J.F., and F.K. reviewed and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 7, 2021
Revised: September 6, 2021
Accepted: November 5, 2021
Published: November 29, 2021

REFERENCES

Boer, J.F. de, Verkade, E., Mulder, N.L., de Vries, H.D., Huijkman, N., Koehorst, M., Boer, T., Wolters, J.C., Bloks, V.W., van de Sluis, B., and Kuipers, F. (2020). A human-like bile acid pool induced by deletion of hepatic Cyp2c70 modulates effects of FXR activation in mice. J. Lipid Res. *61*, 291–305.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Chambers, K.F., Day, P.E., Aboufarrag, H.T., and Kroon, P.A. (2019). Polyphenol effects on cholesterol metabolism via bile acid biosynthesis, CYP7A1: a review. Nutrients *11*, 2588.

Chávez-Talavera, O., Tailleux, A., Lefebvre, P., and Staels, B. (2017). Bile acid control of metabolism and inflammation in obesity, Type 2 diabetes, dyslipidemia, and nonalcoholic fatty liver disease. Gastroenterology *152*, 1679–1694.e3.

Chen, L., Munckhof, I.C.L. van den, Schraa, K., Horst, R. ter, Koehorst, M., Faassen, M. van, Ley, C. van der, Doestzada, M., Zhernakova, D.V., Kurilshikov, A., et al. (2020). Genetic and microbial associations to plasma and fecal bile acids in obesity relate to plasma lipids and liver fat content. Cell Rep. *33*, 108212.

Chiang, J.Y. (2017). Recent advances in understanding bile acid homeostasis. F1000Res. *6*, 2029.

Collins, C., and Didelot, X. (2018). A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput. Biol. *14*, e1005958.

Đanić, M., Stanimirov, B., Pavlović, N., Goločorbin-Kon, S., Al-Salami, H., Stankov, K., and Mikov, M. (2018). Pharmacological applications of bile acids and their derivatives in the treatment of metabolic syndrome. Front. Pharmacol. *9*, 1382.

Dermadi, D., Valo, S., Ollila, S., Soliymani, R., Sipari, N., Pussila, M., Sarantaus, L., Linden, J., Baumann, M., and Nyström, M. (2017). Western diet deregulates bile acid homeostasis, cell proliferation, and tumorigenesis in colon. Cancer Res. *77*, 3352–3363.

Earle, S.G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N.C., Walker, T.M., Spencer, C.C.A., Iqbal, Z., Clifton, D.A., Hopkins, K.L., et al. (2016). Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat. Microbiol. *1*, 16041.

Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., Wang, L.-P., Simmonett, A.C., Harrigan, M.P., Stern, C.D., et al. (2017). OpenMM 7: rapid development of high performance algorithms for molecular dynamics. PLoS Comput. Biol. *13*, e1005659.

Eggink, H.M., Oosterman, J.E., Goede, P., de, Vries, E.M. de, Foppen, E., Koehorst, M., Groen, A.K., Boelen, A., Romijn, J.A., la Fleur, S.E., et al. (2017). Complex interaction between circadian rhythm and diet on bile acid homeostasis in male rats. Chronobiol. Int. *34*, 1339–1353.

Gao, L., Lv, G., Li, R., Liu, W.-T., Zong, C., Ye, F., Li, X.-Y., Yang, X., Jiang, J.-H., Hou, X.-J., et al. (2019). Glycochenodeoxycholate promotes hepatocellular carcinoma invasion and migration by AMPK/mTOR dependent autophagy activation. Cancer Lett. *454*, 215–223.

Gu, Y., Wang, X., Li, J., Zhang, Y., Zhong, H., Liu, R., Zhang, D., Feng, Q., Xie, X., Hong, J., et al. (2017). Analyses of gut microbiota and plasma bile acids enable stratification of patients for antidiabetic treatment. Nat. Commun. *8*, 1785.

Heinken, A., Ravcheev, D.A., Baldini, F., Heirendt, L., Fleming, R.M.T., and Thiele, I. (2019). Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. Microbiome *7*, 75.

