Microbiomics, metabolomics, predicted metagenomics and hepatic steatosis in a population-based study of 1355 adults

Louise JM Alferink, Djawad Radjabzadeh, Nicole S Erler, Dina Vojinovic, Carolina Medina-Gomez, André G Uitterlinden, Robert J de Knegt, Najaf Amin, M Arfan Ikram, Jessica C Kiefte-de Jong, Harry LA Janssen, Herold J Metselaar, Cornelia M. van Duijn, Robert Kraaij, Sarwa Darwish Murad

Hepatology. 2020 Jun 12. In Press.

Abstract

Introduction Previous small studies have appraised the gut microbiome (GM) in steatosis, but large-scale studies are lacking. We studied the association of GM diversity and composition, plasma metabolites, predicted functional metagenomics and steatosis.

Methods This is a cross-sectional analysis of the prospective population-based Rotterdam Study. We used 16Sribosomal-RNA gene sequencing and determined taxonomy using the Silva-reference database. Alpha-and beta-diversity were calculated using Shannon-index and Bray-Curtis dissimilarities. Differences were tested across steatosis using PerMANOVA. Hepatic steatosis was diagnosed by ultrasonography. We subsequently selected genera using regularised regression. The functional metagenome was predicted based on the GM using KEGG-pathways. Serum metabolomics were assessed using high-throughput proton nuclear magnetic resonance. All analyses were adjusted for age, sex, BMI, alcohol, diet, and proton-pump inhibitors.

Results We included 1355 participants of which 472 had steatosis. Alpha-diversity was lower in steatosis (P=1.1·10^{-9}) and beta-diversity varied across steatosis strata (P=0.001). Lasso selected 37 genera of which three remained significantly associated after adjustment (*Coprococcus3:* β=–65; *Ruminococcus Gauvreauiigroup*: β=62; and *Ruminococcus Gnavusgroup*: β=45, *Q*-value =0.037). Predicted metagenome analyses revealed that pathways of secondary bile-acid synthesis and biotin metabolism were present and D-alanine metabolism was absent in steatosis. Metabolic profiles showed positive associations for aromatic-and branched chain amino acids and glycoprotein acetyls with steatosis and *R. Gnavusgroup*, whereas these metabolites were inversely associated with alpha-diversity and *Coprococcus3*.

Conclusion We confirmed, for the first time on a large-scale, the lower microbial diversity and association of *Coprococcus* and *Ruminococcus Gnavus* with steatosis. We additionally showed that steatosis and alpha-diversity share –opposite– metabolic profiles.

Introduction

Hepatic steatosis is the most common liver disease worldwide[280] and its presence is implicated in the development of advanced liver disease as well as cardiovascular disease.[182] Indeed, at present hepatic cirrhosis due to alcoholic and non-alcoholic fatty liver disease together is the number one indication for liver transplantation,[399] and its co-incidence with other liver diseases, such as viral hepatitis, worsens disease severity.[400] Aetiology of hepatic steatosis is multifactorial, but unhealthy lifestyle habits, such as excessive alcohol consumption in alcoholic fatty liver disease and an unhealthy diet in non-alcoholic fatty liver disease (NAFLD), are key in the development of this disease.[17] Aside from these well-known risk factors, there is an increasing interest in the contribution of the gut microbiome to steatogenesis.[401] The gut microbiome refers to the collective genomic content of microbiota in the intestinal tract consisting of over 100 trillion microbes, being mainly bacteria (>99%).[402] These gut microbes are crucial for life as they perform essential functions such as energy harvest, host nutrition, and immunomodulation.[403]

Hepatic steatosis is closely related to obesity.[6] There is a myriad of studies that demonstrate the importance of the gut microbiome in the development and progression of obesity.[404] In addition, Le Roy and colleagues have shown that diet-induced steatosis in mice was directly dependent on the gut microbial composition. Furthermore, the authors showed that the steatosis phenotype was transmissible via the gut microbiome in germ-free mice.[405] Subsequent studies in humans tried to link the presence of steatosis with specific bacteria, but these endeavours to define a so-called 'core microbiome' in steatosis have been highly inconsistent. This inconsistency is possibly due to the limited sample size, heterogeneous study populations (paediatric vs adult), different phenotypes (simple steatosis vs steatofibrosis), and diverse analytical approaches of these studies.[406-417] Importantly, most of these studies did not correct for important confounders such as diet, body mass index (BMI), and alcohol consumption, making it hard to draw inferences on the independent association between microbes and steatosis.

The close connection between the gut microbiome and the liver is referred to as the 'gut-liver axis' – as microbiome-derived metabolites and other bacterial products can easily reach the liver via the portal vein.[401] Zhu et al. for instance proposed that ethanol-producing bacteria in the gut microbiome of children contributed to the pathogenesis of steatosis, mimicking alcoholic fatty liver disease.[409] In addition, it has been shown that gut microbiota are the major driver of circulating lipid levels such as triglycerides and other fatty acids that are known risk factors for steatosis.[6,418]

In this large epidemiologic cohort study, we aimed to characterize the composition of the gut microbiome in individuals with steatosis. In addition, we examined predicted functional pathways of the gut microbiome in steatosis and we assessed the metabolic profiles of steatosis and the gut microbiome by means of high-throughput proton Nuclear Magnetic Resonance (NMR) metabolomics.

Subjects and methods

Study Population

This study is embedded in the Rotterdam Study, a prospective cohort study in Ommoord, a suburb in the city of Rotterdam, the Netherlands. A detailed description on study design and rationale can be found in a recent overview paper.[248] This study is cross-sectional, all analyses are performed in cohort III visit 2 (May 2012 to June 2014), in which participants were asked to collect faecal samples for microbiome analyses. Inclusion criteria for participation in the current study were availability of gut microbiome and ultrasound data and age of 45 years or above at initial visit. The Rotterdam Study has been conducted in accordance with the Declaration of Helsinki and has been approved by the institutional review board (Medical Ethics Committee) of the Erasmus Medical Centre and by the review board of The Netherlands Ministry of Health, Welfare and Sports. All participants provided written informed consent.

Liver imaging

Abdominal ultrasound was performed by an experienced nurse ultrasonographist (PvW) using the Hitachi HI VISION 900. The definition of hepatic steatosis was overt hyperechogeneity of the liver parenchyma as compared to that of the kidney.[104] We dichotomized hepatic steatosis, because of the poor performance for the detection of mild steatosis and known subjective ultrasonographic grading of steatosis.[419] We analysed a subset of participants in which the steatosis was unlikely to be secondary, i.e. ruling out excessive alcohol consumption, steatogenic drugs, and viral hepatitis. The remaining group was eligible for the diagnosis non-alcoholic fatty liver disease (NAFLD). In addition, liver stiffness measurements (LSM) were carried out in all participants using transient elastography (Fibroscan®, EchoSens, Paris, France). Those with steatosis and elevated LSM (of 8 kilopascals [kPa] or higher) were assumed to have more advanced fatty liver disease, i.e. (non-alcoholic) steatofibrosis (NASF).[54] The use and reliability criteria of transient elastography have been published in more detail previously.[78]

Faecal samples

Participants were instructed to collect faecal samples in sterile tubes at home in a standardized fashion. Participants were also asked to disclose recent antibiotic use (past twelve months), journeys abroad (past month), and probiotic use (past three months). Samples were then returned to the Erasmus MC through postal mail. Upon arrival, samples were recorded and stored at –20°C. An aliquot of approximately 300mg was homogenized in

stool stabilizing buffer according to the manufacturer's protocol (Arrow Stool DNA; Isogen Life Science, De Meern, the Netherlands). Homogenized samples were bead in lysing Matrix B tubes containing 0.1mm silica beads (MP Biomedicals®, LLC, Bio Connect Life Sciences BV, Huissen, the Netherlands). Samples were then centrifuged and the supernatant was subjected to automated DNA isolation according to the manufacturer's protocol (Arrow; DiaSorin S.P.A., Saluggia, Italy). Isolated DNA was then stored at –20°C.

Metadata

All covariates used in this study were derived by an extensive home interview, by drawing fasting blood samples, by automated linkage with the local pharmacy and by completion of an external-validated 389-item food frequency questionnaire. Detailed information on these covariates can be found in the *Supplementary Methods.*

Gut microbiome data set generation

For a detailed description of the gut microbiome dataset generation, we refer to the recent paper of *Radjabzadeh et al*.[420] In short, 16S ribosomal RNA (rRNA) gene amplification of variable regions 3 and 4 was used to sequence all samples by the Illumina MiSeq® platform v3. The sequence data were rarefied at 10,000 reads per sample and reads were clustered into operational taxonomic units (OTU) with 97% similarity. These OTUs were classified using the RDP classifier (version 2.12)[421] and Silva rRNA database project® (v128).[422] It has been demonstrated that quality-filtering 16S amplicon sequence reads can greatly improve accuracy of microbial community analysis, we therefore used a cut-off threshold excluding OTUs with a total read count of less than 0.005% of the total reads.[423] Alpha diversity was calculated at this step. OTUs presented in less than 1% of the samples were also removed in order to harmonize the dataset.

Functional metagenome

In order to predict bacterial function in steatosis, we used Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt (v.1.1.0)) algorithm.[424] We then used HUMAnN2 (v0.99) to identify Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways.

Metabolite assessment

All metabolites were quantified using EDTA plasma samples by high-throughput proton NMR metabolomics (Nightingale Health, Helsinki, Finland). Details on this methodology

have been published previously.[425] We quantified metabolic measures, including various sizes of lipoproteins, serum lipids, various fatty acids, branched-chain amino-acids (BCAAs), aromatic amino acids (AAAs) and glycolysis-related metabolites. In this study we analysed 146 metabolic measures.

Statistical analyses

Imputation analyses and participant characteristics

We first excluded participants of whom sequencing failed due to technical issues and those with duplicate samples, samples with unknown travel time or over 3 days of travel time between collection and analysis, and participants with recent antibiotic use. Thereafter, we excluded participants that had no ultrasound data available or that had unreliable food frequency questionnaires (i.e. <500 or >7500 kilocalories/day). Missing values in the meta-data were imputed using multiple imputation under the fully conditioned specification to reduce bias due to missing data. We created 50 imputed datasets using the R Package mice. The results from these analyses were pooled using Rubin's rules in order to take into account the added uncertainty due to the missing of data. The *Supplementary Method* contains a detailed description of the imputation process and references of the statistical methods described in this method section.
Characteristics of the study population after imputation were described across steatosis strata. Differences were appraised using analyses of variance (ANOVA) for approximately normally distributed variables, Kruskall-Wallis tests for continuous non-normally distributed data, and Chi-squared tests for categorical data.

Description of the microbiome

The median relative abundance of all phyla was evaluated. In addition, we examined the Firmicutes-to-Bacteroidetes-ratio (FBR). Richness and alpha-diversity, as determined by Shannon and Inverse-Simpson indices, were calculated. Difference in alpha-diversity between steatosis strata was visualized using violin plots and evaluated using Wilcoxon rank sum test with continuity correction. The beta-diversity was studied using principal coordinate analysis applied to Bray-Curtis dissimilarity using the function "capscale" from the R package vegan on genus level.[426] Sample clustering of genera was tested for association with steatosis, using permutational multivariate analysis of variance (PerMANOVA, n=999) on Bray-Curtis dissimilarities. We evaluated the percentage of explained variation in the gut microbiome variability by steatosis (r^2= 0.0030) and NAFLD (r^2= 0.0029).
In order to investigate the association between steatosis and all genera simultaneously, logistic regression was performed. To prevent overfitting regression coefficients of the gen-

era were regularized using lasso.[427] This technique extends ordinary regression by imposing penalty terms on the regression coefficients resulting in shrinkage towards zero, such that coefficients of covariates that do not sufficiently contribute to the model are set to zero. To additionally take into account the compositional nature of the microbiome we followed the approach of Lin et al., which is implemented in the R package zeroSum.[428,429] For this selection of genera, relative abundances were log-transformed (log10[1+genus]) and the model was adjusted for batch-effect and travel time of samples. The degree of penalization was determined by 10-fold cross-validation, which was repeated 20 times in each imputed dataset to average out effects of random sampling. The degree of penalization resulting in the on average smallest cross-validation error per imputed dataset was then applied in the analysis of that respective dataset. Genera with coefficients different from zero in at least half of the imputed datasets were selected as predictor variables in the final logistic (unrestricted) regression model.

