Links between gut microbiome composition and fatty liver disease in a large population sample

Matti O. Ruuskanen[1,2]*, Fredrik Åberg[3,4], Ville Männistö[5,6], Aki S. Havulinna[2,7], Guillaume Méric[8,9], Yang Liu[8,10], Rohit Loomba[11,12], Yoshiki Vázquez-Baeza[13,14], Anupriya Tripathi[15,16,17], Liisa M. Valsta[2], Michael Inouye[8,18], Pekka Jousilahti[2], Veikko Salomaa[2], Mohit Jain[12,19], Rob Knight[13,14,20,21], Leo Lahti[22], Teemu J. Niiranen[1,2,23]

[1]Department of Internal Medicine, University of Turku, Turku, Finland
[2]Department of Public Health Solutions, Finnish Institute for Health and Welfare, Helsinki, Finland
[3]Transplantation and Liver Surgery Clinic, Helsinki University Hospital, University of Helsinki, Helsinki, Finland
[4]The Transplant Institute, Sahlgrenska University Hospital, Gothenburg, Sweden
[5]Department of Medicine, Kuopio University Hospital, University of Eastern Finland, Kuopio, Finland
[6]Department of Experimental Vascular Medicine, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands
[7]Institute for Molecular Medicine Finland, FIMM - HiLIFE, Helsinki, Finland
[8]Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia
[9]Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia
[10]Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia
[11]Department of Medicine, NAFLD Research Center, La Jolla, CA, USA
[12]Department of Medicine, University of California, San Diego, La Jolla, CA, USA
[13]Jacobs School of Engineering, University of California, San Diego, La Jolla, CA, USA
[14]Center for Microbiome Innovation, University of California San Diego, La Jolla, California, USA
[15]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California, USA
[16]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California, USA
[17]Division of Biological Sciences, University of California, San Diego, La Jolla, California, USA
[18]Department of Public Health and Primary Care, Cambridge University, Cambridge, United Kingdom
[19]Department of Pharmacology, University of California San Diego, La Jolla, California, USA
[20]Department of Pediatrics, School of Medicine, University of California San Diego, La Jolla, California, USA
[21]Department of Computer Science & Engineering, University of California San Diego, La Jolla, California, USA
[22]Department of Future Technologies, University of Turku, Turku, Finland
[23]Division of Medicine, Turku University Hospital, Turku, Finland

*Correspondence: Matti Ruuskanen, matti.ruuskanen@utu.fi

Running head: Gut microbiome composition and fatty liver

Abstract

Background. Fatty liver disease is the most common liver disease in the world. It is characterized by a build-up of excess fat in the liver that can lead to cirrhosis and liver failure. The link between fatty liver disease and gut microbiome has been known for at least 80 years. However, this association remains mostly unstudied in the general population because of underdiagnosis and small sample sizes. To address this knowledge gap, we studied the link between the Fatty Liver Index (FLI), a well-established proxy for fatty liver disease, and gut microbiome composition in a representative, ethnically homogeneous population sample in Finland. We based our models on biometric covariates and phylogenetically transformed gut microbiome compositions from shallow metagenome sequencing.

Results. Our classification models were able to discriminate between individuals with a high FLI (≥ 60, indicates likely liver steatosis) and low FLI (< 60) in our validation set, consisting of 30% of the data not used in model training, with an average AUC of 0.75. In addition to age and sex, our models included differences in 11 microbial groups from class *Clostridia*, mostly belonging to orders *Lachnospirales* and *Oscillospirales*. Pathway analysis of representative genomes of the FLI-associated taxa in (NCBI) *Clostridium* subclusters IV and XIVa indicated the presence of *e.g.*, ethanol fermentation pathways.

Conclusions. Through modeling the fatty liver index, our results provide with high resolution associations between gut microbiota composition and fatty liver in a large representative population cohort. Our results lend further support to the role of endogenous ethanol producers in the development of fatty liver.

Keywords: Metagenomics, human gut, fatty liver, fatty liver index, population sample

Background

Fatty liver disease affects roughly a quarter of the world's population (Younossi et al., 2016). It is characterized by accumulation of fat in the liver cells and is intimately linked with pathophysiology of metabolic syndrome (Marchesini et al., 2003; Chalasani et al., 2012; Yki-Järvinen, 2014). Fatty liver disease can be broadly divided into two variants: non-alcoholic fatty liver disease (NAFLD), attributed to high caloric intake, and alcohol associated fatty liver disease, attributed to high alcohol consumption. Even though the rate of progressions and underlying causes of both diseases might be different, they can be broadly sub-divided into those who have fat accumulation in the liver with no or minimal inflammation or those who have additional features of cellular injury and active inflammation with or without fibrosis typically seen in peri-sinusoidal area. (Toshikuni et al., 2014). Patients with steatohepatitis may progress to cirrhosis and hepatocellular carcinoma and have increased risk of liver-related morbidity and mortality, globally amounting to hundreds of thousands of deaths (Rinella and Charlton, 2016).

The human gut harbors up to 10^{12} microbes per gram of content (Gilbert et al., 2018) and is intimately connected with the liver. Thus, it is no surprise that gut microbiome composition appears to have a strong connection with liver disease (Caussy et al., 2019). Numerous studies over the past 80 years have reported associations between gut microbial composition and liver disease (Compare et al., 2012). For example, gut permeability and overgrowth of bacteria in the small intestine (Miele et al., 2009), changes in *Gammaproteobacteria* and *Erysipelotrichi* abundance during choline deficiency (Spencer et al., 2011), elevated abundance of ethanol-producing bacteria (Zhu et al., 2013; Yuan et al., 2019), metagenomic signatures of specific bacterial species (Loomba et al., 2017; Jiao et al., 2019) have all been linked to NAFLD in small case-control patient samples. However, the microbial signatures often overlap between NAFLD and metabolic diseases, while those of more serious liver disease such as

steatohepatitis and cirrhosis are more clear (Aron-Wisnewsky et al., 2020). For example, oral taxa appear to invade the gut in liver cirrhosis (Qin et al., 2014), and this phenotype can accurately be detected by analyzing the fecal microbiome composition (AUC = 0.87 in a validation cohort; Caussy et al., 2019). Furthermore, we recently demonstrated good prediction accuracy for incident liver disease diagnoses (AUC = 0.83 for non-alcoholic liver disease, AUC = 0.96 for alcoholic liver disease, during ~15 years), showing that the signatures of serious future liver disease are easy to detect (Liu et al., 2020).