Heintz-Buschart, A., and Wilmes, P. (2018). Human gut microbiome: function matters. Trends Microbiol. *26*, 563–574.

Hoogerland, J.A., Lei, Y., Wolters, J.C., de Boer, J.F., Bos, T., Bleeker, A., Mulder, N.L., van Dijk, T.H., Kuivenhoven, J.A., Rajas, F., et al. (2019). Glucose-6-phosphate regulates hepatic bile acid synthesis in mice. Hepatology *70*, 2171–2184.

Horst, R. ter, Munckhof, I.C.L. van den, Schraa, K., Aguirre-Gamboa, R., Jaeger, M., Smeekens, S.P., Brand, T., Lemmers, H., Dijkstra, H., Galesloot, T.E., et al. (2020). Sex-specific regulation of inflammation and metabolic syndrome in obesity. Arterioscler. Thromb. Vasc. Biol. *40*, 1787–1800.

Jia, W., Xie, G., and Jia, W. (2018). Bile acid–microbiota crosstalk in gastrointestinal inflammation and carcinogenesis. Nat Rev Gastroenterol. Hepatol. *15*, 111–128.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

Kim, B.-T., Kim, K.-M., and Kim, K.-N. (2020). The effect of ursodeoxycholic acid on small intestinal bacterial overgrowth in patients with functional dyspepsia: a pilot randomized controlled trial. Nutrients *12*, 1410.

Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nat. Commun. *10*, 5416.

Krautkramer, K.A., Fan, J., and Bäckhed, F. (2021). Gut microbial metabolites as multi-kingdom intermediates. Nat. Rev. Microbiol. *19*, 77–94.

Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr. *60*, 2256–2268.

Krissinel, E., and Henrick, K. (2005). Multiple alignment of protein structures in three dimensions. In Computational Life Sciences, First International Symposium, CompLife 2005, pp. 67–78, Konstanz, Germany, September 25–27, 2005. Proceedings.

Kuipers, F., Bloks, V.W., and Groen, A.K. (2014). Beyond intestinal soap—bile acids in metabolic control. Nat. Rev. Endocrinol. *10*, 488–498.

Kurilshikov, A., van den Munckhof, I.C.L., Chen, L., Bonder, M.J., Schraa, K., Rutten, J.H.W., Riksen, N.P., de Graaf, J., Oosting, M., Sanna, S., et al. (2019). Gut microbial associations to plasma metabolites linked to cardiovascular phenotypes and risk. Circ. Res. *124*, 1808–1820.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lou, D., Wang, B., Tan, J., Zhu, L., Cen, X., Ji, Q., and Wang, Y. (2016). The three-dimensional structure of Clostridium absonum 7α-hydroxysteroid dehydrogenase: new insights into the conserved arginines for NADP(H) recognition. Sci. Rep. *6*, 22885.

Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L. (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. *3*, e104.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics *29*, 2722–2728.

Mars, R.A.T., Yang, Y., Ward, T., Houtti, M., Priya, S., Lekatz, H.R., Tang, X., Sun, Z., Kalari, K.R., Korem, T., et al. (2020). Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. Cell *182*, 1460–1473.e17.

Mende, D.R., Letunic, I., Huerta-Cepas, J., Li, S.S., Forslund, K., Sunagawa, S., and Bork, P. (2017). proGenomes: a resource for consistent functional

and taxonomic annotations of prokaryotic genomes. Nucleic Acids Res. *45*, D529–D534.

Mirdita, M., Ovchinnikov, S., and Steinegger, M. (2021). ColabFold - Making protein folding accessible to all. BioRxiv. https://doi.org/10.1101/2021.08.15.456425.

Mossavar-Rahmani, Y., Kamensky, V., Manson, J.E., Silver, B., Rapp, S.R., Haring, B., Beresford, S.A.A., Snetselaar, L., and Wassertheil-Smoller, S. (2019). Artificially sweetened beverages and stroke, coronary heart disease, and all-cause mortality in the Women's Health Initiative. Stroke *50*, 555–562.