To take into account potential confounding we considered the following covariates (based on previous studies)[6] for additional adjustment of the logistic regression for steatosis: age, sex, BMI, education level, energy intake, alcohol units, smoking status, dietary quality scores, physical activity, diabetes, proton-pump inhibitor (PPI) use, serum triglycerides and high-density lipoprotein (HDL) cholesterol. The selection of covariates on the basis of previous studies in which these covariates were associated with steatosis.[6] Due to the limited number of steatosis cases, however, we were only able to include a subset that we expected to have the highest confounding potential: age, sex, BMI, technical variables, alcohol intake, dietary quality scores and PPI-use. This selection was based on pathophysiological background knowledge. Models were evaluated for multicollinearity using the variable inflation factor. Predicted probabilities for steatosis from this model were plotted against the relative abundance of selected significant bacteria to facilitate the interpretation of the results.

Predicted functional metagenomics

Predicted metagenomics pathways were tested for significance by performing logarithmic linear discriminant analysis (LDA) in linear discriminant analysis effect size with the significance level of 0.05 and the logarithmic LDA score threshold equal to 2.0.

Metabolic profiles

The metabolic profile (as exposure) was derived using linear or logistic regression analyses for steatosis, alpha-diversity, and significant genera while adjusting for abovementioned covariates plus the technical variable: metabolic batch. The adjusted effect estimates were plotted in a heatmap. We dealt with zeros in metabolic measures by adding half of the

minimum non-zero values to all values, we then transformed (log10) and standardized all metabolic measures (n=146). As most of the metabolites we studied concern lipid-related measures, sensitivity analyses excluding participants with lipid-lowering medication was performed.

NAFLD analyses

We performed a subgroup analysis on the associations between the microbiome, metabolites, and predicted metagenomics, excluding all participants with potential secondary causes for steatosis (i.e. NAFLD and NASF). Although we have included alcohol intake as covariate in the multivariable models, we would like to see if the gut microbiome, metabolic profile and predicted functional metagenomic in NAFLD is different from the overall steatosis group, or whether similar pathways and bacteria are involved.

Statistical Significance

Adjusted *P*-values (referred to as *Q*-values) were calculated using the Benjamini & Hochberg correction taking together all genera or metabolites within the same model,[430] to account for the inflated type I error that arises due to multiple testing. All computations were done using R version 3.5.2 (R Core Team [2018]), the package Vegan version 2.5.2 [2018], and the package ZeroSum version 1.1.1.

Results

Participant Characteristics

The flowchart of the study is depicted in Figure 1. In total, 1739 out of 2440 individuals participated (71.3%). The final study population comprised 1355 participants. Population characteristics (Supplementary Table 1) were as follows: mean age was 62.4 (5.9) years, mean BMI was 27.4 (4.5), 57.8% was female, 34.8% had steatosis, and the majority was of European descent (96.7%). Characteristics of the steatosis strata are presented in Table 1.

Description of the microbiome

We identified 11 different phyla in our dataset. The relative abundance of these phyla is illustrated in *Figure 2* and median relative abundances are presented in *Supplementary Table 2*. The most prevalent phylum was *Firmicutes* (81.2%), followed by *Bacteroidetes* (8.6%), *Actinobacteria* (2.7%), and *Proteobacteria* (1.4%). Phyla *Verrucomicrobia*, *Tenericutes*,

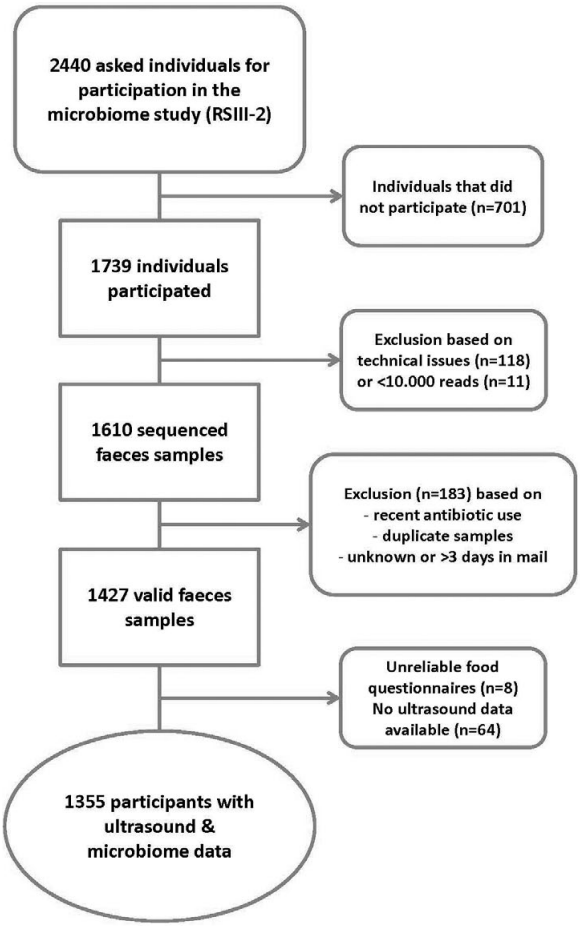


Figure 1: Flowchart of the included study population

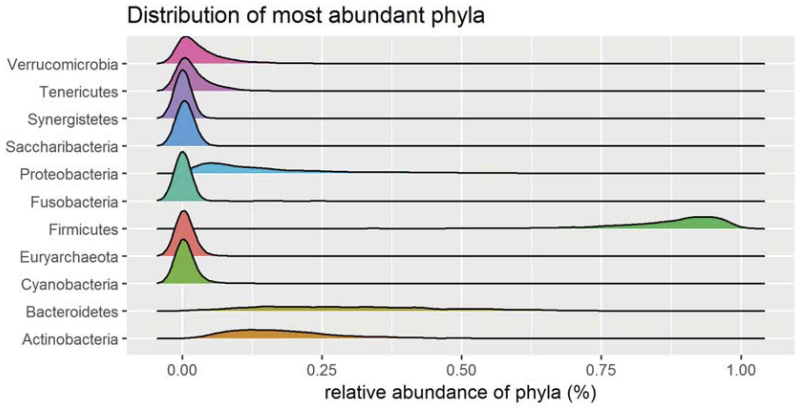
Proteobacteria and *Euryarchaeota* were less prevalent in steatosis. FBR was not different across steatosis strata (*Figure 3*; *P*=0.335). Richness and alpha-diversity were significantly lower in steatosis (*Figure 3, P-values* $\leq 2.0 \cdot 10^{-6}$). The principle coordinate analyses plots are depicted in *Supplementary Figure 1*. Although in the plots the variation between the two groups was not clearly visible, PerMANOVA showed that sample clustering was significantly different for steatosis (P<0.001), taking into account the sequentially added technical variables, age, BMI, and sex. Lasso regularized regression identified 37 genera. These genera are listed in *Supplementary Table 3* with their median relative abundances. In subsequent multivariable logistic regression analyses for steatosis including these 37 genera; three of them had a statistically significant association (*Table 2*). *Coprococcus3* was inversely associated with steatosis, whereas *Ruminococcus Gauvreauiigroup* and *Ruminococcus Gnavusgroup* were positively associated with steatosis. Relative abundances were

Table 1: Population Characteristics

	No steatosis n=883 (65.2%)	Steatosis n=472 (34.8%)	*P*-value*
Age (years)	62.19 (5.92)	62.69 (5.73)	0.136
Female (%)	60.9	51.9	0.002
European (%)	96.1	97.9	0.191
Education Level (%)			0.095
Low	40.1	43.3	
Intermediate	27.3	29.9	
High	32.6	26.9	
Smoking status (%)			0.030
Never	36.7	31.8	
Past or Current	63.3	68.2	
Alcohol (units/day)	0.81 [0.16, 1.92]	0.91 [0.20, 2.18]	0.514
Physical Activity (METh/wk)	50.85 [22.85, 84.85]	41.45 [17.50, 76.59]	0.019
Energy intake (kcal/day)	2276 [1906, 2725]	2214 [1831, 2738]	0.972
Dietary Quality Score (0–15)	7 [6, 8]	7 [6, 8]	0.047
BMI (kg/m^2)	26.12 (3.90)	29.86 (4.49)	<0.001
Waist-to-hip ratio	0.86 [0.80, 0.93]	0.94 [0.88, 1.00]	<0.001
HOMA-IR	2.10 [1.49, 3.01]	3.94 [2.62, 5.95]	<0.001
Total Cholesterol (mmol/L)	5.63 (1.10)	5.49 (1.12)	0.035
HDL-Cholesterol	1.60 (0.47)	1.33 (0.37)	<0.001
Triglycerides	1.16 [0.90, 1.55]	1.58 [1.19, 2.08]	<0.001
Diabetes Mellitus (%)	5.6	18.2	<0.001
Hypertension (%)	52.3	69.2	<0.001
Liver stiffness measurements (kPa)†	4.40 [3.60, 5.40]	4.90 [3.90, 6.10]	<0.001
LSM≥8.0 kPa n (%)	17 (1.9)	32 (6.8)	<0.001
Lipid-lowering drug use	24.2	32.8	<0.001
Proton-pump inhibitor use	15.5	21.8	<0.001
Microbiome batch #0/1	78.8 / 21.2	78.8 / 21.2	1.000
Time in mail (%)			0.777
1 day	64.8	66.3	
2 days	29.8	28.0	
3 days	5.4	5.7	
Secondary causes for steatosis n (%)‡	145 (20.1)	83 (22.7)	0.378
Viral hepatitis	5 (0.6)	3 (0.6)	1.000
Excessive alcohol use	128 (17.6)	74 (20.2)	0.328
Steatogenic drug use	14 (1.6)	7 (1.5)	1.000
NAFLD§ n (%)	-	283 (60.0)	-

Data are presented as median value (P25-P75), as mean value (SD) or as percentage.*P-value is assessed using ANOVA, Kruskall-Wallis or Chi-squared tests. †n=1195 reliable measurements (n=783 for no steatosis and n=412 for steatosis), ‡data for n=1085 participants, § data for n=857 participants.
Abbreviations: HOMA-IR: homeostasis model of insulin resistance; kcal: kilocalories; kPa: kilopascals; LSM: liver stiffness measurement; METeqh/wk: metabolic equivalent task hours per week; NAFLD: non-alcoholic fatty liver disease

A



B

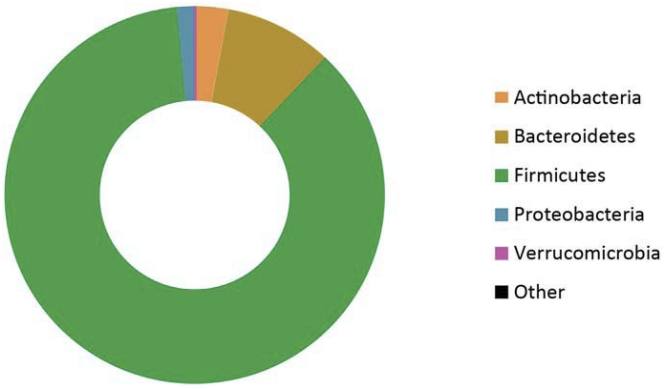


Figure 2: Relative abundance of phyla in the total population

A) This figure depicts the distribution of the relative abundance per phylum (using squared root transformation). The relative abundance of all phyla together add up to one. On the y-axis, the relative abundance is shown. On the x-axis the distribution of the relative abundance in the total population is shown. B) This donut plot depicts the median relative abundance per phylum in percentages. The relative abundance of all phyla here add up to 100%. Please see Supplementary Table 2 for the absolute median percentages and the P25-P75.

plotted across steatosis strata (*Supplementary Figure 2A*). There was no evidence for sex differences nor for interaction of these genera with steatosis. The predicted probability of steatosis across relative abundances of these genera, for exemplary men and women with fixed BMI values (while fixing all other covariates in the model to reference values) were plotted (*Supplementary Figure* 3).

Figure 3: Visualization of richness and alpha-diversity and FBR using violin plots across steatosis strata.
Violin plots of the distribution of A-B) alpha-diversity as assessed by the Shannon-index and Inverse Simpson, C) microbial richness, and D) log-transformed Firmicutes-to-Bacteroidetes-ratio by steatosis strata. Violin plots have additional value to boxplots since they show the full distribution of the data, whereas boxplots only show a summary. But the basic idea is similar. The statistical difference between strata was assessed using Wilcoxon rank sum test with continuity correction and was as follows: A) Shannon: P= 1.1·10–9; B) Inverse Simpson: P= 2.0·10–6; C) Microbial Richness: P=7.3·10–9 and; D) Firmicutes-to-Bacteroidetes-ratio: P=0.335.
Abbreviations: InvSimpson: Inverse Simpson; logFBratio: log-transformed Firmicutes-to-Bacteroidetes-ratio.