The mechanisms underlying the contribution of gut microbiome content with fatty liver disease are thought to be primarily linked to gut bacterial metabolism. Bacterial metabolites can indeed be translocated from the gut through the intestinal barrier into the portal vein and transported to the liver, where they interact with liver cells, and can lead to inflammation and steatosis (Safari and Gérard, 2019). Short-chain fatty acid production, conversion of choline into methylamines, modification of bile acids (BA) into secondary BA, and ethanol production, all of which are mediated by gut bacteria, are also known to be aggravating factors for NAFLD (Safari and Gérard, 2019). Recent studies have also suggested that endogenous ethanol production by gut bacteria could lead to an increase in gut membrane permeability (Yuan et al., 2019). This can facilitate the translocation of bacterial metabolites and cell components such as lipopolysaccharides from the gut to the liver, leading to further inflammation and possible development of NAFLD (Carpino et al., 2019).

Liver biopsy assessment is the current gold standard for diagnosis of fatty liver disease and its severity (Li et al., 2018), but it is also impractical and unethical in a population-based setting. Ultrasound and MRI based assessment can help detect presence of fatty liver, however, this data is not available in our cohort. Regardless, recent studies have shown that indices based on anthropometric measurements and

standard blood tests can be a reliable tool for non-invasive diagnosis of fatty liver particularly in population-based epidemiologic studies (Koehler et al., 2013; Vanni and Bugianesi, 2015).

Here, we designed and conducted computational analyses to examine the links between fatty liver and gut microbiome composition in a representative population sample of 7211 extensively phenotyped Finnish individuals (Salosensaari et al., 2020). Because fatty liver disease is generally underdiagnosed in the general population (Alexander et al., 2018), we used population-wide measurements of BMI, waist circumference, blood triglycerides and gamma-glutamyl-transferase (GGT) to calculate a previously validated Fatty Liver Index (FLI) for each participant as a proxy for fatty liver (Bedogni et al., 2006). In parallel, we used shallow shotgun sequencing to analyze gut microbiome composition (Hillmann et al., 2018), which also enabled the use of phylogenetic and pathway prediction methods. In this work, we describe high-resolution associations between fatty liver and individual gut microbial taxa and clades, which are generalizable at the population level.

Results

Bacterial community structure is correlated with Fatty Liver Index in a population sample

To investigate the link between fatty liver disease (using FLI as a proxy; **Figure 1A, 1B**) and gut microbial composition, we used linear regression (adjusted $R^2 = 0.29$) on the three first principal component (PC) axes of the fecal bacterial beta-diversity (between individuals), sex, age, and alcohol to model FLI. $\log_{10}$(FLI) significantly correlated with all three bacterial PC axes, sex, age, and alcohol use (all $P < 1\times10^{-6}$). Correlations between FLI and archaeal PC axes were not significant ($P > 0.05$). The effect size estimate on $\log_{10}$(FLI) was a magnitude larger for PC1 (0.11 ± 0.008) than for PC2 (0.04 ± 0.008) and PC3 (-0.06 ± 0.008). The relationships between FLI and the bacterial PC components representing their beta-diversity are visualized for each of the three components in **Figure**

1C. In our analyses, we classified our reads against the Genome Taxonomy Database (GTDB; Parks et al., 2018), and thus the taxonomy discussed in this study follows the standardized GTDB taxonomy, unless otherwise noted.

Bacterial clades with the highest positive loadings on PC1 (and therefore associated with higher FLI values) included members of the *Lachnospirales* and *Oscillospirales* taxonomic orders, of the *Bacilli* class, and of the *Ruminococcaceae, Bacteroidaceae* and *Lachnospiraceae* families (**Figure S2**). These observations led us to further analyses within a machine learning framework to estimate the relative contributions of individual bacterial taxa to the differences in FLI between study participants.

Bacterial lineages within the NCBI *Clostridium* subclusters IV and XIVa associate with FLI

In our machine learning framework, we used the known covariates in addition to individual archaeal and bacterial "balances" as the predicting features. Briefly, each balance represents a single internal node in a phylogenetic tree, and its value is a log-ratio of the abundances of the two clades descending this node (for details see methods, and Silverman et al., 2017). Continuous FLI and differences between FLI groups (FLI < 60, N = 4359 and FLI ≥ 60, N = 1910; see **Figures 1A, 1B**) were modeled with gradient boosting regression or classification using Leave-One-Group-Out Cross-Validation (LOGOCV) between participants from different regions.

After feature selection and Bayesian hyperparameter optimization, the correlation between the predictions of the final regression models (age, sex, self-reported alcohol use, and 18 bacterial balances as features; each trained on the data from 5/6 regions) and true values in unseen data from the omitted region averaged R^2 = 0.30 (0.26 – 0.33). After feature selection and optimization, the main