Mullee, A., Romaguera, D., Pearson-Stuttard, J., Viallon, V., Stepien, M., Freisling, H., Fagherazzi, G., Mancini, F.R., Boutron-Ruault, M.-C., Kühn, T., et al. (2019). Association Between soft drink consumption and mortality in 10 European countries. JAMA Intern. Med. *179*, 1479–1490.

Naumann, S., Haller, D., Eisner, P., and Schweiggert-Weisz, U. (2020). Mechanisms of interactions between bile acids and plant compounds-a review. Int. J. Mol. Sci. *21*, 6495.

Ozdal, T., Sela, D.A., Xiao, J., Boyacioglu, D., Chen, F., and Capanoglu, E. (2016). The reciprocal interactions between polyphenols and gut microbiota and effects on bioaccessibility. Nutrients *8*, 78.

Pathak, P., Xie, C., Nichols, R.G., Ferrell, J.M., Boehme, S., Krausz, K.W., Patterson, A.D., Gonzalez, F.J., and Chiang, J.Y.L. (2018). Intestine farnesoid X receptor agonist and the gut microbiota activate G-protein bile acid receptor-1 signaling to improve metabolism. Hepatology *68*, 1574–1588.

Queipo-Ortuño, M.I., Boto-Ordóñez, M., Murri, M., Gomez-Zumaquero, J.M., Clemente-Postigo, M., Estruch, R., Cardona Diaz, F.C., Andrés-Lacueva, C., and Tinahones, F.J. (2012). Influence of red wine polyphenols and ethanol on the gut microbiota ecology and biochemical biomarkers. Am. J. Clin. Nutr. *95*, 1323–1334.

Sato, Y., Atarashi, K., Plichta, D.R., Arai, Y., Sasajima, S., Kearney, S.M., Suda, W., Takeshita, K., Sasaki, T., Okamoto, S., et al. (2021). Novel bile acid biosynthetic pathways are enriched in the microbiome of centenarians. Nature. https://doi.org/10.1038/s41586-021-03832-5.

Scholtens, S., Smidt, N., Swertz, M.A., Bakker, S.J., Dotinga, A., Vonk, J.M., Dijk, F. van, Zon, S.K. van, Wijmenga, C., Wolffenbuttel, B.H., and Stolk, R.P. (2015). Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int. J. Epidemiol. *44*, 1172–1180.

Sembries, S., Dongowski, G., Mehrländer, K., Will, F., and Dietrich, H. (2006). Physiological effects of extraction juices from apple, grape, and red beet pomaces in rats. J. Agric. Food Chem. *54*, 10269–10280.

Song, Z., Cai, Y., Lao, X., Wang, X., Lin, X., Cui, Y., Kalavagunta, P.K., Liao, J., Jin, L., Shang, J., and Li, J. (2019). Taxonomic profiling and populational patterns of bacterial bile salt hydrolase (BSH) genes based on worldwide human gut microbiome. Microbiome *7*, 9.

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028.

Steiner, C., Othman, A., Saely, C.H., Rein, P., Drexel, H., Eckardstein, A. von, and Rentsch, K.M. (2011). Bile acid metabolites in serum: intraindividual variation and associations with coronary heart disease, metabolic syndrome and diabetes mellitus. PLoS One *6*, e25006.

Sun, L., Xie, C., Wang, G., Wu, Y., Wu, Q., Wang, X., Liu, J., Deng, Y., Xia, J., Chen, B., et al. (2018). Gut microbiota and intestinal FXR mediate the clinical benefits of metformin. Nat. Med. *24*, 1919–1929.

Tanaka, N., Nonaka, T., Tanabe, T., Yoshimoto, T., Tsuru, D., and Mitsui, Y. (1996). Crystal structures of the binary and ternary complexes of 7α-Hydroxysteroid dehydrogenase from Escherichia coli. Biochemistry *35*, 7715–7730.

Tian, Y., Gui, W., Koo, I., Smith, P.B., Allman, E.L., Nichols, R.G., Rimal, B., Cai, J., Liu, Q., and Patterson, A.D. (2020). The microbiome modulating activity of bile acids. Gut Microbes *11*, 979–996.