Predicted functional metagenome

Predicted analyses revealed four pathways that were present and two that were absent in steatosis (*Figure 4*). The most present pathways were that of "Secondary bile acid biosynthesis" (ko00121) and that of "biotin (vitamin H or B7) metabolism" and acetyl-CoA carboxylase (ko00780). Another bile acid-related pathway "Taurine and hypotaurine" (ko00430) was absent in steatosis as well as "D-alanine metabolism" (ko00473).

Metabolic profiles

Figure 5 A-E shows the independent metabolic profiles of steatosis, alpha diversity, and the three significant genera. The top-hits for steatosis were BCAAs isoleucine and leucine. Both were independently associated with higher steatosis prevalence (*Figure 5E*). In addition, valine (also a BCAA), alanine (a non-essential amino-acid), and the AAAs tyrosine and phenylalanine were significantly associated with higher steatosis prevalence. The acute

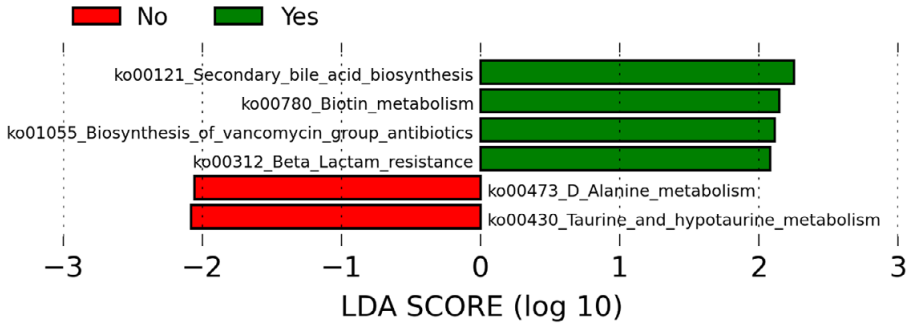


Figure 4: Predicted metagenomics pathways for steatosis.
This figure depicts six metabolic pathways that were either present (Yes in green) or absent in steatosis (No in red). The predicted metagenomics are based on 16S rRNA gene sequencing data from our study cohort. Linear discriminant effect size was calculated and shown on the x-axis based on PICRUSt. A metabolic pathway was considered significant if the false discovery rate was lower than 0.05.

phase reaction marker glycoprotein acetyls was positively associated with steatosis (*Figure 5E*). Furthermore, multiple significant positive associations were detected for very low-density lipoproteins (VLDL) particles of all sizes (*Figure 5A*), saturated fatty acids, mono-unsaturated fatty acids and total triglycerides (*Figure 5D*). Positive associations were found for large and extra-large HDL particles, whereas there was an inverse association for small HDL particles (*Figure 5C*). Additionally, effect estimates for steatofibrosis were plotted, and we found that glucose and lactate were significantly associated with both steatosis and steatofibrosis. Glutamine had a significant inverse association with steatofibrosis, but not with steatosis (*Figure 5E*).

The metabolic profiles of richness and alpha-diversity mirrored those of steatosis, having the –often significant– opposite direction with metabolites (*Figure 5*). As for the significant genera, the only significant association found was that of *Ruminococcus Gnavusgroup* with glycoprotein acetyls (*Figure 5E*).

The sensitivity analysis of metabolic profiles, excluding participants with lipid-lowering medication, showed attenuated but largely similar results (*Supplementary Figure 4*).

NAFLD analyses

The subgroup NAFLD included 857 participants of which 283 (33%) had NAFLD. Results on differences in relative abundance of phyla, FBR, and alpha-diversity resembled across NAFLD strata resembled steatosis (*Supplementary Table 4* and *Supplementary Figure 5*). Beta-diversity was also significantly different across NAFLD strata (sequentially adjusted PerMANOVA: P=0.003). Similar significant results were found for multivariable regression analysis in NAFLD (*Supplementary Table 5*). Relative abundance of these genera associated with NAFLD and NASF were plotted in *Supplementary Figure 2C* and *2D*. Relative

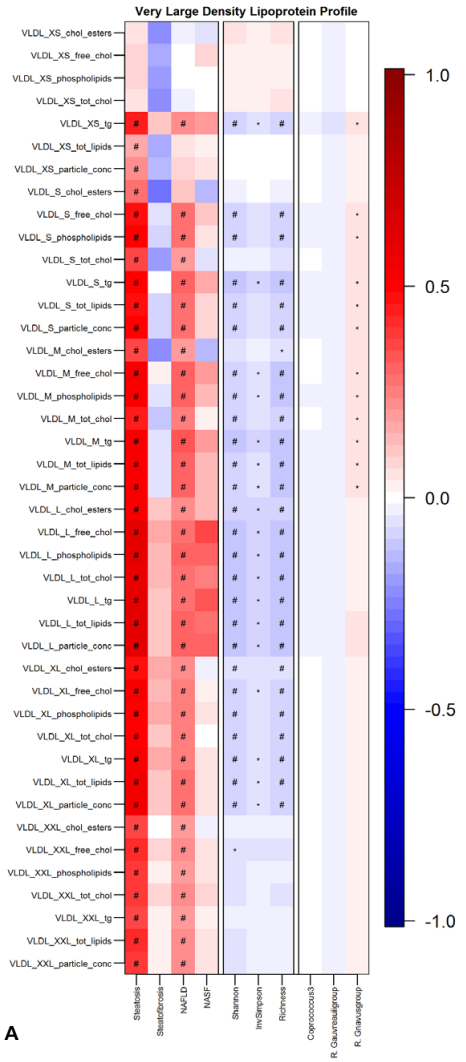
Table 2: Multivariable logistic regression analysis of lasso-selected genera with steatosis as dependent variable

Phylum			Variables	β	95%CI	P-value	Q-value
Class	Order	Family					
			intercept	−5.48	-	-	-
			batch 0	ref			
			batch 1	−0.002	−0.33, 0.33	0.992	-
			Time in mail	−0.02	−0.24, 0.21	0.884	-
			age (per year)	0.004	−0.02, 0.03	0.708	-
			sex (women)	−0.35	−0.63, −0.07	**0.013**	-
			BMI (per kg/m²)	0.21	0.18, 0.25	**<0.001**	-
			alcohol (per unit)	0.11	0.01, 0.21	**0.025**	-
			PPI use	0.33	−0.02, 0.67	0.060	-
			dietary quality - score	−0.03	−0.11, 0.06	0.530	-
Actinobacteria							
Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	*Bifidobacterium*	−2.87	−12.33, 6.59	0.552	0.796
Coriobacteriia	Coriobacteriales	Coriobacteriaceae	*Collinsella*	31.36	−11.09, 73.81	0.148	0.399
Bacteroidetes							
Bacteroidia	Bacteroidales	Bacteroidaceae	*Bacteroides*	−1.07	−8.30, 6.16	0.771	0.839
		BacteroidalesS247gr	*Unknowngenus*	−17.76	−51.24, 15.71	0.298	0.537
		Prevotellaceae	*Prevotella9*	−2.12	−9.87, 5.62	0.591	0.810
		Rikenellaceae	*Alistipes*	−2.24	−28.27, 23.79	0.866	0.871
Firmicutes							
Bacilli	Lactobacillales	Lactobacillaceae	*Lactobacillus*	−1.91	−20.13, 16.31	0.837	0.871
		Streptococcaceae	*Streptococcus*	−5.10	−13.44, 3.24	0.231	0.494
		Christensenellaceae	*ChristensenellaceaeR7gr*	−9.89	−23.39, 3.62	0.151	0.399
		Clostridiaceae	*Clostridiumsensustricto1*	−3.07	−21.01, 14.87	0.737	0.839
			Blautia	−2.53	−9.79, 4.74	0.495	0.796

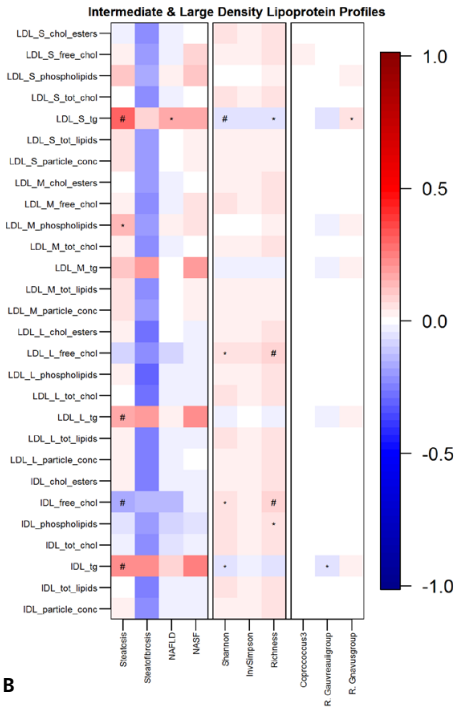
			Butyrivibrio	−7.13	−29.88, 15.62	0.539	0.796
			Coprococcus3	**−65.26**	**−106.78, −23.74**	**0.002**	**0.037**
			Dorea	−4.55	−25.34, 16.23	0.668	0.823
			Fusicatenibacter	−4.81	−20.98, 11.35	0.559	0.796
			Lachnoclostridium	−4.65	−31.85, 22.56	0.738	0.839
		Lachnospiraceae	*LachnospiraceaeND3007gr*	−73.91	−140.92, −6.90	0.031	0.227
			Eligensgr	−18.68	−54.40, 17.03	0.305	0.537
			R. Gauvreauiigr	**61.72**	**24.90, 98.54**	**0.001**	**0.037**
			R. Gnavusgr	**44.68**	**15.23, 74.14**	**0.003**	**0.037**
			Halliigr	−2.19	−16.08, 11.71	0.758	0.839
Clostridia	Clostridiales		*Rectalegr*	1.93	−6.83, 10.68	0.666	0.823
		Peptostreptococcaceae	*Romboutsia*	−21.16	−44.15, 1.84	0.071	0.377
			Faecalibacterium	−6.28	−13.75, 1.19	0.100	0.399
			RuminococcaceaeNK4A214gr	−19.80	−73.21, 33.60	0.467	0.786
			RuminococcaceaeUCG002	−15.73	−34.73, 3.27	0.104	0.399
		Ruminococcaceae	*RuminococcaceaeUCG014*	−6.84	−15.74, 2.07	0.132	0.399
			Ruminococcus1	−5.92	−29.58, 17.74	0.624	0.823
			Ruminococcus2	15.99	3.03, 28.95	0.016	0.144
			Subdoligranulum	−7.10	−18.95, 4.75	0.240	0.494
			Coprostanoligenesgr	13.35	−6.10, 32.79	0.178	0.440
		Erysipelotrichaceae	*Catenibacterium*	−2.36	−30.88, 26.16	0.871	0.871
		Acidaminococcaceae	*Phascolarctobacterium*	19.20	−6.53, 44.92	0.144	0.399
Negativicutes	Selenomonadales	Veillonelaceae	*Megamonas*	20.95	−18.77, 60.67	0.301	0.537
Proteobacteria							
	Aeromonadales	Succinivibrionaceae	*Succinivibrio*	21.54	−7.41, 50.50	0.145	0.399
Gammaproteo-bacteria			*EscherichiaShigella*	−6.54	−12.86, −0.23	0.042	0.261
	Enterobacteriales	Enterobacteriaceae	*Klebsiella*	−14.68	−38.38, 9.02	0.225	0.494

Logistic regression analyses with steatosis as dependent variable. All genera (log10(1+genus)) are analyzed simultaneously. Results are given as regression coefficient (95%CI), nominal *P*-value and *Q*-value-corrected for multiple comparisons using Benjamini & Hochberg.

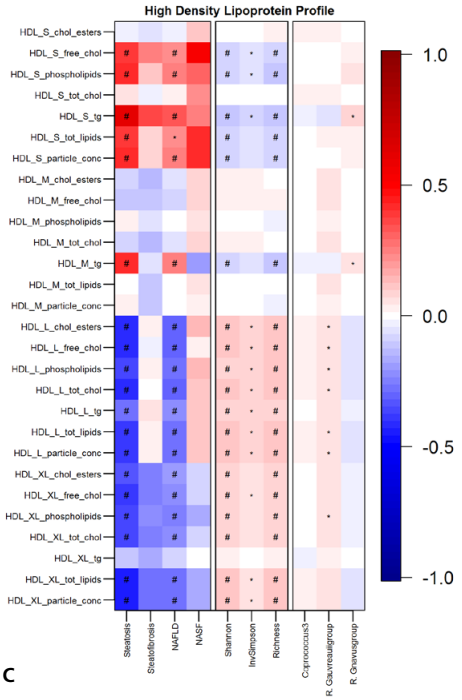
Abbreviations: BMI: body mass index; PPI: proton-pump inhibitor.