157 classification models (age, sex, and 11 bacterial balances as features; each trained on the data from 5/6
158 regions) averaged AUC = 0.75 (**Table S1**) and AUPRC = 0.56 (baseline at 0.30; **Table S2**) on (unseen)
159 test data from the omitted region. Models trained using only the covariates averaged AUC = 0.71
160 (AUPRC = 0.47) and using only the 11 bacterial balances they averaged AUC = 0.66 (AUPRC = 0.47)
161 on test data. Alternative models were constructed by excluding participants with FLI between 30 and
162 60 (N = 1583) and discerning between groups of FLI < 30 (N = 2776) and FLI $\geqslant$ 60 (N = 1910). These
163 models averaged AUC = 0.80 (AUPRC = 0.75, baseline at 0.41) on their respective test data. They
164 averaged AUC = 0.76 (AUPRC = 0.68) when using only the covariates, and AUC = 0.70 (AUPRC =
165 0.63) when using only the 20 bacterial balances.
166
167 Because training data from all 6 regions was used to prevent overfitting in the selection of core features
168 for all of the models, and similarly in searching for common hyperparameters, participants from the
169 validation region of each model (in the training partition) partly influenced these parameters. Thus, we
170 also constructed classification models discerning between the FLI < 60 and FLI $\geqslant$ 60 groups, where
171 data of the validation region was completely excluded in the feature selection and hyperparameter
172 optimization of each LOGOCV model. These models, using their individual feature sets and
173 hyperparameters, averaged AUC = 0.75 and AUPRC = 0.57 (baseline at 0.30) on test data from their
174 respective validation regions (**Table S3**). Using only covariates, they averaged AUC = 0.71 (AUPRC =
175 0.47), and AUC = 0.67 (AUPRC = 0.48) with only the bacterial balances.
176
177 To facilitate interpretability of the results, we primarily continued examining the main classification
178 models using a common set of core features. In these models, the median effect sizes of the features on
179 the model predictions at their minimum and maximum values were highest for age, followed by sex,
180 and the 11 balances in the phylogenetic tree (**Figures S2, S3**). All 11 associated balances were in

phylum *Firmicutes*, class *Clostridia*, and largely in the NCBI *Clostridium* subclusters IV and XIVa (**Figure 2).** The specific taxa represented standardized GTDB genera (NCBI in brackets) *Negativibacillus* (*Clostridium*), *Clostridium M* (*Lachnoclostridium* / *Clostridium*), *CAG-81* (*Clostridium*), *Dorea* (*Merdimonas* / *Mordavella* / *Dorea* / *Clostridium* / *Eubacterium*), *Faecalicatena* (*Blautia* / *Ruminococcus* / *Clostridium*), *Blautia* (*Blautia*), *Sellimonas* (*Sellimonas* / *Drancourtella*), *Clostridium Q* (*Lachnoclostridium* [*Clostridium*]) and *Tyzzerella* (*Tyzzerella* / *Coprococcus*). Notably, all but one of the features in the main classification models (n226) were identified in the feature selection for the alternative models (constructed otherwise identically, but FLI < 30 was compared against FLI ≥ 60 in different data partitions), together with 10 additional balances (**Figure S4**). Only one of the balances in the alternative models was outside phylum *Firmicutes* (n1712 in *Bacteroidota*), and in addition, 4 balances were outside class *Clostridia* (n481 in *Negativicutes*; n826, n1009 and n918 in *Bacilli*).

In addition to blood test results, FLI is based on anthropometric markers linked to metabolic syndrome, waist circumference and BMI. This prompted us to attempt to dissect the index and identify which of the covariates and associated microbial balances from the phylogenetic tree can be linked to blood GGT and triglycerides measurements (see **Figure 1B**), and therefore would be more specific to hepatic steatosis and liver damage. To do so, we performed feature selection (similarly to continuous FLI) for GGT and triglycerides measurements in subsets of participants grouped by age, sex, and BMI. The feature selection identified two balances within the NCBI *Clostridia* XIVa subcluster (identified as n336 and n330) which were important for both GGT and triglyceride level prediction, and thus likely specific to liver function (**Figure 2**). Bacterial taxa were positively linked to liver function in these balances, and included (NCBI species) *Clostridium clostridioforme*, *C. bolteae*, *C. citroniae*, *C. saccharolyticum* and *C. symbiosum*.

205

206 **Ethanol and acetate production pathways are identified in representative bacterial**
207 **genomes from taxa linked to liver function**

208 The values of predictive balances in the phylogenetic tree cannot be summarized for individual taxa,
209 which means that only a qualitative investigation of the associations between their metabolism and
210 fatty liver was possible in this study. We identified genetic pathways predicted to encode for SCFA
211 (acetate, propanoate, butanoate) and ethanol production, BA metabolism, and choline degradation to
212 trimethylamine in representative genomes from the taxa we identified to be linked to liver function
213 (**Figure S3**). These specific processes were chosen because they have been previously identified to
214 have a mechanistic link to NAFLD (see *e.g.,* Safari and Gérard, 2019).

215

216 Acetate and ethanol production pathways appeared to be more abundant in the representative genomes
217 of the taxa which had a positive association with FLI. In the liver function specific clades, n336 and
218 n330, MetaCyc pathways for pyruvate fermentation to ethanol III (PWY-6587) and L-glutamate
219 degradation V (via hydroxyglutarate; P162-PWY; produces acetate and butanoate) were present only in
220 genomes positively associated with FLI. In balance n336, also heterolactic fermentation (P122-PWY;
221 produces ethanol and lactate) was more often encoded in the FLI-associated clade (3/5) than the
222 opposing clade (1/2). In representative genomes from the non-liver-specific balance n355, potential
223 ethanol producers (PWY-6587) were seen in the positively associated clade, but for most balances such
224 trends were not clear in the qualitative analysis. Furthermore, we did not detect any of these pathways
225 in the representative genomes of two individual taxa positively associated with FLI, *Negativibacillus*
226 *sp000435195* and *Phocea massiliensis* (**Figure S3**).

227

Discussion

The pathophysiology of fatty liver disease in general, and NAFLD in particular, is complex and its clinical diagnosis can be difficult (Haas et al., 2016). In this study, we leveraged multi-omics data from a large population study (FINRISK02) to identify broad links between the overall gut microbiome composition and fatty liver disease, using FLI as a recognized proxy (**Figure 1C**), and identified specific microbial taxa and lineages predictive of a higher FLI (**Figure 2**). Considering that the predictive ability of FLI for clinically diagnosed NAFLD ranges between AUC = 0.81 – 0.93, in populations of Caucasian ethnicity such as ours (Vanni and Bugianesi, 2015), our models were able to reasonably predict the FLI group with AUC = 0.75 (AUPRC = 0.56, baseline at 0.30), while extrapolating to a validation region not used in training of the model.