Tibshirani, R., and Walther, G. (2012). Cluster validation by prediction strength. J. Comput. Graph. Stat. *14*, 511–528.

Tigchelaar, E.F., Zhernakova, A., Dekens, J.A.M., Hermes, G., Baranska, A., Mujagic, Z., Swertz, M.A., Muñoz, A.M., Deelen, P., Cenit, M.C., et al. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open *5*, e006772.

Udayappan, S., Manneras-Holm, L., Chaplin-Scott, A., Belzer, C., Herrema, H., Dallinga-Thie, G.M., Duncan, S.H., Stroes, E.S.G., Groen, A.K., Flint, H.J., et al. (2016). Oral treatment with Eubacterium hallii improves insulin sensitivity in db/db mice. NPJ Biofilms Microbiomes *2*, 16009.

Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E.M., Disz, T., Gabbard, J.L., et al. (2017). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res *45*, D535–D542.

Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257.

Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., Weinberger, A., Fu, J., Wijmenga, C., Zhernakova, A., and Segal, E. (2019). Structural variation in the gut microbiome associates with host health. Nature *568*, 43–48.

Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science *352*, 565–569.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Fecal and blood samples of Lifelines-DEEP	Tigchelaar et al., 2015	PMID: 26319774
Fecal samples of 300-Obesity	Kurilshikov et al., 2019	PMID: 30971183
Blood samples of 300-Obesity	Horst et al., 2020	PMID: 32460579
Critical commercial assays		
AllPrep DNA/RNA Mini Kit	QIAGEN	80204
Quant-iT PicoGreen dsDNA Assay	Life Technologies	P7589
Chemicals		
Cholic acid	Sigma-Aldrich	SKU- C1129
Taurocholic acid	Sigma-Aldrich	SKU- T4009
Glycocholic acid	Sigma-Aldrich	SKU- G2878
Deoxycholic acid	Sigma-Aldrich	SKU- D2510
Taurodeoxycholic acid	Sigma-Aldrich	SKU- T0557
Glycodeoxycholic acid	Sigma-Aldrich	SKU- G9910
Chenodeoxycholic acid	Sigma-Aldrich	SKU- C9377
Taurochenodeoxycholic acid	Sigma-Aldrich	SKU- T6260
Glycochenodeoxycholic acid	Sigma-Aldrich	SKU- G0759
Ursodeoxycholic acid	Sigma-Aldrich	SKU- U5127
Tauroursodeoxycholic acid	Merck	CAT No. 580549
Glycoursodeoxycholic acid	Sigma-Aldrich	SKU- G0759
Lithocholic acid	Sigma-Aldrich	SKU- L6250
Taurolithocholic acid	Sigma-Aldrich	SKU- T7515
Glycolithocholic acid	IsoSciences	Cat No. 13231UNL
D4-cholic acid	CDN Isotopes	Prod No. D-2452
D4-chenodeoxycholic acid	CDN Isotopes	Prod No. D-2772
D4-glycochenodeoxycholic acid	CDN Isotopes	Prod No. D-5673
D4-glycocholic acid	CDN Isotopes	Prod No. D-3878
D4-taurochenodeoxycholic acid	Medical Isotopes	Cat No. D2122
D4-taurocholic acid	Medical Isotopes	Cat No. D3770
D4-tauroursodeoxycholic acid	IsoSciences	Prod No. 13106
D6-taurodeoxycholic acid	IsoSciences	Prod No. 13228
Deposited data		
Metagenomic sequencing data of LifeLines-DEEP	This study	European Genomics-Phenome Archive, EGAS00001001704
Metagenomic sequencing data of 300-Obesity	This study	European Genomics-Phenome Archive, EGAS00001003508
Software and algorithms		
Bowtie2 (version 2.3.4.3)	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Trimmomatic (version 0.39)	Bolger et al., 2014	http://www.usadellab.org/cms/?page=trimmomatic
KneadData (version 0.7.4)	Huttenhower lab	https://huttenhower.sph.harvard.edu/kneaddata/
Kraken2 (version 2.1.2)	Wood et al., 2019	https://ccb.jhu.edu/software/kraken2/
Bracken (version 2.6.2)	Lu et al., 2017	https://ccb.jhu.edu/software/bracken/
ICRA	Zeevi et al., 2019	https://github.com/segalab/SGVFinder