Very Large Density Lipoprotein Profile
VLDL_XS_chol_esters
VLDL_XS_free_chol
VLDL_XS_phospholipids
VLDL_XS_tot_chol
VLDL_XS_tg
VLDL_XS_tot_lipids
VLDL_XS_particle_conc
VLDL_S_chol_esters
VLDL_S_free_chol
VLDL_S_phospholipids
VLDL_S_tot_chol
VLDL_S_tg
VLDL_S_tot_lipids
VLDL_S_particle_conc
VLDL_M_chol_esters
VLDL_M_free_chol
VLDL_M_phospholipids
VLDL_M_tot_chol
VLDL_M_tg
VLDL_M_tot_lipids
VLDL_M_particle_conc
VLDL_L_chol_esters
VLDL_L_free_chol
VLDL_L_phospholipids
VLDL_L_tot_chol
VLDL_L_tg
VLDL_L_tot_lipids
VLDL_L_particle_conc
VLDL_XL_chol_esters
VLDL_XL_free_chol
VLDL_XL_phospholipids
VLDL_XL_tot_chol
VLDL_XL_tg
VLDL_XL_tot_lipids
VLDL_XL_particle_conc
VLDL_XXL_chol_esters
VLDL_XXL_free_chol
VLDL_XXL_phospholipids
VLDL_XXL_tot_chol
VLDL_XXL_tg
VLDL_XXL_tot_lipids
VLDL_XXL_particle_conc
Steatosis
Steatofibrosis
NAFLD
NASF
Shannon
InvSimpson
Richness
Coprococcus3
R. Gauvreauiigroup
R. Gnavusgroup
1.0
0.5
0.0
-0.5
-1.0

A


Intermediate & Large Density Lipoprotein Profiles
LDL_S_chol_esters
LDL_S_free_chol
LDL_S_phospholipids
LDL_S_tot_chol
LDL_S_tg
LDL_S_tot_lipids
LDL_S_particle_conc
LDL_M_chol_esters
LDL_M_free_chol
LDL_M_phospholipids
LDL_M_tot_chol
LDL_M_tg
LDL_M_tot_lipids
LDL_M_particle_conc
LDL_L_chol_esters
LDL_L_free_chol
LDL_L_phospholipids
LDL_L_tot_chol
LDL_L_tg
LDL_L_tot_lipids
LDL_L_particle_conc
IDL_chol_esters
IDL_free_chol
IDL_phospholipids
IDL_tot_chol
IDL_tg
IDL_tot_lipids
IDL_particle_conc
Steatosis
Steatofibrosis
NAFLD
NASF
Shannon
InvSimpson
Richness
Coprococcus3
R. Gauvreauiigroup
R. Gnavusgroup
1.0
0.5
0.0
-0.5
-1.0

B


High Density Lipoprotein Profile
HDL_S_chol_esters
HDL_S_free_chol
HDL_S_phospholipids
HDL_S_tot_chol
HDL_S_tg
HDL_S_tot_lipids
HDL_S_particle_conc
HDL_M_chol_esters
HDL_M_free_chol
HDL_M_phospholipids
HDL_M_tot_chol
HDL_M_tg
HDL_M_tot_lipids
HDL_M_particle_conc
HDL_L_chol_esters
HDL_L_free_chol
HDL_L_phospholipids
HDL_L_tot_chol
HDL_L_tg
HDL_L_tot_lipids
HDL_L_particle_conc
HDL_XL_chol_esters
HDL_XL_free_chol
HDL_XL_phospholipids
HDL_XL_tot_chol
HDL_XL_tg
HDL_XL_tot_lipids
HDL_XL_particle_conc
Steatosis
Steatofibrosis
NAFLD
NASF
Shannon
InvSimpson
Richness
Coprococcus3
R. Gauvreauiigroup
R. Gnavusgroup
1.0
0.5
0.0
-0.5
-1.0

C

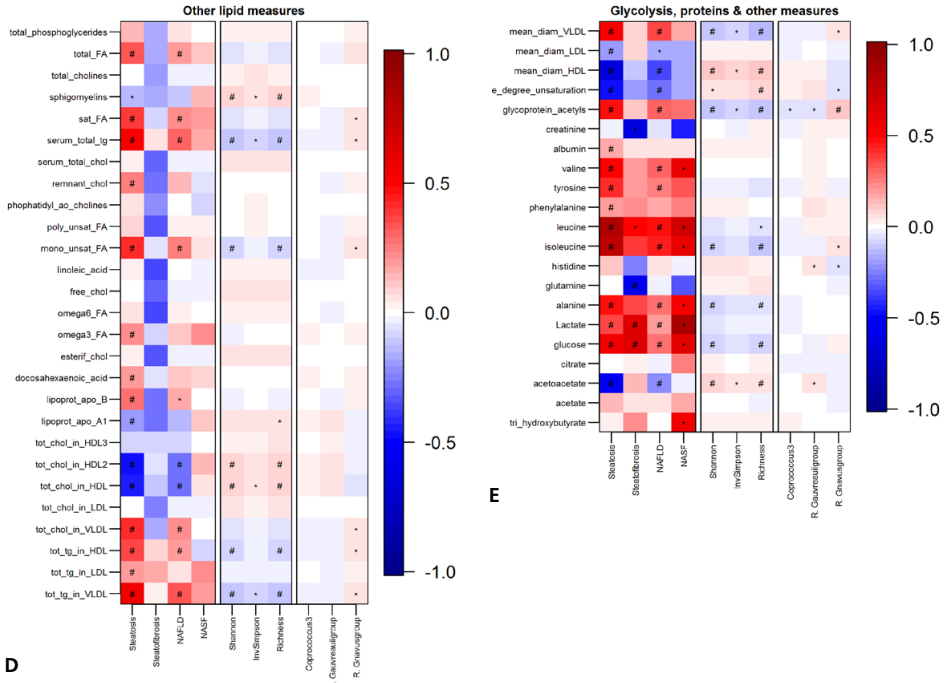


Figure 5: Metabolic profiles of steatosis, alpha-diversity and lasso-selected genera.
The colors in columns represent the standardized betas of the metabolites (as exposure) with steatosis, steatofibrosis, NAFLD, NASF, alpha-diversity and the significant genera (as various outcomes) in separate multivariable logistic regression models. The betas were adjusted for covariates: sex†‡, age†‡, technical variables†‡, BMI†‡, Dietary Quality-Score, PPI use, and alcohol units†. We corrected for multiple testing using Benjamini & Hochberg, significant Q-values are depicted by #, nominally significant values by *. †Steatofibrosis (n=32) and ‡NASF (n=14) were adjusted for less covariates in the multivariable models because of risk of overfitting.
Abbreviations: BMI: body mass index; chol: cholesterol; conc: concentration; diam: diameter; FA: fatty acids; HDL: high-density lipoprotein; IDL: intermediate density lipoprotein; InvSimpson: Inverse Simpson; L: large; LDL: low-density lipoprotein; M: medium; NAFLD: non-alcoholic fatty liver disease; NASF: non-alcoholic steatofibrosis; PPI: proton-pump inhibitor; S: small; sat: saturated; tg: triglycerides; tot: total; VLDL: very low density lipoprotein; (X)XL: (extra) extra large

abundance of *Coprococcus3* was markedly lower in NASF as compared to simple steatosis. Predicted metagenomics in NAFLD largely overlapped the results for steatosis for "secondary bile acid biosynthesis" (ko00121), "biotin metabolism" (ko00780) and "D-alanine metabolism" (ko00473), as depicted in *Supplementary Figure 6*. Lastly, the metabolic profile for NAFLD was almost identical to that of steatosis, as was that of NASF compared to steatofibrosis (*Figure 5*).

Discussion

Over the last decade, the role of the gut microbiome in human health has gained global interest. The liver, which receives approximately 70 percent of the blood outflow from the gut, is a preeminent example of the cross talk between our other genome –the gut microbiome– and the rest of the body. There have been many studies that appraised the gut microbiome in steatosis.[406-417] However, results were rather inconsistent, which is not surprising given the relatively small sample size of the studies (maximum n=156), heterogeneous study designs and the fact that the gut microbiome is highly variable between individuals. With our study we had the unique possibility to examine the diversity and composition of the gut microbiome and its associated metabolites and predicted metagenomics in the context of (non-alcoholic) steatosis and its advanced subtype steatofibrosis in a large population-based cohort of 1355 adults.

We have made several interesting observations. First, we confirmed that the gut microbiome of individuals with steatosis was less diverse than without steatosis. Second, the composition of the steatotic gut microbiome was significantly different from that of the controls. Thirty-seven genera contributed to this difference of which *Coprococcus3*, *Ruminococcus Gauvreauiigroup*, and *Ruminococcus Gnavusgroup* remained independently associated with steatosis after adjustment for confounders, such as BMI and diet, and multiple testing. Third, the composition of the gut microbiome in steatosis was predicted to contribute to secondary bile acid synthesis and biotin metabolism and to counteract D-alanine metabolism. Fourth, the metabolomic top-hits for steatosis included higher BCAAs, AAAs, glycoprotein acetyls and a detrimental lipid profile. Interestingly, incremental alpha-diversity was associated with lower BCAAs, glycoprotein acetyls and a favourable lipid profile, and hence had an opposite metabolic profile than steatosis.

Steatosis and obesity are tightly linked, especially in the context of NAFLD.[17] It has been proposed that a high ratio of Firmicutes-to-Bacteroidetes contributes to higher energy harvest in obese individuals.[431] In our study there was an eminent median relative abundance of *Firmicutes* in the total population, but no difference in FBR across steatosis or NAFLD strata. However, in line with our results, other reports could not always confirm this finding. Therefore it is unlikely that this ratio is solely responsible for obesogenic shifts in the microbiome.[432]

At genus level, lasso penalty selected 37 genera from four different phyla that were associated with steatosis. Three of these bacteria, all from the *Lachnospiraceae* family, remained significantly associated with steatosis. The anaerobic *Coprococcus3* was associated with lower steatosis prevalence, which is consistent with previous reports on the gut microbiome and hepatic steatosis.[409,415-417,433,434] Looking at the unadjusted relative abundance of *Coprococcus3*, this was markedly lower as well, particularly in NASF. Indeed, an elegant study of Hoyles et al. found that the genus Coprococcus was associated with less hepatocyte

ballooning as assessed by liver biopsy.[417] And although there was no significant association with the NMR metabolomics for this genus in our study, *Coprococcus3* was nominally significantly associated with lower glycoprotein acetyls. Indeed, the genus Coprococcus has previously been described to be lower in inflammatory bowel disease, confirming its anti-inflammatory profile.[435] Interestingly, other studies revealed that Coprococcus was associated with lower fasting glucose[433] and higher microbial richness.[436]

The two other significant genera *Ruminococcus Gauvreauiigrou*p and *Ruminococcus Gnavusgroup* are –although within the same family as *Coprococcus3*– associated with higher prevalence of steatosis. *R. Gauvreauiigroup* has not been described in relation to steatosis nor to other phenotypes before. It is known, however, that this genus is a strict anaerobe that produces mainly acetate as by-product of glucose fermentation.[437] *R. Gauvreauiigroup* was nominally significant associated with acetoacetate, but not with acetate. In contrast, *R. Gnavusgroup* has been described in relation with steatosis before.[417] A significant association of *R. Gnavusgroup* with acute phase reaction marker – glycoprotein acetyls. This is in line with the current knowledge that presence of *R. Gnavusgroup* has been previously described in individuals with lower microbial richness,[436] atherosclerotic cardiovascular disease,[438] and inflammatory bowel disease.[439]

Most of the other 34 lasso-selected genera have been previously described in relation to liver traits. For example, an interesting study of Boursier et al. found that Ruminococcus was positively and Prevotella was negatively associated with biopsy-proven fibrosis in NAFLD.[412] In our study, *Ruminococcus2* showed a trend towards a positive association with steatosis (*P*-value=0.016). Of note, there are many different Ruminococcus species which can express different functions. Another study showed a reduction of Prevotella after an animal-based diet and an increase after a high in fibre diet.[440] In our study, *Prevotella9* had a non-significant inverse association with steatosis. The previous study also showed that an animal-based diet was associated with a significant increase in gene expression among beta-lactamases. This is also in line with our study, in which we found an enriched beta-lactam resistance pathway (ko00312) in steatosis. This is of particular interest as our group recently showed that a diet high in animal protein was independently associated with liver steatosis.[344]

Other enriched predicted functional pathways in (non-alcoholic) steatosis included those involved in secondary bile acid biosynthesis (ko00121) and biotin metabolism (ko00780). Biotin is an essential cofactor of biotin-dependent carboxylases, such as acetyl-CoA carboxylase, which is implicated in the biosynthesis of fatty acids and hepatic steatosis.[441] Interestingly, a recent paper from Jiao et al. found that both absolute serum bile acids as well as percentage of secondary bile acids was elevated in the gut of non-alcoholic steatohepatitis patients.[442] However, in that study different bile acid-related pathway, 'taurine and hypo-taurine' (ko00430), was enriched. Nonetheless, a number of differences between their study and ours exist, such as sample size (n=27 vs n=1355), ethnicity (Asian vs European), and disease stage (non-alcoholic steatohepatitis vs NAFLD) that could have

contributed to the different findings and hampers direct comparison. Lastly, D-alanine metabolism (ko00473), a pathway that concerns the transformation from L-alanine to D-alanine, was predicted to be absent in steatosis and NAFLD. A lowered D-alanine metabolism and subsequent accumulation of L-alanine could lead to an increase in aspartate, glutamate and alanine – metabolites that were indeed significantly upregulated in steatosis and NAFLD in our study.