Our additional analyses support these results. Excluding participants with intermediate FLI (between 30 – 60) increased the accuracy slightly (to AUC = 0.8 and AUPRC = 0.75, baseline at 0.41). However, discerning between participants with probable fatty liver disease (FLI $\geq$ 60) from others presents a clinically more relevant target for detecting changes in microbiome composition associated with development of the disease. In another set of models, we negated the influence of validation region data in the individual models also for feature selection and hyperparameter optimization during training. This led to individualized sets of features and parameters in the models, but the average performance of the models was almost identical on validation region samples in the test data (AUC = 0.75 and AUPRC 0.57, baseline at 0.30). The aim of our study was to find patterns in microbiome composition which would be generalizable across the 6 sampled geographic regions in Finland and easy to interpret. Thus, we consider the use of all training data to define the common core feature set justified. This goal also guided our overall modeling architecture and likely led to a lower performance than if we instead performed interpolation within a smaller scale (*e.g.*, He et al., 2018).

When interpreting results, several different levels of associations can be considered according to types of fatty liver disease and the gut microbiome composition. Because FLI has been mostly validated with simple steatosis and NAFLD (Bedogni et al., 2006; Vanni and Bugianesi, 2015), we can conservatively contextualize our findings with previous associative work that used these diagnoses or clinical manifestations, only.

FLI modeling reveals consistent associations between gram-positive *Clostridia* and fatty liver disease

We found significant linear correlations between the first three bacterial PC-axes of our samples (a measure of beta diversity) and FLI (see results and Figure 1C). Previous studies have shown differences in beta diversity in relation to NAFLD (Kim et al., 2019). However, FLI used in our study as a proxy for liver disease also includes features such as BMI and waist circumference, which associate with metabolic syndrome and type 2 diabetes (Aron-Wisnewsky et al., 2020). Links between these diseases and gut microbiome composition are well documented in previous studies (Castaner et al., 2018). It is thus not surprising that bacterial beta diversity and FLI were correlated, but unfortunately this simple correlation does not enable untangling the relative contributions of fatty liver disease and other metabolic diseases to the differences in bacterial beta diversity.

Several studies have reported highly specific changes in microbial abundances in relation to NAFLD (Wigg et al., 2001; Mouzaki et al., 2013; Zhu et al., 2013; Shen et al., 2017). In summary, while also conflicting results have been reported, generally increases in *Lactobacillus* and *Escherichia* genera, and a decrease in *Coprococcus* genus have been most often associated with a NAFLD diagnosis (Sharpton et al., 2019). Furthermore, increased abundance of several gram-positive bacteria belonging

to the *Clostridium* genus have often been positively linked with NAFLD (Jiang et al., 2015; Loomba et al., 2017). Differences in unconstrained between-samples (beta) diversity have been also documented for persistent NAFLD (Kim et al., 2019) and along the NAFLD-cirrhosis spectrum (Caussy et al., 2019).

In our study, abundances of bacteria from the *Coprococcus* genus were not specifically associated with FLI, although the genus was nested inside our predictive balances. Strikingly, we did not identify any bacterial associations with FLI outside of the *Firmicutes* phylum. A possible reason for this might be the higher relative abundance of phylum *Firmicutes* at high latitudes, where Finland is (Suzuki and Worobey, 2014). Among the associations we identified, *Faecalicatena gnavus* (NCBI: *Ruminococcus gnavus*) was positively linked with FLI as part of 3 predictive balances, and associated in previous studies with liver cirrhosis (Qin et al., 2014). Interestingly, none of the oral *Firmicutes*, such as *Veillonella*, suggested to invade the gut, were identified in our own analyses. This might be caused by using FLI as a proxy, which is likely not closely associated with advanced liver disease, such as cirrhosis, and thus would target an earlier phase of liver disease progression.

Two individual taxa, *Negativibacillus sp000435195* and *Phocea massiliensis*, were highly predictive of FLI group (**Figures 2, S2**), but not of its liver function-specific components. The associations of these taxa with fatty liver disease have not been documented previously. However, a decreasing abundance of both bacteria, *Negativibacillus sp000435195* (NCBI: *Clostridium* sp. CAG:169) and *Phocea massiliensis* (NCBI: *Phocea massiliensis*), were seen when the intake of meat and refined cereal was reduced isocalorically in favor of fruit, vegetables, wholegrain cereal, legumes, fish and nuts in overweight and obese subjects in Italy (Meslier et al., 2020). While comparisons between these studies are difficult due to annotation, bacteria such as *Faecalicatena gnavus* (NCBI:

Ruminococcus gnavus) and *Clostridium Q saccharolyticum* (NCBI: *Clostridium saccharolyticum*) were also found to respond negatively to the Mediterranean diet. Together with their positive association with FLI in our study, these observations would warrant further study on these species as plausible biomarkers for healthy diet choices.

Most taxa in our study with a positive association with FLI belonged to the (broadly defined) *Clostridium* NCBI genus, which supports several previous observations (Jiang et al., 2015; Loomba et al., 2017). However, taxonomic standardization according to GTDB has identified the *Clostridium* genus as the most phylogenetically inconsistent of all bacterial genera in the NCBI taxonomy, and divides it into a total of 121 monophyletic genera in 29 distinct families (Parks et al., 2018). These reassignments, although more accurate and sensible, complicate comparisons to previous research studies. However, our results strongly suggest that this finer taxonomic resolution might robustly reveal novel discoveries. Thus, while (shallow) shotgun metagenomic sequencing is often more costly than amplicon sequencing, this might be justified by the increased resolution which is required to accurately identify specific taxon-based associations (*e.g.*, Hillmann et al., 2018, 2020).