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SGVFinder (version 1.0)	Zeevi et al., 2019	https://github.com/segalab/SGVFinder
R (version 4.0.1)	R Core Team	https://www.r-project.org/
Python (version 2.7.16)	Python Core Team	https://www.python.org/
PATRIC (version 3.6.6)	Wattam et al., 2017	https://www.patricbrc.org/
AlphaFold (version 2.0.0)	Jumper et al., 2021	https://github.com/deepmind/alphafold
ColabFold (version 1.0-alpha)	Mirdita et al., 2021	https://github.com/sokrypton/ColabFold
MMseqs2 (version 13-45111)	Steinegger and Söding, 2017	https://github.com/soedinglab/MMseqs2
OpenMM 7 (version 7.5.1)	Eastman et al., 2017	https://openmm.org/
PDBeFold (version 2.59)	Krissinel and Henrick, 2004, 2005	https://www.ebi.ac.uk/msd-srv/ssm/
Analyst® MD (version 1.6.2)	SCIEX	https://sciex.com/products/in-vitro-diagnostics/enhancements-and-options/analyst-md-software
R code for statistical analysis and visualization (version 1.0.0)	This study	https://doi.org/10.5281/zenodo.5599104
Other		
Progenome (version 1.0)	(Mende et al., 2017)	http://progenomes1.embl.de/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources, software, reagents and data sharing should be directed to the Lead Contact, Jingyuan Fu (j.fu@umcg.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Raw metagenomic sequencing data of LifeLines-DEEP and 300-Obesity are publicly available from the European Genome-Phenome Archive via accession number EGAS00001001704 and EGAS00001003508, respectively.
- The code used for statistical analysis is available via https://doi.org/10.5281/zenodo.5599104.
- Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

LifeLines-DEEP cohort

LifeLines-DEEP (LLD) (Tigchelaar et al., 2015) is a sub-cohort of LifeLines (Scholtens et al., 2015), a large population-based prospective cohort that enrolled 167,729 participants from the north of Netherlands, established to explore the risk factors of complex diseases. In LLD, 1,539 individuals were included and multi-layers of omics data were collected. In the current study, high-quality metagenomic sequencing data, 78 dietary factors, 5 smoking factors and 44 drug usage factors were available for 1,135 individuals (474 males and 661 females). The average age of LLD participants was 45.04 years old (18–81, SE = 0.40) and the average BMI was 25.26 (16.67–48.56, SE = 0.12).

300-Obesity cohort

The 300-Obesity (300-OB) cohort was established by Radboud University Medical Center, Nijmegen, the Netherlands (Horst et al., 2020). In total, 302 individuals (167 males and 135 females) aged 55–81 years with a high BMI > 27 were enrolled in 300-OB. The average age of 300-OB participants was 67.1 years old (54–81, SE = 0.31) and their average BMI was 30.7 (26.3–45.5, SE = 0.20). All participants were included between 2014 and 2016.

For both cohorts, the inclusion of samples considered gender balance. The influence of gender has been assessed and adjusted for relevant analysis.

Ethical approval

The LifeLines-DEEP study has been approved by the Institutional ethics Review Board (IRB) of the University Medical Center Groningen (ref. M12.113965), the Netherlands. The 300-Obesity study has been approved by the IRB CMO Regio Arnhem-Nijmegen (nr. 46846.091.13).