Several studies have already provided insights into the molecular signature of steatosis.[443] We found a particular increase in bacterial biosynthetic potential for BCAAs isoleucine and leucine, AAAs tyrosine and phenylalanine, and glycoprotein acetyls. Our results are in line with a large Finnish population-based study that found that circulating lipids, fatty acids, and amino acids –similar to our study– were associated with and preceded the development of hepatic steatosis in young adults.[443] Interestingly, a previous study shows that BCAAs tend to be higher in males than in females.[444] However, we have included sex as covariate in our multivariable model to control for potential confounding. Nevertheless, we cannot completely exclude the possibility of sex differences in the triad metabolomics, sex and steatosis. In addition, glycoprotein acetyls (mainly alpha 1), a composite marker of the most abundant acute phase proteins in the circulation, was positively associated with (non-alcoholic) steatosis and steatofibrosis in our study. Interestingly, this marker has been associated with various other inflammatory and metabolic comorbidities,[445] and with an overall reduced life expectancy.[446] Intriguingly, a large study of Wurtz et al. has shown that aberrations in BCAAs and AAAs precede insulin resistance –key in the development of steatosis– whereas glutamine levels were inversely related to insulin resistance and risk of type II diabetes.[445] Indeed, glutamine was inversely associated with steatofibrosis in our study. Intriguingly, alpha-diversity had an opposite metabolic profile compared to steatosis. Whether this may infer a role for gut microbial richness in the metabolic profile of steatosis is reserved for future longitudinal studies to discover.

The use of microbiomics, predicted metagenomics, and serum metabolomics combined with reliable liver imaging in the context of a large population-based study allowed us to perform a unique in-depth analysis on the gut-liver axis in (non-alcoholic) fatty liver disease. Nonetheless, the results of this study have to be interpreted in light of the following limitations. First, to optimize participation rate, we chose to collect stool samples via postal mail, but sample collection at room temperature and subsequent travel time can affect the composition and diversity of the samples. Second, we (as many others) examined the microbiome of faecal samples, which does not necessarily reflect the microbial variation throughout the whole gut. Also, we used 16S rRNA sequencing, which has the capacity to capture a broad shift in microbial diversity, but metagenomic approaches are in general more precise. Third, this study is cross-sectional by design and hence we cannot make inferences on the cause-effect relations. Fourth, assessing all bacteria in one regression model is challenging due to overfitting. Regularization methods, such as the use of Lasso

penalties are gaining popularity in the field of microbiomics, but as far as we know, we are the first study to use this method in context of the steatosis phenotype. However, overfitting might still have taken place in the phenotypes steatofibrosis and NASF because of the little number of cases. Fifth, replication in another large cohort with similar phenotype data does not exist (to our knowledge), but is needed to confirm our findings. Sixth, the linear regression analyses between the three genera and metabolites should be interpreted with caution because of the zero-inflated distribution (which was less prominent after log-transformation). Lastly, the gold standard for the diagnosis of steatosis and steatofibrosis is a liver biopsy. However, performing a liver biopsy in presumed healthy individuals is ethical debatable. We therefore used LSM, which has been shown to be a reliable non-invasive proxy for fibrosis. In addition, ultrasound is not a perfect measure for steatosis either, it has a low sensitivity for the presence of mild steatosis. However, we chose to dichotomize steatosis into present or absent instead of grading steatosis, because the sensitivity for moderate/severe steatosis is good. In addition, ultrasound is easy-to perform, widely available and inexpensive.

We hope future large-scale studies were to confirm our findings. Then, future studies on a more fundamental level over the course of time can test the causal inferences between the gut microbiome, circulating metabolites and steatosis. Also it would be interesting to further look at the differences between the various forms of steatosis: is the gut microbiome and its metabolites a common denominator in the pathophysiology behind steatosis? Ultimately, such information could help developing precision medicine in steatosis. In a large-scale setting the combination use of elastography and controlled attenuation parameter would be suitable to investigate steatosis and steatofibrosis non-invasively. In addition, it would be interesting if future studies could combine genetics, microbiomics and metabolomics. It is known that certain host genetic variants predispose an individual towards microbiome dysbiosis, which could also be an important factor in lipid metabolism and steatosis.[447] For example, it is known that polymorphisms PNPLA3-I148M and TM6SF2 are associated with an increased risk of NASF by means of interfering lipid pathways.[32-34] It would be interesting to see if associations still hold true when correcting for these kind of intrinsic covariates.

In conclusion, this work offers a comprehensive understanding of microbial diversity, composition, metabolomics and predicted metagenomics with steatosis, NAFLD, and its advanced subtype steatofibrosis in a community-dwelling population. We confirmed the lower microbial diversity and association of previously described 'beneficial' *Coprococcus* and 'harmful' *Ruminococcus Gnavus* with steatosis. Moreover, our study showed an exact opposite association between NMR metabolites and steatosis versus NMR metabolites and alpha-diversity, suggesting that they share –opposite– metabolic profiles. If future large-scale studies were to confirm our findings, our results could help developing precision medicine in steatosis.

Supplementary Files

Supplementary Methods

Metadata

An extensive home interview was carried out by professionally trained assistants, to obtain information on education level, smoking habits, and physical activity (using the LASA Physical Activity Questionnaire, expressed in metabolic equivalent of task (MET)hours/week).[299] Information on medication use (lipid lowering drugs, anti-diabetic medication, PPIs, amiodarone, systemic corticosteroids, methotrexate, and tamoxifen) was retrieved using automated linkage with the local pharmacy with which 98% of the participants was registered. Semi-quantitative (389-item) food frequency questionnaires were used to extract information on alcohol consumption, energy intake, and dietary quality score. This score contained the following components on intake of: I) vegetables (≥200g/day), II) fruit (≥200g/day), III) whole-grain products (≥90g/day), IV) legumes (≥135g/week), V) unsalted nuts (≥15g/day), VI) fish (≥100g/week), VII) dairy (≥350g/day), VIII) tea (≥150mL/day), IX) whole grains ≥50% of total grains, X) unsaturated fats and oils ≥ of total fats, XII) red and processed meat <300g/week, XIII) sugar-containing beverages (≤150mL/day), XIV) alcohol (≤10 g/day), and XV) salt (≤6 g/day). The food questionnaire was filled in 5.5 years prior to ultrasound and fecal evaluation, but a previous paper from our group showed that dietary consumption and specifically alcohol consumption was generally stable over time.[321] Excessive alcohol consumption was defined as ≥ 20 grams per day for women and ≥ 30 grams per day for men. Height (m) and weight (kg) were measured during the visit at the research centre, subsequent BMI was calculated (kg/m^2). Automatic enzyme procedures were used to measure blood lipids and glucose. Automatic immunoassays were used to measure insulin, hepatitis B surface antigen and anti-hepatitis C virus (Roche Diagnostic®, GmbH, Mannheim, DE). We calculated the homeostasis model assessment of insulin resistance (HOMA-IR), i.e. fasting glucose (mmol/dl) times fasting insulin (mU/L) divided by 22.5,[106] to proxy degree of insulin resistance in participants. Diabetes was diagnosed if the participant used anti-diabetic drugs and/or fasting glucose was 7.0 mmol/L or above. Hypertension was defined as the use of drugs for elevated blood pressure and/or elevated blood pressure, being either a systolic pressure of ≥ 140 mmHg or a diastolic pressure of ≥ 90 mmHg.

Statistical Analysis

Imputation analyses and participant characteristics

Missing values in the metadata (ranging from 0.00% to 19.26% per variable) were imputed using multiple imputation under the fully conditioned specification to reduce bias

due to missing data.[301] We created 50 imputed datasets using the R Package mice.[323] The results from these analyses were pooled using Rubin's rules in order to take into account the added uncertainty due to the missing of data.[254]

Description of the microbiome

The beta-diversity, which reflects the between-individual variability, was studied using principal coordinate analysis (PCoA) applied to Bray-Curtis dissimilarity using the function "capscale" from package Vegan[448] on genus level.

To additionally take into account the compositional nature of the microbiome we followed the approach of Lin et al.[428], which is implemented in the R package zeroSum.[449]

Predicted functional metagenomics

Predicted metagenomics pathways were tested for significance by performing logarithmic linear discriminant analysis (LDA) in linear discriminant analysis effect size with the significance level of 0.05 and the logarithmic LDA score threshold equal to 2.0.[450]

Statistical Significance

Adjusted *P*-values (referred to as *Q*-values) were calculated using the Benjamini & Hochberg correction[430] taking together all genera or metabolites within the same model, to account for the inflated type I error that arises due to multiple testing. All computations were done using R version 3.5.2 (R Core Team [2018]), the package Vegan version 2.5.2 [2018],[448] and the package ZeroSum version 1.1.1.[428,449]

Detailed information on the imputation process

	Multiple imputation
Software used	R version 3.5.1
Imputation method and key settings	Fully conditional specification (package mice version 3.1.0)*;* maximum iterations: 20
No. of imputed data sets created	50
Analyses variables	age; sex; steatosis; steatogenic medication; BMI; proton-pump inhibitors; time in mail; batch; 22 principle components of genera; smoking status; diabetes mellitus; high-density lipid cholesterol; glucose; energy intake; physical activity; waist-hip ratio; Dutch dietary guideline score; triglycerides; alcoholic units; education level
Auxiliary variables	Fibroscan probe; anti-diabetic drugs; liver stiffness (with IQR); alanine aminotransferase; total cholesterol; homeostasis model assessment of insulin resistance; systolic blood pressure; diastolic blood pressure; platelet count; spleen size; gamma-glutamyl transferase; insulin
Treatment of not normally distributed continuous variables	Predictive mean matching
Treatment of normally distributed variables	Linear regression
Treatment of binary/categorical variables	(proportional odds) Logistic regression
Population	For the imputation we used reliable FFQs. We included only participants with ultrasound data. Missing FFQs-related covariates (e.g. diet quality, energy intake) were imputed.