Bacterial taxa associated with a high FLI have a genetic potential to exacerbate the development of fatty liver disease

We identified several plausible new associations between individual taxa and clades of bacteria and fatty liver. All taxa were from class *Clostridia*, which are obligate anaerobes. We observed that reference genomes from the bacterial taxa positively associated with FLI in the liver-specific balances harbored several genetic pathways necessary for ethanol production. Specifically, genes predicted to enable the fermentation of pyruvate to ethanol (MetaCyc PWY-6587) appeared to be common. Endogenous production of ethanol has been known to both induce hepatic steatosis and increase

intestinal permeability (de Faria Ghetti et al., 2018) and several of the taxa identified in our study have also been experimentally shown to produce ethanol, such as *C. M asparagiforme*, *C. M bolteae*, *C. M clostridioforme / C. M clostridioforme A* (Mohan et al., 2006), and *C. Q Saccharolyticum* (Murray et al., 1982). The relative abundances of these putatively ethanol-producing taxa were predictive of FLI groups in previously unseen data. However, the self-reported alcohol consumption from the participants was not among the best predictors for the FLI groups, as it was excluded in the feature selection step.

All reference genomes from taxa positively associated with FLI in balance n330 harboured genes predicted to encode for the L-glutamate fermentation V (P162-PWY; **Figure S3**) pathway, which results in the production of acetate and butanoate. Glutamate fermentation could lead to increased microbial protein fermentation in the gut, which has been previously been linked with obesity, diabetes and NAFLD (Diether and Willing, 2019). Recently, the combined intake of fructose and microbial acetate production in the gut was experimentally observed to contribute to lipogenesis in the liver in a mouse model (Zhao et al., 2020). Interestingly, *C. Q saccharolyticum* (in our study, a FLI-associated species deriving from balance n330), was experimentally shown to ferment various carbohydrates, including fructose, to acetate, hydrogen, carbon dioxide, and ethanol (Murray et al., 1982). Furthermore, while our own pathway analysis did not detect BA modification pathways in the reference genome of *C. Q saccharolyticum*, a strain of this species has been highlighted as a probable contributor to NAFLD development through the synthesis of secondary BA (Jiao et al., 2019). The links between dietary intake and gene regulation, combined with microbial fermentation in the gut warrant further mechanistic experiments to elucidate their links with fatty liver, and likely other metabolic diseases.

Intriguingly, NAFLD-associated ethanol producing bacteria in previous cohort studies have all been gram-negatives, such as (NCBI-defined) *Klebsiella pneumoniae* (Yuan et al., 2019) and *Escherichia coli* (Zhu et al., 2013). In our population sample, instead of gram-negatives, bacteria from the *C. M bolteae, C. M clostridioforme / C. M clostridioforme A* and *C. M citroniae* species (linked in our study with FLI as deriving from balance n336) have been described as opportunistic pathogens (Dehoux et al., 2016), and are hypothesized to exacerbate fatty liver development similarly through endogenous ethanol production. This result suggests that geographical (He et al., 2018) and ethnic (Deschasaux et al., 2018) variability might also strongly affect gut microbiome composition and its associations with disease. In addition to putative endogenous ethanol producers, we identified other FLI-associated taxa deriving from balance n330, for which reference genomes harbored a genetic pathway predicted to encode for the ability to ferment L-lysine to acetate and butyrate. While the production of these SCFAs is often considered beneficial for gut health, other metabolism of proteolytic bacteria might negatively contribute to fatty liver disease (Canfora et al., 2019).

Through modeling a previously validated risk index for fatty liver, we could associate specific members of the gut microbiome with the disease across geographical regions in this representative sample of the general population in Finland. In addition, sex and age of participants were also strongly predictive of the FLI group in our models (**Figures 2, S2, Table S1**). Their similar positive associations with fatty liver disease are known from previous studies (*e.g.*, Cheng et al., 2013; Lonardo et al., 2019). The associated microbial balances could be used to improve the predictions above the baseline of these covariates on 5/6 regions in Finland. For example, in the model cross-validated with Lapland the balances were more predictive of FLI group than the covariates by themselves, and their combination increased the AUC further. Yet, when testing the model where Turku/Loimaa region was used for cross-validation, the microbial balances were slightly predictive of

FLI group but failed to improve the AUC over the covariates (**Table S1**). This pattern might stem from the cultural and genetic west-east division in Finland (Näyhä, 1989; Kerminen et al., 2017), with a closer proximity of the Helsinki/Vantaa region to eastern regions than Turku/Loimaa, in both terms. It is thus likely that further incorporation and investigation on the use of spatial information in microbiome modeling would elucidate these geographical patterns in taxa-disease associations.

Conclusions

Modeling an established risk index for fatty liver enabled the detection of associations between the disease and gut microbiome composition, even to the level of individual taxa. These taxa and clades were all from the obligately anaerobic gram-positive class *Clostridia*, from several redefined GTDB genera previously included in the polyphyletic NCBI genus *Clostridium*. Many of the representative genomes of taxa positively associated with fatty liver had genomic potential for endogenous ethanol production. This suggests a possible mechanistic link to the pathophysiology of liver disease through increased gut permeability and induction of hepatic steatosis. Further mechanistic links with microbial production of SCFAs, especially acetate, and fatty liver development are also likely. Our models were able to predict the FLI group of participants across geographical regions in Finland, showing that the associations are robust and mostly generalizable in the sampled population.

Methods

Survey details and sample collection

Cardiovascular disease risk factors have been monitored in Finland since 1972 by conducting a representative population survey every five years (Borodulin et al., 2018). In the FINRISK 2002 survey, a stratified random population sample was conducted on six geographical regions in Finland.

These are North Karelia and Northern Savo in eastern Finland, Turku and Loimaa regions in southwestern Finland, the cities of Helsinki and Vantaa in the capital region, the provinces of Northern Ostrobothnia and Kainuu in northwestern Finland, and the province of Lapland in northern Finland.

Briefly, at baseline examination the participants filled out a questionnaire form, and trained nurses carried out a physical examination and blood sampling in local health centers or other survey sites. Data was collected for physiological measures, biomarkers, and dietary, demographic and lifestyle factors. Stool samples were collected by giving willing participants a stool sampling kit with detailed instructions. These samples were mailed overnight between Monday and Thursday under Finnish winter conditions to the laboratory of the Finnish Institute for Health and Welfare, where they were stored at -20ºC. In 2017, the samples were shipped still unthawed to University of California San Diego for microbiome sequencing.