METHOD DETAILS

Bile acid quantification

Levels of 15 BAs and C4 concentrations in fasting plasma were quantified by liquid chromatography–mass spectrometry procedures, as previously described (Eggink et al., 2017; Hoogerland et al., 2019). In brief, for sample preparation, 25 μL plasma was mixed with 250 μL internal standard solution to precipitate proteins. Samples were centrifuged at 15800 × g and the supernatant poured into a clean glass tube. The fluid was evaporated under nitrogen at 40 °C. Before measuring samples were reconstituted in 200 μL 50% methanol in water. For the quantitative determination of BAs, we used a Nexera X2 Ultra High-Performance Liquid Chromatography system (SHIMADZU, Kyoto, Japan), coupled to a SCIEX QTRAP 4500 MD triple quadrupole mass spectrometer (SCIEX, Framingham, MA, USA) (UHPLC-MS/MS). The LC-MS/MS system is controlled by Analyst® MD 1.6.2 software. BAs were separated with a ACQUITY UPLC BEH C18 Column (1.7 μm, 2.1 mm × 100 mm) equipped with a ACQUITY UPLC BEH C18 VanGuard Pre-Column (1.7 μm, 2.1 mm × 5 mm), (Waters, Milford, MA, USA). Separation was achieved in 28 minutes using 10 mM ammonium acetate in 20% acetonitrile (mobile phase A) and 10 mM ammonium acetate in 80% acetonitrile (mobile phase B), flow 0.4 ml/min. Chemicals used for bile acids profiling are listed in the key resources table.

The proportions of 15 BAs (with suffix '_p') were calculated by dividing by total BA concentration. Additionally, different total BA concentrations and ratios were calculated (Chen et al., 2020): (1) Total BA (Total_BAs) = sum of all BA concentrations, (2) total primary BA (Total_primary_BAs) = sum of all primary BA concentrations, (3) total secondary BA (Total_secondary_BAs) = sum up of all secondary BA concentrations, (4) ratio of Secondary BAs to primary BAs ratio (Secondary_primary_ratio) = Total_primary_BAs/Total_secondary_BAs, (5) ratio of CA to CDCA concentrations (CA_CDCA_ratio) = (CA + TCA + GCA)/(CDCA + TCDCA + GCDCA), (6) ratio of unconjugated BA to conjugated BA concentrations = (CA + CDCA + DCA + LCA)/(TCA + GCA + TCDCA + GCDCA + TDCA + GDCA + TLCA + GLCA), (7) ratio of dehydroxylated CA to deconjugated CA concentrations (CA_dehydro_deconju_ratio) = (DCA + TDCA + GDCA)/(CA + TCA + GCA) and (8) ratio of taurine conjugated BA to glycine conjugated BA concentrations (Taurine_glycine_ratio) = (TCA + TCDCA + TDCA + TLCA)/(GCA + GCDCA + GDCA + GLCA).

Metagenomic sequencing and quality control

Microbial DNA was isolated from fecal samples of LLD and 300-OB and sequenced as previously described following the similar protocal (Kurilshikov et al., 2019; Zhernakova et al., 2016). In brief, we isolated DNA from fecal samples with the AllPrep DNA/RNA Mini Kit (Qiagen, Hilden, Germany; catalog No. 80204) as well as the mechanical lysis, and then performed metagenomic shotgun sequencing using Illumina HiSeq platform (Illumina, San Diego, CA). We removed host genome–contaminated reads and low-quality reads from the raw metagenomic sequencing data using KneadData (version 0.7.4), Bowtie2 (version 2.3.4.3) (Langmead and Salzberg, 2012) and Trimmomatic (version 0.39) (Bolger et al., 2014). In brief, the data-cleaning procedure includes two main steps: (1) filtering out the human genome–contaminated reads by aligning raw reads to the human reference genome (GRCh37/hg19) and (2) removing adaptor sequences and low-quality reads using Trimmomatic with default settings (SLIDINGWINDOW:4:20 MINLEN:50).

Taxonomic abundance

We generated the taxonomic relative abundance for both LLD and 300-OB samples from the cleaned metagenomic reads using Kraken2 (version 2.1.2) (Wood et al., 2019) and Bracken (version 2.6.2) (Lu et al., 2017).