Supplementary Table 1: Imputed and Original characteristics

	Original data	Imputed data
Age (years)	62.36 (5.86)	*no missing*
Female (%)	57.8	*no missing*
European (%)	96.7	96.7
Education Level (%)		
Low	41.2	41.2
Intermediate	28.2	28.2
High	30.6	30.6
Smoking status (%)		
Never	35.4	35.0
Past or Current	64.6	65.0
Alcohol (units/d)	0.83 [0.17, 2.05]	0.83 [0.17, 2.05]
Physical Activity (METeq/wk)	49.00 [21.00, 82.73]	49.00 [21.00, 82.78]
Energy intake (kcal/day)	2254 [1887, 2734]	2251 [1882, 2734]
Dietary Quality Score (0–15)	7.00 [6.00, 8.00]	7.00 [6.00, 8.00]
BMI (kg/m²)	27.42 (4.49)	*no missing*
Waist-to-hip ratio	0.89 [0.82, 0.96]	0.89 [0.82, 0.96]
HOMA-IR	2.56 [1.70, 4.08]	2.56 [1.70, 4.08]
Total Cholesterol (mmol/L)	5.58 (1.11)	5.58 (1.11)
HDL-Cholesterol	1.51 (0.45)	1.51 (0.45)
Triglycerides	1.29 [0.96, 1.74]	1.30 [0.96, 1.74]
Diabetes Mellitus (%)	10.0	10.0
Hypertension (%)	58.3	58.3
Liver stiffness measurements (kPa)*	4.55 [3.70, 5.60]	4.55 [3.70, 5.60]
Lipid-lowering drug use	27.2	*no missing*
Proton-pump inhibitor use	17.7	*no missing*
Microbiome batch 0/1	78.8 / 21.2	*no missing*
Time in mail (%)		*no missing*
1 day	65.3	
2 days	29.2	
3 days	5.5	
Steatosis n (%)	472 (34.8)	*no missing*
Viral hepatitis†	8 (0.6)	*not imputed*
Excessive alcohol use‡	202 (18.5)	*not imputed*
Steatogenic drug use	21 (1.5)	*not imputed*

*Liver stiffness not yet filtered for reliability here. †Data available for 1338 participants. ‡Data available for 1094 participants.
Abbreviations: HOMA-IR: homeostasis model of insulin resistance; kcal: kilocalories; kPa: kilopascals; LSM: liver stiffness measurement; METeqh/wk: metabolic equivalent task hours per week; NAFLD: non-alcoholic fatty liver disease

Supplementary Table 2: Relative abundance of phyla between steatosis strata

	Total population (n=1355)	No steatosis (n=883)	Steatosis (n=472)	*P*-value*
Verrucomicrobia	$3.1{\cdot}10^{-2}$ (0 – 0.21)	$3.4{\cdot}10^{-2}$ (0 – 0.23)	$2.1{\cdot}10^{-2}$ (0 – 0.16)	0.040
Tenericutes	0 (0 – $8.7{\cdot}10^{-2}$)	$1.1{\cdot}10^{-2}$ (0 – 0.10)	0 (0 – $4.6{\cdot}10^{-2}$)	<0.001
Synergistetes	0 (0 – 0)	0 (0 – 0)	0 (0 – 0)	0.982
Saccharibacteria	0 (0 – $1.1{\cdot}10^{-2}$)	0 (0 – $1.1{\cdot}10^{-2}$)	0 (0 – $1.1{\cdot}10^{-2}$)	0.148
Proteobacteria	1.4 (0.33 – 5.3)	1.4 (0.37 – 5.8)	1.2 (0.29 – 4.4)	0.035
Fusobacteria	0 (0 – 0)	0 (0 – 0)	0 (0 – 0)	0.239
Firmicutes	81.2 (69.3 – 88.6)	81.2 (69.6 – 88.4)	81.1 (69.1 – 88.8)	0.693
Euryarchaeota	0 (0 – 0)	0 (0 – $1.0{\cdot}10^{-2}$)	0 (0 – 0)	0.008
Cyanobacteria	0 (0 – 0)	0 (0 – 0)	0 (0 – 0)	0.227
Bacteroidetes	8.6 (3.1 – 17.9)	8.8 (3.4 – 17.7)	8.0 (2.8 – 18.0)	0.323
Actinobacteria	2.7 (1.2 – 5.3)	2.8 (1.2 – 5.1)	2.6 (1.2 – 5.5)	0.998

Relative abundance (%) median (P25-P75) are given. * *P*-value based on Wilcoxon Rank Sum test comparing steatosis with no steatosis per phylum.

Supplementary Table 3: Relative abundance of selected genera between steatosis strata

Phylum *Class* *Order* *Family* *Genus (median % (P25-P75))*	No steatosis (n=883)	Steatosis (n=472)
Actinobacteria		
Actinobacteria *Bifidobacteriales* *Bifidobacteriaceae* *Bifidobacterium*	1.36 (0.36 – 3.40)	1.32 (0.27 – 3.72)
Coriobacteriia *Coriobacteriales* *Coriobacteriaceae* *Collinsella*	0.09 (0.01 – 0.39)	0.12 (0.01 – 0.54)
Bacteroidetes		
Bacteroidia *Bacteroidales* *Bacteroidaceae* *Bacteroides*	3.97 (1.51 – 9.48)	3.96 (1.18 – 8.63)
Bacteroidia *Bacteroidales* *BacteroidalesS247group* *unknowngenus*	0 (0 – 0.13)	0 (0 – 0.05)
Bacteroidia *Bacteroidales* *Prevotellaceae* *Prevotella9*	0 (0 – 0.75)	0 (0 – 0.72)
Bacteroidia *Bacteroidales* *Rikenellaceae* *Alistipes*	0.43 (0.14 – 1.09)	0.33 (0.07 – 1.13)

Supplementary Table 3 (*continued*)

Firmicutes		
Bacilli *Lactobacillales* *Lactobacillaceae* *Lactobacillus*	0.01 (0 – 0.09)	0.01 (0 – 1.01)
Bacilli *Lactobacillales* *Streptococcaceae* *Streptococcus*	0.40 (0.12 – 1.47)	0.46 (0.14 – 1.98)
Clostridia *Clostridiales* *Christensenellaceae* *ChristensenellaceaeR7group*	1.58 (0.55 – 3.85)	0.81 (0.19 – 2.29)
Clostridia *Clostridiales* *Clostridiaceae1* *Clostridiumsensustricto1*	0.35 (0.04 – 1.38)	0.18 (0.02 – 1.05)
Clostridia *Clostridiales* *Lachnospiraceae* *Blautia*	8.40 (4.77 – 14.27)	10.0 (5.52 – 16.64)
Firmicutes		
Clostridia Clostridiales Lachnospiraceae Butyrivibrio	0 (0 – 0.03)	0 (0 – 0)
Clostridia Clostridiales Lachnospiraceae Coproccocus3	0.60 (0.27 – 1.08)	0.63 (0.23 – 1.11)
Clostridia Clostridiales Lachnospiraceae Dorea	1.61 (0.94 – 2.74)	1.93 (0.98 – 3.20)
Clostridia Clostridiales Lachnospiraceae Fusicatenibacter	1.68 (0.77 – 3.02)	1.70 (0.69 – 3.22)
Clostridia Clostridiales Lachnospiraceae Lachnoclostridium	0.55 (0.32 – 0.99)	0.70 (0.34 – 1.37)
Clostridia Clostridiales Lachnospiraceae LachnospiraceaeND3007group	0.42 (0.13 – 0.82)	0.33 (0.12 – 0.61)
Clostridia Clostridiales Lachnospiraceae Eligensgroup	0.26 (0.05 – 0.82)	0.15 (0.01 – 0.56)
Clostridia Clostridiales Lachnospiraceae R. Gauvreauiigroup	0.36 (0.07 – 0.81)	0.42 (0.08 – 0.96)

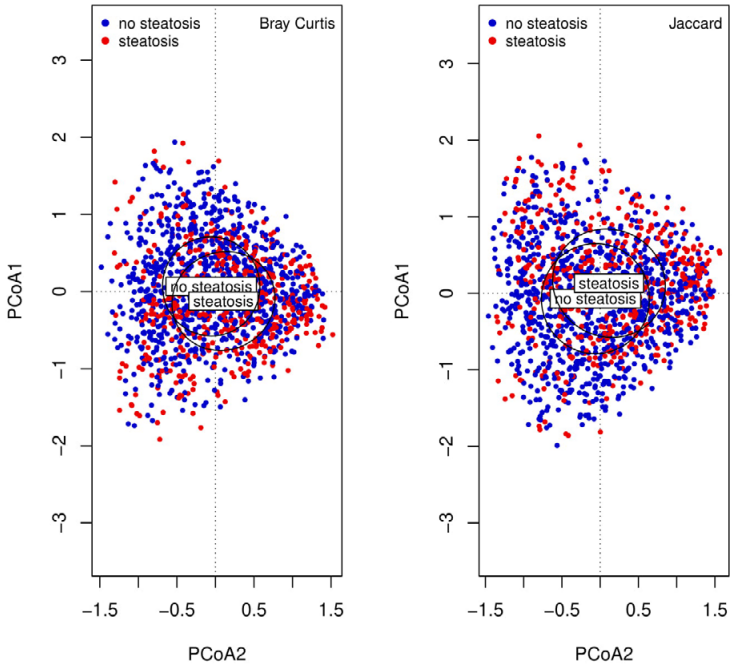
Supplementary Table 3 (*continued*)

Clostridia Clostridiales Lachnospiraceae R. Gnavusgroup	0 (0 – 0.04)	0 (0 – 0.08)
Clostridia Clostridiales Lachnospiraceae Halliigroup	2.26 (1.21 – 4.18)	2.84 (1.36 – 5.11)
Clostridia Clostridiales Lachnospiraceae Rectalegroup	3.08 (1.29 – 6.25)	4.10 (1.81 – 7.95)
Clostridia Clostridiales Peptostreptococcaceae Romboutsia	4.71 (0.08 – 1.46)	3.33 (0.03 – 1.17)
Clostridia Clostridiales Ruminococcaceae RuminococcaceaeNK4A214group	0.42 (0.14 – 0.91)	0.26 (0.05 – 0.68)
Firmicutes		
Clostridia *Clostridiales* *Ruminococcaceae* *RuminococcaceaeUCG002*	1.48 (0.58 – 3.20)	0.93 (0.29 – 2.20)
Clostridia *Clostridiales* *Ruminococcaceae* *RuminococcaceaeUCG014*	2.08 (0.33 – 4.84)	0.83 (0.11 – 3.25)
Clostridia *Clostridiales* *Ruminococcaceae* *Ruminococcus1*	0.65 (0.22 – 1.53)	0.51 (0.09 – 1.27)
Clostridia *Clostridiales* *Ruminococcaceae* *Ruminococcus2*	1.76 (0.60 – 3.25)	1.51 (0.35 – 3.78)
Clostridia *Clostridiales* *Ruminococcaceae* *Subdoligranulum*	2.59 (1.43 – 4.47)	2.35 (1.23 – 4.43)
Clostridia *Clostridiales* *Ruminococcaceae* *Coprostanoligenesgroup*	1.61 (0.95 – 2.75)	1.51 (0.78 – 2.61)
Erysipelotrichia *Erysipelotrichales* *Erysipelotrichaceae* *Catenibacterium*	0 (0 – 0)	0 (0 – 0)
Negativicutes *Selenomonadales* *Acidaminococcaceae* *Phascolarctobacterium*	0.11 (0 – 0.74)	0.06 (0 – 0.65)

Supplementary Table 3 (*continued*)

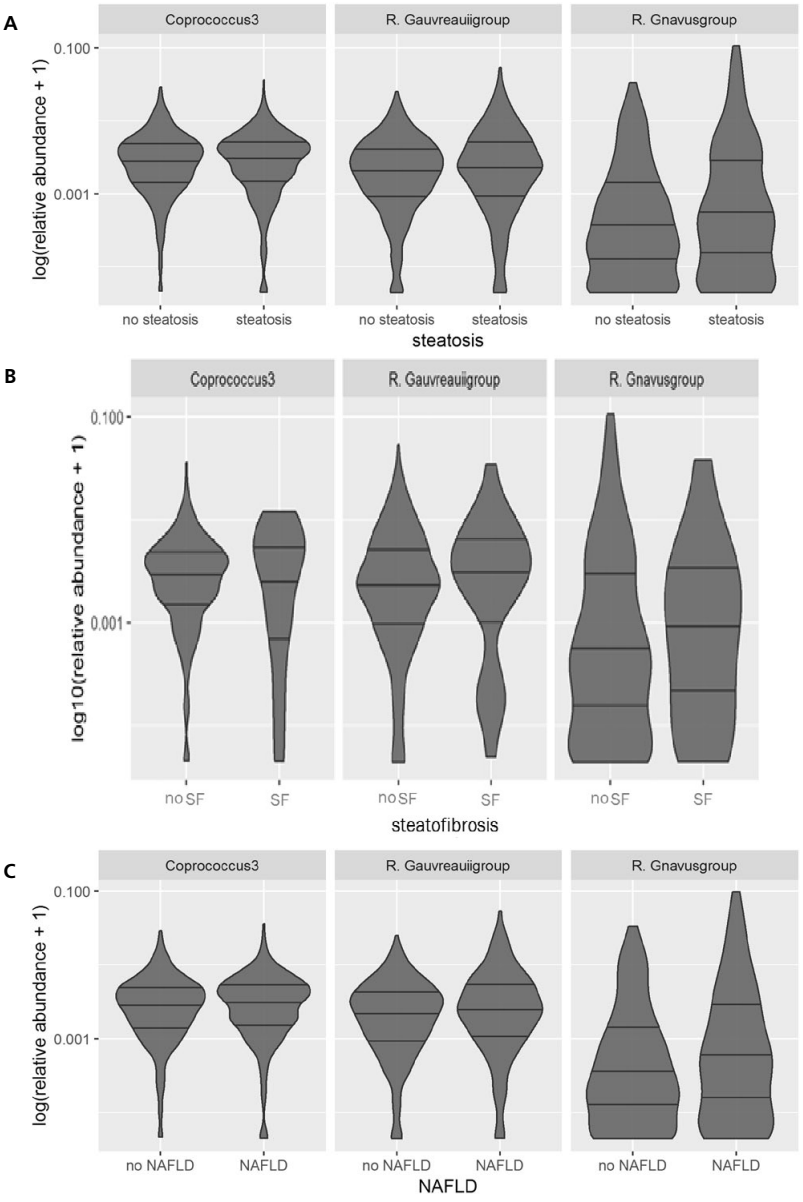
Negativicutes *Selenomonadales* *Veillonelaceae* *Megamonas*	0 (0 – 0)	0 (0 – 0)
Proteobacteria		
Gammaproteobacteria *Aeromonadales* *Succinivibrionaceae* *Succinivibrio*	0 (0 – 0)	0 (0 – 0)
Gammaproteobacteria *Enterobacteriales* *Enterobacteriaceae* *EscherichiaShigella*	0.45 (0.04 – 4.23)	0.28 (0.03 – 2.56)
Gammaproteobacteria *Enterobacteriales* *Enterobacteriaceae* *Klebsiella*	0 (0 – 0)	0 (0 – 0)