Further details of the FINRISK cohorts and sampling have been extensively covered in previous publications (Borodulin et al., 2015; Salosensaari et al., 2020). The Coordinating Ethics Committee of the Helsinki University Hospital District approved our study protocol. All participants have given their written informed consent.

Stool DNA extraction and shallow shotgun metagenome sequencing

A miniaturized version of the Kapa HyperPlus Illumina-compatible library prep kit (Kapa Biosystems) was used for library generation, following the previously published protocol (Sanders et al., 2019). DNA extracts were normalized to 5 ng total input per sample in an Echo 550 acoustic liquid handling robot (Labcyte Inc). A Mosquito HV liquid-handling robot (TTP Labtech Inc was used for 1/10 scale enzymatic fragmentation, end-repair, and adapter-ligation reactions). Sequencing adapters were based

on the iTru protocol (Glenn et al., 2019), in which short universal adapter stubs are ligated first and then sample-specific barcoded sequences added in a subsequent PCR step. Amplified and barcoded libraries were then quantified by the PicoGreen assay and pooled in approximately equimolar ratios before being sequenced on an Illumina HiSeq 4000 instrument to an average read count of approximately 900,000 reads per sample.

Taxonomic matching and phylogenetic transforms

To improve the taxonomic assignments of our reads, we used a custom index (Méric et al., 2019) based on the Genome Taxonomy Database (GTDB) release 89 (Parks et al., 2018, 2020) taxonomic redefinitions for read classification with default parameters in Centrifuge 1.0.4 (Kim et al., 2016). After read classification, all following steps were performed with R version 3.5.2 (R Core Team, 2018). To reduce the number of spurious read assignments, and to facilitate more accurate phylogenetic transformations, only reads classified at the species level, matching individual GTDB reference genomes, were retained. Samples with less than 50,000 reads, from pregnant participants or recorded antibiotic use in the past 6 months were removed, resulting in a final number of 6,269 samples. We first filtered taxa not seen with more than 3 counts in at least 1% of samples and those with a coefficient of variation ≤ 3 across all samples, following (McMurdie and Holmes, 2013) with a slight adaption from 20% of samples to 1% of samples, because of our larger sample size. The complete bacterial and archaeal phylogenetic trees of the GTDB release 89 reference genomes, constructed from an alignment of 120 bacterial or 122 archaeal marker genes (Parks et al., 2018), were then combined with our taxa tables. The resulting trees were thus subset only to species which were observed in at least one sample in our data. The read counts were transformed to phylogenetic node balances in both trees with PhILR (Silverman et al., 2017). The default method for PhILR inputs a pseudocount of 1 for taxa absent in an individual sample before the transform.

In this study, we did not specifically and solely use relative abundances at various taxonomic levels, as is common practice for microbiome studies. Instead, we applied a PhILR transformation to our microbial composition data (Silverman et al., 2017), introducing the concept of microbial "balances". Indeed, evolutionary relationships of all species harbored in each microbiome sample can be represented on a phylogenetic tree, with species typically shown as external nodes that are related to each other by multiple branches connected by internal nodes. In this context, the value of a given microbial "balance" is defined as the log-ratio of the geometric mean abundance between two groups of microbes descending from the same corresponding internal node on a microbial phylogenetic tree. This phylogenetic transform was used because it i) addresses the compositionality of the metagenomic read data (Gloor et al., 2017), ii) permits simultaneous comparison of all clades without merging the taxa by predefined taxonomic levels, and iii) enables evolutionary insights into the microbial community. The links between microbes and their environment, such as the human gut, is mediated by their functions. Different functions are known to be conserved at different taxonomic resolutions, and most often at multiple different resolutions (Louca et al., 2018). Thus, associations between the microbes and the response variable are likely not best explained by predefined taxonomic levels. In the absence of functional data, concurrently analyzing all clades (partitioned by the nodes in the phylogenetic tree) would likely enable the detection of the associations at the appropriate resolution depending on the function and the local tree topography.

Covariates

Because fatty liver disease is underdiagnosed at the population level (Alexander et al., 2018) and our sampling did not have extensive coverage of liver fat measurements, we chose to use the Fatty Liver index (Bedogni et al., 2006) as a proxy for fatty liver. Furthermore, the index performs well in cohorts

of Caucasian ethnicity, such as ours, to diagnose the presence of NAFLD (Vanni and Bugianesi, 2015). We calculated FLI after Bedogni et al., (2006): ($e^{0.953*\log_e(\text{triglycerides mg/dL}) + 0.139*\text{BMI} + 0.718*\log_e(\text{GGT}) + 0.053*\text{waist circumference} - 15.745}$) / ($1 + e^{0.953*\log_e(\text{triglycerides mg/dL}) + 0.139*\text{BMI} + 0.718*\log_e(\text{GGT}) + 0.053*\text{waist circumference} - 15.745}$) * 100. We chose the cutoff at FLI ≥ 60 to identify participants likely to be diagnosed with hepatic steatosis (positive likelihood ratio = 4.3 and negative likelihood ratio = 0.5 in Bedogni et al., 2006). Triglycerides, gamma glutamyl transferase (GGT), BMI and waist circumference measurements had near complete coverage for the participants in our data. Self-reported alcohol use was calculated as grams of pure ethanol per week. Cases with missing values were omitted in linear regression models. At least one feature used for FLI calculation was missing for 20 participants (0.3%) and the self-reported alcohol use was missing for 247 participants (3.9%). In the machine learning framework, missing values for FLI and self-reported alcohol use were mean imputed. However, for the feature selection to identify liver function-specific balances, GGT, triglycerides and BMI were not imputed but observations where any of these were missing were simply removed.

Beta-diversity and linear modeling of FLI

Beta-diversity was calculated as Euclidian distance of the PhILR balances through Principal Component Analysis (PCA) on bacterial and archaeal balances separately with ‘rda’ in vegan 2.5.6 (Oksanen et al., 2018). A linear regression model was constructed for FLI with ‘lm’ in base R (R Core Team, 2018) with $\log_{10}$-transformed FLI as the dependent variable and with first three bacterial PCs, sex, age, and self-reported alcohol use as the independent variables. Archaeal PCs were dropped from the model because none of them were significantly correlated with FLI (all $P > 0.05$). Variation of the samples on the top two bacterial PC axes by their effect sizes in the model were plotted together with a unit vector of $\log_{10}$(FLI) to show their correlation.