Detection of structural variations

Structural variants (SVs) are highly variable genomic segments within bacterial genomes that can be absent from the metagenomes of some individuals and present with variable abundance in other individuals. Based on the cleaned metagenomic reads, we detected the microbial SVs of all 1,437 samples from LLD and 300-OB using SGVFinder with default parameters. SGVFinder was devised and described by (Zeevi et al., 2019) and can detect two types of SV – deletion SVs (dSVs) and variable SVs (vSVs) – from metagenomic data. If the deletion percentage of the genomic segment across the population is < 25%, the standardized coverage will be calculated for this SV (vSV). If the deletion percentage is > 25% and < 75%, only the presence or absence status of this genomic segment will be kept (dSV). If the deletion percentage of a region is > 75%, the region is excluded from the analysis. The SV-calling procedure includes two major steps: (1) resolving ambiguous reads with multiple alignments according to the mapping quality and genomic coverage using the iterative coverage–based read assignment algorithm and reassigning the ambiguous reads to the most likely reference with high accuracy and (2) splitting the reference genomes into genomic bins and then examining the coverage of genomic bins across all samples to identify highly variable genomic segments and detect SVs. We used the reference database provided by SGVFinder, which is based on the proGenomes database (http://progenomes1.embl.de/) (Mende et al., 2017). In total, we detected 5,666 dSVs and 2,616 vSVs from 55 bacteria using default parameters. All bacterial species with SV calling were present in at least 5% of total samples.

Functional annotation

The reference genomes were downloaded from progenome (http://progenomes1.embl.de/) (Mende et al., 2017) and annotated using the web-based genome annotation service provided by PATRIC (version 3.6.6, https://www.patricbrc.org/) (Wattam et al., 2017).

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical tests were performed using R (version 4.0.1). Details of statistical tests are also provided in Results and Figure legends.

Association analysis

We assessed the normality of continuous variables with Shapiro-Wilk normality test and most variables did not show a normal distribution. Before association analysis, all continuous variables were standardized to follow a standard normal distribution ($N\sim(0, 1)$) using empirical normal quantile transformation. Associations between SVs and BAs were assessed in LLD and 300-OB using linear models with the following formula (*model 1*):

$$BA \sim SV + Age + Sex + BMI + Read\ number + Species\ relative\ abundance$$

To identify lineage-independent associations between SVs and BAs, in *model 2* we added the principal components (PCs) of corresponding species-level population genetic makeup to the linear model, along with the covariates included in *model 1*. The number of PCs of each species included were determined by the following procedure: (1) cumulatively summing up PCs until the explained variance proportion reaches 60%, the counted PCs would be added as covariates in *model 2*; (2) if the explained variance proportion was still smaller than 60% when PC10 was counted, the top 10 PCs were added as covariates. The formula of *model 2* was:

$$BA \sim SV + Age + Sex + BMI + Read\ number + Species\ relative\ abundance + PCs$$

The association between species relative abundance and BA parameters were assessed in LLD and 300-OB using linear model with following formula:

$$BA \sim Species\ relative\ abundance + Age + Sex + BMI + Read\ number$$

The association results of LLD and 300-OB were further integrated via meta-analysis with a random-effect model, whereas the statistical heterogeneities were estimated with I^2. To control the false discovery rate (FDR), Benjamini-Hochberg and Bonferroni P-value corrections were performed using the *p.adjust()* function in R. The association analysis and P-value correction were conducted for vSVs, dSVs and species relative abundance separately. The replicable significant SV–BA associations were confirmed with following four criteria: (1) $P_{LLD} < 0.05$, (2) $P_{300\text{-}OB} < 0.05$, (3) $FDR_{meta} < 0.05$ and (4) $P_{heterogeneity} > 0.05$.

The differences of BA parameters between SV-based clusters within species were tested using the Kruskal-Wallis rank-sum test. Empiric P values were estimated based on 999 permutations. For the analysis shown in Figures S6B–S6E, the Spearman correlation coefficient was calculated between the effect size in LLD and 300-OB. In Figure S1A, the mean value ± standard deviation is shown.

The association analysis of BA-associated SVs with cardiometabolic phenotypes and stool characteristics in LLD (phenotype number = 10) and/or 300-OB (phenotype number = 14) were conducted with linear regression (quantitative traits) and logistic regression (binary traits) in LLD and 300-OB separately, and covariates (sex, age, BMI and reads number) were added in the models.