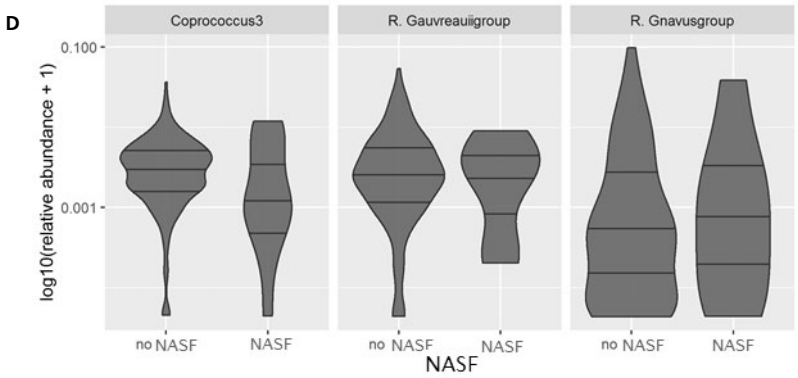
Relative abundance (%) median (P25-P75) are given for the group with steatosis and without steatosis.



Supplementary Figure 1: Principle Coordinate Analysis Plots

The Principle Coordinate Analyses using Bray-Curtis dissimilarity distances (quantitative mode on the left) and Jaccard similarity distances (binary mode on the ride). The blue dots reflect cases without steatosis and the red dots those with steatosis. Both plots show the group centroids (black circles). For the statistical analyses we used PerMANOVA adjusted for the covariates age, sex, BMI and technical covariates. PerMANOVA on both distances showed significant variation between the two groups (R2=0.003; p<0.001 for Bray-Curtis and R2=0.002; p<0.001 for Jaccard).

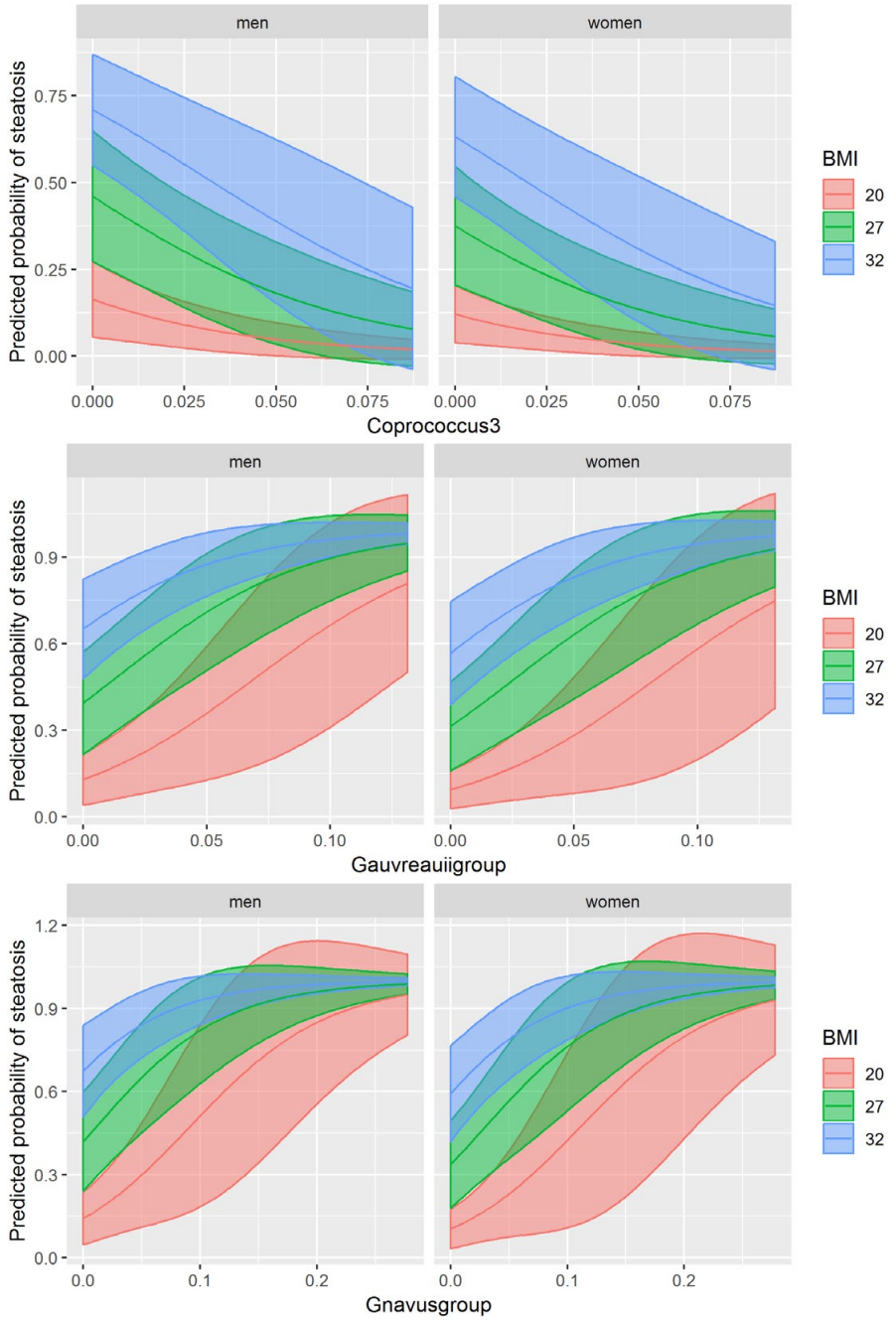

A
Coprococcus3
R. Gauvreauiigroup
R. Gnavusgroup
log(relative abundance + 1)
0.100
0.001
no steatosis
steatosis
steatosis
B
Coprococcus3
R. Gauvreauiigroup
R. Gnavusgroup
log10(relative abundance + 1)
0.100
0.001
no SF
SF
steatofibrosis
C
Coprococcus3
R. Gauvreauiigroup
R. Gnavusgroup
log(relative abundance + 1)
0.100
0.001
no NAFLD
NAFLD
NAFLD



Supplementary Figure 2: Plotted relative abundance for steatosis-associated genera across (A) steatosis strata, (B) steatofibrosis strata, (C) NAFLD strata, and (D) NASF strata

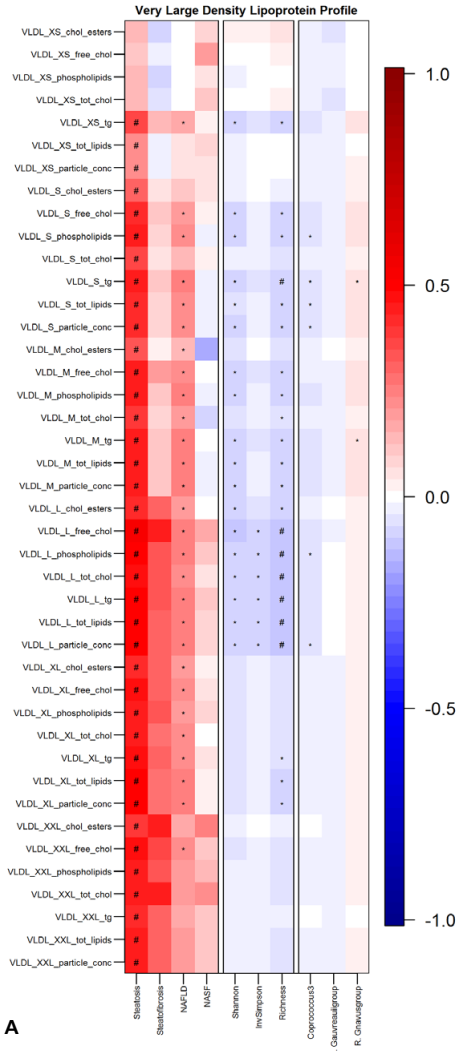
The log10 transformed non-zero values of the significant lasso-selected genera are depicted across A) steatohepatitis, B) NAFLD, and C) NASH strata using violin plots. The log-transformed non-zero values of the significant lasso-selected genera are depicted across steatosis strata using violin plots.

Abbreviations: NAFLD: non-alcoholic fatty liver disease; NASH: non-alcoholic steatohepatitis.

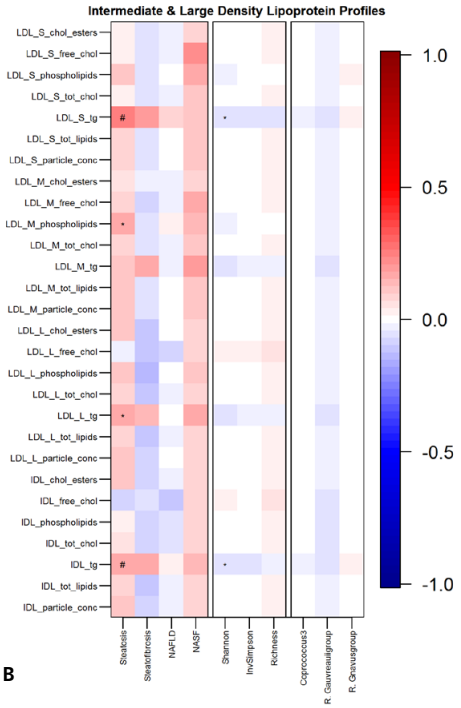


Supplementary Figure 3: Predicted probability of steatosis for relative abundance of Coprococcus3, R Gauvreauiigroup, and R Gnavusgroup

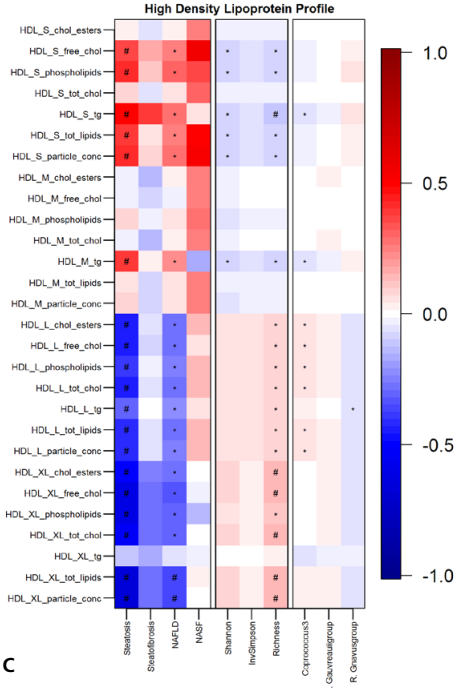
A-C: Predicted probability for exemplary patients plotted by sex and BMI (two significant variables), based on the multivariable model as shown in table 2. On the y-axis: the predicted probability (in %) for steatosis with 95%CI. On the x-axis: the relative abundance of the genus. Abbreviations: BMI: body mass index.