FLI modeling within a machine learning framework

In the machine learning framework, both regression and categorical models were constructed for FLI. The feature selection, hyperparameter optimization and cross-validation methods were identical for both approaches, unless otherwise stated. The continuous or categorical FLI (groups of FLI < 60 and FLI ≥ 60) were modeled with xgboost 0.90.0.2 (Chen and Guestrin, 2016) by using both bacterial and archaeal balances, sex, age, and self-reported alcohol use as preliminary predictor features. We used FLI 60 as the cutoff for ruling in fatty liver (steatosis) for the classification, after Bedogni et al., (2006). The data was first split to 70% train and 30% test sets while preserving sex and region balance. To take into account geographical differences (*e.g.*, He et al., 2018) and to find robust patterns across all 6 sampled regions in Finland between the features and FLI group, we used Leave-One-Group-Out Cross-Validation (LOGOCV) inside the 70% train set to construct 6 separate models in each optimization step. Because of high dimensionality of the data (3423 predictor features) feature selection by filtering was first performed inside the training data, based on random forest permutation as recommended by Bommert et al., (2020). Briefly, permutation importance is based on accuracy, or specifically the difference in accuracy between real and permuted (random) values of the specific variable, averaged in all trees across the whole random forest. The permutation importance in models based on the 6 LOGOCV subsets of the training data were calculated with mlr 2.16.0 (Bischl et al., 2016) and the simple intersect between the top 50 features in all LOGOCV subsets were retained as the final set of features. Thus, the feature selection was influenced by the training data from all 6 geographical regions, but this only serves to limit the number of chosen features because of the required simple intersect. This approach was used to obtain a set of core predictive features which would have potential for generalizability across the regions. The number of features included in the models by this approach was deemed appropriate, since the relative effect size of the last included predictor was very small (< 0.1 change in classification probability across its range).

Bayesian hyperparameter optimization of the xgboost models was then performed with only the selected features. An optimal set of parameters for the xgboost models were searched over all LOGOCV subsets with 'mbo' in mlrMBO 1.1.3 (Bischl et al., 2018), using 30 preliminary rounds with randomized parameters, followed by 100 optimization rounds. Parameters in the xgboost models and their considered ranges were learning rate (eta) [0.001, 0.3], gamma [0.1, 5], maximum depth of a tree [2, 8], minimum child weight [1, 10], fraction of data subsampled per each iteration [0.2, 0.8], fraction of columns subsampled per tree [0.2, 0.9], and maximum number of iterations (nrounds) [50, 5000]. The parameters recommended by these searchers were as following for regression: eta=0.00889; gamma=2.08; max_depth=2; min_child_weight=8; subsample=0.783; colsample_bytree=0.672; nrounds=1810, and for classification: eta=0.00107; gamma=0.137; max_depth=5; min_child_weight=9; subsample=0.207; colsample_bytree=0.793; nrounds=4328. We used Root-Mean-Square Error (RMSE) for the regression models and Area Under the ROC Curve (AUC) for the classification models to measure model fit on the left-out data (region) in each LOGOCV subset. The final models were trained on the LOGOCV subset training data, the data from one region thus omitted per model, and using the selected features and optimized hyperparameters. Validation of these models was conducted against participants only from the region omitted from each model, in the 30% test data which was not used in model training or optimization. Sensitivity analysis was conducted by using only the predictive covariates (sex and age) or balances separately, with the same hyperparameters, data partitions and final validation as for the full models.

Partial dependence interpretation of the FLI classification models

Because the classification models have a more clinically relevant modeling target for the difference between FLI < 60 and FLI ≥ 60, the latter used to rule in fatty liver (Bedogni et al., 2006), we further

interpreted the partial dependence of their predictions. Partial dependence of the classification model predictions on individual features was calculated with 'partial' in pdp 0.7.0 (Greenwell, 2017). The partial dependence of the features on the model predictions was also plotted, overlaying the results from each of the 6 models. For each feature, its relative effect on the model prediction was estimated as medians of the minimum and maximum yhat (output probability of the model for the FLI ≥ 60 class), calculated at the minimum and maximum values of the feature separately in each of the 6 models. The relative effects of the balances were then overlaid as a heatmap on a genome cladogram which covers all balances in the model with ggtree 2.1.1 (Yu et al., 2017).

Construction of alternative classification models to discern between FLI < 30 and FLI ≥ 60 groups

To assess robustness of the models and how removing the participants with intermediate FLI (between 30 and 60) affects model performance, we removed this group (N = 1910) and constructed alternative classification models to discern between the FLI < 30 and FLI ≥ 60 groups. Other than removing the intermediate FLI participants and resulting new random split to the train (70%) and test (30%) sets, these models were constructed identically to the main models, including LOGOCV design, feature selection, and hyperparameter optimization. The recommended parameters for this classification task were eta=0.00102; gamma=3.7; max_depth=2; min_child_weight=5; subsample=0.49; colsample_bytree=0.631; nrounds=3119. Interpretation of partial dependence was also performed identically, but only the relative effects of the model features were plotted without a cladogram.

Exclusion of validation region data from feature selection and hyperparameter optimization

Because training data from all 6 regions is used to inform the selection of optimal features and hyperparameters, the validation region data cannot be considered completely independent from the training of the LOGOCV models. Thus, we constructed a set of classification models for the FLI ≥ 60 and FLI < 60 groups, where all validation region samples also in the training data were excluded from the simple intercept of top50 features in each LOGOCV set and from the subsequent hyperparameter optimization. These models with individualized features and hyperparameters were then tested on the validation region samples in the unseen test data to estimate how model performance was affected. The main test (70%) and train (30%) sets were identical to the main models, but additionally 6 randomized 70/30 splits nested inside the test set (excluding the validation region) were used in hyperparameter optimization to reduce overfitting. Average optimal hyperparameters in the 6 models were eta=0.00106; gamma=4.3; max_depth=2; min_child_weight=7; subsample=0.36; colsample_bytree=0.613; nrounds=1772.