Mediation analysis

The causal relationships between exposure factors, SVs and BAs were inferred by bidirectional mediation analysis using the R package *mediation* (version 4.5.0). To reduce the number of tests, before mediation analysis, we identified lifestyle–SV–BA groups in which all variables correlated with each other as candidate groups with a potential causal relationship. A candidate group had to meet the following criteria: (1) the association between the BA and SV is significant and replicable in both LLD and 300-OB based on *model 1* or *model 2*, (2) the association between BA and lifestyle factor is significant ($P < 0.05$) and (3) the association between lifestyle factor and SV is significant ($P < 0.05$). We then performed bidirectional mediation analysis on the candidate variable groups following the framework described in Figure 7A. For the vSV candidate groups, a linear model was used in each step of mediation analysis. For the dSV candidate groups, a logistic regression model was used when the response variable was a dSV. Finally, the P-values of indirect effects were corrected by FDR estimation.

Distance calculation

We merged the vSV and dSV profiles and calculated Canberra distance between all samples based on the SV profile of each species respectively. We then standardized all matrices by dividing each matrix by its maximum distance value. To quantify the overall microbial genetic kinships between all individuals, we calculated the metagenome-wide genetic dissimilarities between all samples by calculating the distance of shared SVs. To quantify the overall compositional differences of the BA pool, we calculated the Canberra distance between all samples based on BA concentration profile and proportion profile. Distance matrices were computed using the *vegdist()* function from R package *vegan* (version 2.5-6).

PCoA and PERMANOVA

We performed principal coordinates analysis (PCoA) on Canberra distance matrices of SV and BA profiles using the *cmscale()* function from R package *vegan*. We estimated the proportion of BA pool variance explained by basic phenotypes (sex, age and BMI) and cohort factor using permutational multivariate analysis of variance (PERMANOVA) using the *adonis()* function (999 permutations) from R package *vegan*.

To confirm whether the genetic differences between cohorts were confounded by species composition, we selected the top 5 PCs of microbial abundance profile that collectively explained >60% of total compositional variance, then performed PERMANOVA to assess the proportion of genetic variance explained by microbial composition, age, gender, BMI, total read count and cohort factor (999 permutations).

Clustering analysis

Based on the genetic dissimilarity matrix of each species, we clustered the samples using the partitioning around medoid method and assigned samples to clusters with a given cluster number k ($k \in [2, 10]$). The best cluster numbers were determined by prediction strength (PS) (Tibshirani and Walther, 2012), with the highest number of clusters with a PS > 0.55 considered the best cluster number. If there was no PS value > 0.55, we assumed there was no obvious clusters within the corresponding species. The clustering results were then visualized using PCoA plot and t-distributed stochastic neighbor embedding (t-SNE) (Kobak and Berens, 2019).

Protein 3D structure prediction and protein structure-based analysis

We obtained the BA biotransformation genes in the gut microbiome summarized by (Heinken et al., 2019) and performed blastp against reference genomes of species with more than 10 BA associations using the PATRIC BLAST service (https://www.patricbrc.org/app/BLAST). This identified three putative homologue genes located in BA-associated regions (E-value < 10): CK1_08630 from *Ruminococcus* sp. SR1/5, RUMOBE_03494 from *Blautia obeum and* EUBHAL_00727 from *Eubacterium hallii* (EUBHAL_00727). Their protein structures were modeled using the artificial intelligence algorithm AlphaFold2 (Jumper et al., 2021) via ColabFold (Mirdita et al., 2021) and MMseqs2 (Steinegger and Söding, 2017) for predicting protein structure using multiple sequence alignments (Mirdita et al., 2021). Following modeling, the predicted structures were relaxed using amber force fields employing OpenMM 7 (Eastman et al., 2017). The Local Distance Difference Test (LDDT) was used to evaluate the quality of the models (Mariani et al., 2013; Mirdita et al., 2021). Predicted structures were then used in a structure similarity analysis against the entire protein data bank (PDB) to identify crystal structures with similar 3D conformations using the PDBeFold server (https://www.ebi.ac.uk/msd-srv/ssm/) (Krissinel and Henrick, 2004, 2005).