Very Large Density Lipoprotein Profile
VLDL_XS_chol_esters
VLDL_XS_free_chol
VLDL_XS_phospholipids
VLDL_XS_tot_chol
VLDL_XS_tg
VLDL_XS_tot_lipids
VLDL_XS_particle_conc
VLDL_S_chol_esters
VLDL_S_free_chol
VLDL_S_phospholipids
VLDL_S_tot_chol
VLDL_S_tg
VLDL_S_tot_lipids
VLDL_S_particle_conc
VLDL_M_chol_esters
VLDL_M_free_chol
VLDL_M_phospholipids
VLDL_M_tot_chol
VLDL_M_tg
VLDL_M_tot_lipids
VLDL_M_particle_conc
VLDL_L_chol_esters
VLDL_L_free_chol
VLDL_L_phospholipids
VLDL_L_tot_chol
VLDL_L_tg
VLDL_L_tot_lipids
VLDL_L_particle_conc
VLDL_XL_chol_esters
VLDL_XL_free_chol
VLDL_XL_phospholipids
VLDL_XL_tot_chol
VLDL_XL_tg
VLDL_XL_tot_lipids
VLDL_XL_particle_conc
VLDL_XXL_chol_esters
VLDL_XXL_free_chol
VLDL_XXL_phospholipids
VLDL_XXL_tot_chol
VLDL_XXL_tg
VLDL_XXL_tot_lipids
VLDL_XXL_particle_conc
Steatosis
Steatofibrosis
NAFLD
NASF
Shannon
InvSimpson
Richness
Coprococcus3
R. Gauvreauiigroup
R. Gnavusgroup
1.0
0.5
0.0
-0.5
-1.0

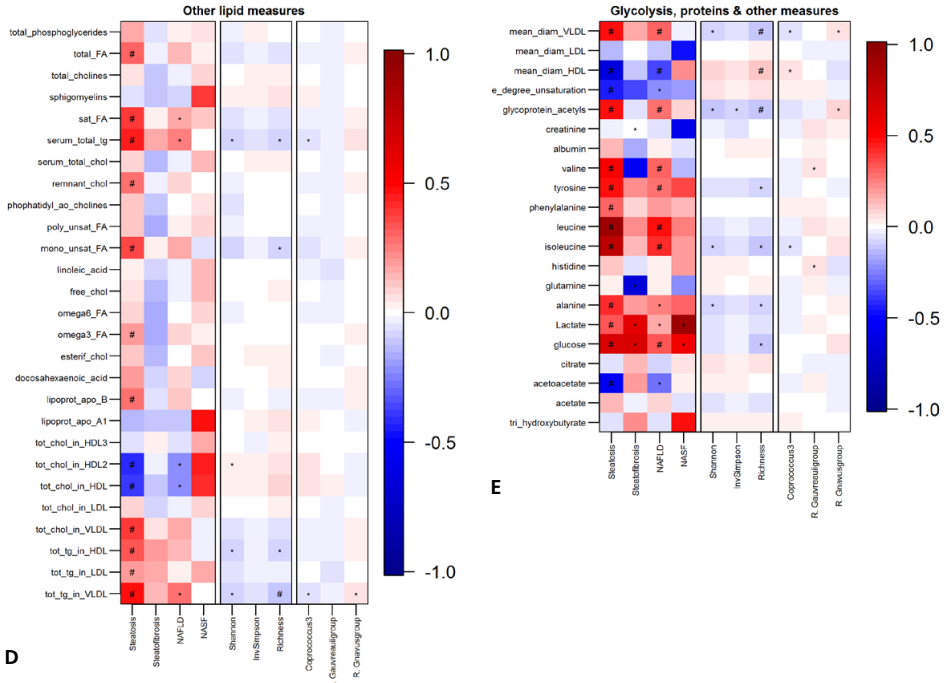
A


Intermediate & Large Density Lipoprotein Profiles
LDL_S_chol_esters
LDL_S_free_chol
LDL_S_phospholipids
LDL_S_tot_chol
LDL_S_tg
LDL_S_tot_lipids
LDL_S_particle_conc
LDL_M_chol_esters
LDL_M_free_chol
LDL_M_phospholipids
LDL_M_tot_chol
LDL_M_tg
LDL_M_tot_lipids
LDL_M_particle_conc
LDL_L_chol_esters
LDL_L_free_chol
LDL_L_phospholipids
LDL_L_tot_chol
LDL_L_tg
LDL_L_tot_lipids
LDL_L_particle_conc
IDL_chol_esters
IDL_free_chol
IDL_phospholipids
IDL_tot_chol
IDL_tg
IDL_tot_lipids
IDL_particle_conc
Steatosis
Steatofibrosis
NAFLD
NASF
Shannon
InvSimpson
Richness
Coprococcus3
R. Gauvreauiigroup
R. Gnavusgroup
1.0
0.5
0.0
-0.5
-1.0

B


High Density Lipoprotein Profile
HDL_S_chol_esters
HDL_S_free_chol
HDL_S_phospholipids
HDL_S_tot_chol
HDL_S_tg
HDL_S_tot_lipids
HDL_S_particle_conc
HDL_M_chol_esters
HDL_M_free_chol
HDL_M_phospholipids
HDL_M_tot_chol
HDL_M_tg
HDL_M_tot_lipids
HDL_M_particle_conc
HDL_L_chol_esters
HDL_L_free_chol
HDL_L_phospholipids
HDL_L_tot_chol
HDL_L_tg
HDL_L_tot_lipids
HDL_L_particle_conc
HDL_XL_chol_esters
HDL_XL_free_chol
HDL_XL_phospholipids
HDL_XL_tot_chol
HDL_XL_tg
HDL_XL_tot_lipids
HDL_XL_particle_conc
Steatosis
Steatofibrosis
NAFLD
NASF
Shannon
InvSimpson
Richness
Coprococcus3
R. Gauvreauiigroup
R. Gnavusgroup
1.0
0.5
0.0
-0.5
-1.0

C



Supplementary Figure 4: Metabolic profiles of steatosis, alpha-diversity and lasso-selected genera after exclusion of participants that use lipid-lowering medication (n=986)

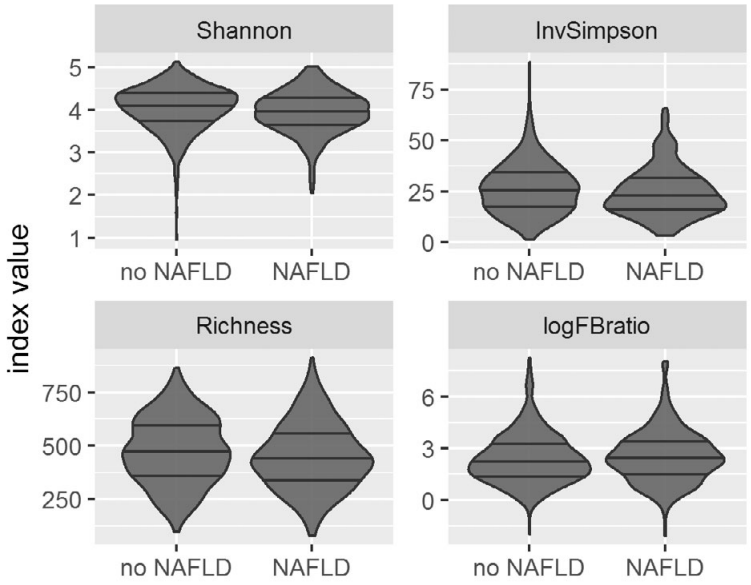
The colors in columns represent the standardized effect estimates (betas) of the metabolites with steatosis, steatohepatitis, NAFLD, NASH, alpha-diversity and significant genera. The betas were adjusted for covariates in model 1 (sex†‡, age†‡, technical variables†‡, BMI†‡, Dietary Quality-Score, PPI use, alcohol units†). We corrected for multiple testing using Benjamini & Hochberg, significant Q-values are depicted by #, nominally significant values by *. †Steatohepatitis adjustment, ‡NASH adjustment.

Abbreviations: BMI: body mass index; chol: cholesterol; conc: concentration; diam: diameter; FA: fatty acids; HDL: high-density lipoprotein; HOMA-IR: homeostasis model of insulin resistance; IDL: intermediate density lipoprotein; InvSimpson: Inversed Simpson; L: large; LDL: low-density lipoprotein; M: medium; PPI: proton-pump inhibitor; S: small; sat: saturated; tg: triglycerides; tot: total; VLDL: very low density lipoprotein; (X)XL: (extra) extra large

Supplementary Table 4: Relative abundance of phyla between NAFLD strata

	No NAFLD (n=574)	NAFLD (n=283)	*P*-value*
Verrucomicrobia	$3.2{\cdot}10^{-2}$ (0 – 0.24)	$2.3{\cdot}10^{-2}$ (0 – 0.18)	0.637
Tenericutes	$1.0{\cdot}10^{-2}$ (0 – 0.10)	0 (0 – $4.3{\cdot}10^{-2}$)	<0.001
Synergistetes	0 (0 – 0)	0 (0 – 0)	0.584
Saccharibacteria	0 (0 – $1.1{\cdot}10^{-2}$)	0 (0 – $1.1{\cdot}10^{-2}$)	0.232
Proteobacteria	1.5 (0.34 – 6.1)	1.1 (0.29 – 3.5)	0.018
Fusobacteria	0 (0 – 0)	0 (0 – 0)	0.121
Firmicutes	81.2 (70.2 – 88.3)	82.3 (72.1 – 88.4)	0.274
Euryarchaeota	0 (0 – $7.6{\cdot}10^{-3}$)	0 (0 – 0)	0.146
Cyanobacteria	0 (0 – 0)	0 (0 – 0)	0.581
Bacteroidetes	8.6 (3.3 – 17.5)	7.5 (2.9 – 16.2)	0.150
Actinobacteria	2.9 (1.2 – 5.3)	2.8 (1.2 – 6.1)	0.796

Relative abundance (%) median (P25-P75) are given. * *P*-value based on Wilcoxon Rank Sum test comparing steatosis with no steatosis per phylum.
Abbreviations: NAFLD: non-alcoholic fatty liver disease.

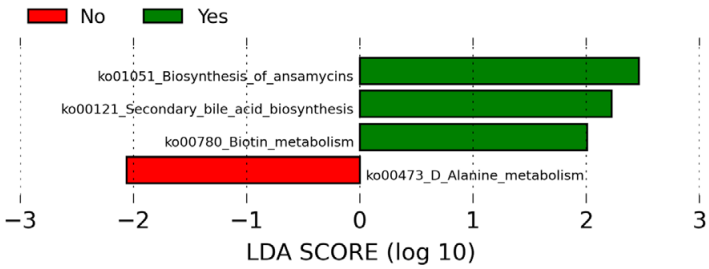


Supplementary Figure 5: Alpha-diversity and Firmicutes-to-Bacteroidetes-ratio in NAFLD
Violin plots of the distribution of A-B) alpha-diversity as assessed by the Shannon-index and Inverse Simpson, C) microbial richness, and D) log-transformed FBR across steatosis strata. The statistical difference between strata was assessed using Wilcoxon rank sum test with continuity correction and was as follows (Shannon: P= 0.003; Inverse Simpson: P= 0.021; Richness: P=0.015 and; FBR: P=0.876).
Abbreviations: InvSimpson: Inversed Simpson; logFBratio: log-transformed Firmicutes-to-Bacteroidetes-ratio.

Supplementary Table 5: Multivariable logistic regression analysis of genera that were associated with steatosis (n=1355) and are now tested against steatosis and NAFLD as dependent variable (n= 857).

	Steatosis 472 / 1355		NAFLD 283 / 857	
	β	95%CI	β	95%CI
Coprococcus3	−51.76	−89.98, −13.54	−43.85	−90.29, 2.58
Gauvreauiigroup	64.29	31.03, 97.56	61.71	22.17, 101.25
Gnavusgroup	52.12	24.22, 80.02	50.26	18.51, 82.0

Logistic regression analyses with steatosis or NAFLD as dependent variable, both adjusted for age, sex, BMI, alcohol, PPI-use, dietary quality and technical variables (batch and time in mail). All genera (log10(1+genus)) are analyzed simultaneously. Results are given as regression coefficient (95%CI).
Abbreviations: BMI: body mass index; PPI: proton-pump inhibitor; NAFLD: non-alcoholic fatty liver disease.



Supplementary Figure 6: Predicted metagenomics pathways for NAFLD (n=857)
Predicted functional composition of metagenomes based on 16S rRNA gene sequencing data in our study cohort. Linear discriminant effect size-based on the PICRUSt dataset revealed six significant (FDR<0.05) metabolic pathways that were predicted to be present in NAFLD (Yes in green) and that were predicted to be absent in NAFLD (No in red).