Identification of predictive features specific to liver function

Because the FLI also incorporates BMI and waist circumference, and they strongly contribute to the index (Bedogni et al., 2006), we deemed it necessary to further investigate which of the identified balances were specific to liver function. The participants were first grouped by age (< 40, 40 – 60, and 60 <), sex (female or male) and BMI (< 25, 25 – 30, and 30 <) into 18 categories (N = 105 ~ 711 per category). We performed feature selection similarly to the FLI models by fitting random forest regressors for GGT and triglycerides with mlr 2.16.0 (Bischl et al., 2016). This was done separately in each of the 18 categories, and in each category, we again used LOGOCV with the regions to obtain 6 runs per category. Finally, the features predictive of GGT or triglycerides in each category were

584 selected as the intersect of top 50 features in the 6 LOGOCV iterations by permutation importance. The
585 intersect of features predictive of GGT or triglycerides in any of the categories and the features
586 predictive of categorical FLI were identified as specific to liver function.
587
588 **Pathway inference for taxa associated with FLI**
589 Our taxonomic matching of the reads is based on the genomes of GTDB (release 89; Parks et al.,
590 2018), which are all complete or nearly complete and available in online databases. This enables us to
591 estimate the likely genetic content, and thus, the metabolic potential of the microbes associated with
592 FLI. We use this approach because the sequencing depth of our samples does not allow assembling
593 contigs and (metagenome-assembled) genomes, required for pathway predictions. Because of the
594 compositional phylogenetic transform, among other features of our data, previously developed
595 approaches such as PICRUSt (Douglas et al., 2019) could not be used here.
596
597 The genomes of all 336 bacteria under at least one of the predictive balances were downloaded from
598 NCBI. 119 of these genomes were originally not annotated, which is a requirement for pathway
599 prediction. Therefore, Prokka v1.14.6 (Seemann, 2014) was used to annotate the 119 unannotated
600 genomes as a preliminary step. Pathway predictions were then performed for all 336 genomes with
601 mpwt v0.5.3 (Belcour et al., 2019) multiprocessing tool for the PathoLogic pipeline of Pathway Tools
602 23.0 (Karp et al., 2019). Pathways for ethanol and short chain fatty acid (acetate, butyrate, propionate)
603 production, bile acid metabolism, and choline degradation to trimethylamine were identified from
604 MetaCyc pathway classifications (Caspi et al., 2018; **Table S4**). The prevalence of these processes was
605 then assessed in the analyzed genomes and summarized per process to consider the possible links of the
606 taxa with fatty liver pathophysiology. Finally, the presence of individual pathways for acetate and
607 ethanol production was also outlined for each genome.

Declarations

Ethics approval and consent to participate

The Coordinating Ethics Committee of the Helsinki University Hospital District approved our study protocol (Ref. 558/E3/2001). All participants have given their written informed consent.

Consent for publication

Not applicable.

Availability of data and material

The analysis code written for this study is included with the Supplementary Information. The datasets generated during and analyzed during the current study are not public, but are available based on a written application to the THL Biobank as instructed in: https://thl.fi/en/web/thl-biobank/for-researchers

Competing interests

V.S. has consulted for Novo Nordisk and Sanofi and received honoraria from these companies. He also has ongoing research collaboration with Bayer AG, all unrelated to this study. R.L. serves as a consultant or advisory board member for Anylam/Regeneron, Arrowhead Pharmaceuticals, AstraZeneca, Bird Rock Bio, Boehringer Ingelheim, Bristol-Myer Squibb, Celgene, Cirius, CohBar, Conatus, Eli Lilly, Galmed, Gemphire, Gilead, Glympse bio, GNI, GRI Bio, Inipharm, Intercept, Ionis, Janssen Inc., Merck, Metacrine, Inc., NGM Biopharmaceuticals, Novartis, Novo Nordisk, Pfizer, Prometheus, Promethera, Sanofi, Siemens, and Viking Therapeutics. In addition, his institution has received grant support from Allergan, Boehringer-Ingelheim, Bristol-Myers Squibb, Cirius, Eli Lilly and Company, Galectin Therapeutics, Galmed Pharmaceuticals, GE, Genfit, Gilead, Intercept, Grail,

Janssen, Madrigal Pharmaceuticals, Merck, NGM Biopharmaceuticals, NuSirt, Pfizer, pH Pharma, Prometheus, and Siemens. He is also co-founder of Liponexus, Inc.

Funding

This research was supported in part by grants from the Finnish Foundation for Cardiovascular Research, the Emil Aaltonen Foundation, the Paavo Nurmi Foundation, the Urmas Pekkala Foundation, the Finnish Medical Foundation, the Sigrid Juselius Foundation, the Academy of Finland (#321356 to A.H.; #295741, #307127 to L.L.; #321351 to T.N.) and the National Institutes of Health (R01ES027595 to M.J.). R.L. receives funding support from NIEHS (5P42ES010337), NCATS (5UL1TR001442), NIDDK (U01DK061734, R01DK106419, P30DK120515, R01DK121378, R01DK124318), and DOD PRCRP (W81XWH-18-2-0026). Additional support was provided by Illumina, Inc. and Janssen Pharmaceutica through their sponsorship of the Center for Microbiome Innovation at UCSD.

Authors' contributions

M.R., F.Å., V.M., V.S., R.K, L.L and T.N designed the work. A.H., L.V., G.M., P.J., V.S., M.J and R.K. acquired the data. M.R., L.L. and T.N. analyzed the data. M.R. wrote the manuscript in consultation with all authors. M.I., P.J., V.S., R.K., L.L. and T.N. supervised the work. All authors gave final approval of the version to be published.

Acknowledgements

We thank all participants of the FINRISK 2002 survey for their contributions to this work, and Tara Schwartz for assistance with laboratory work.

Figures

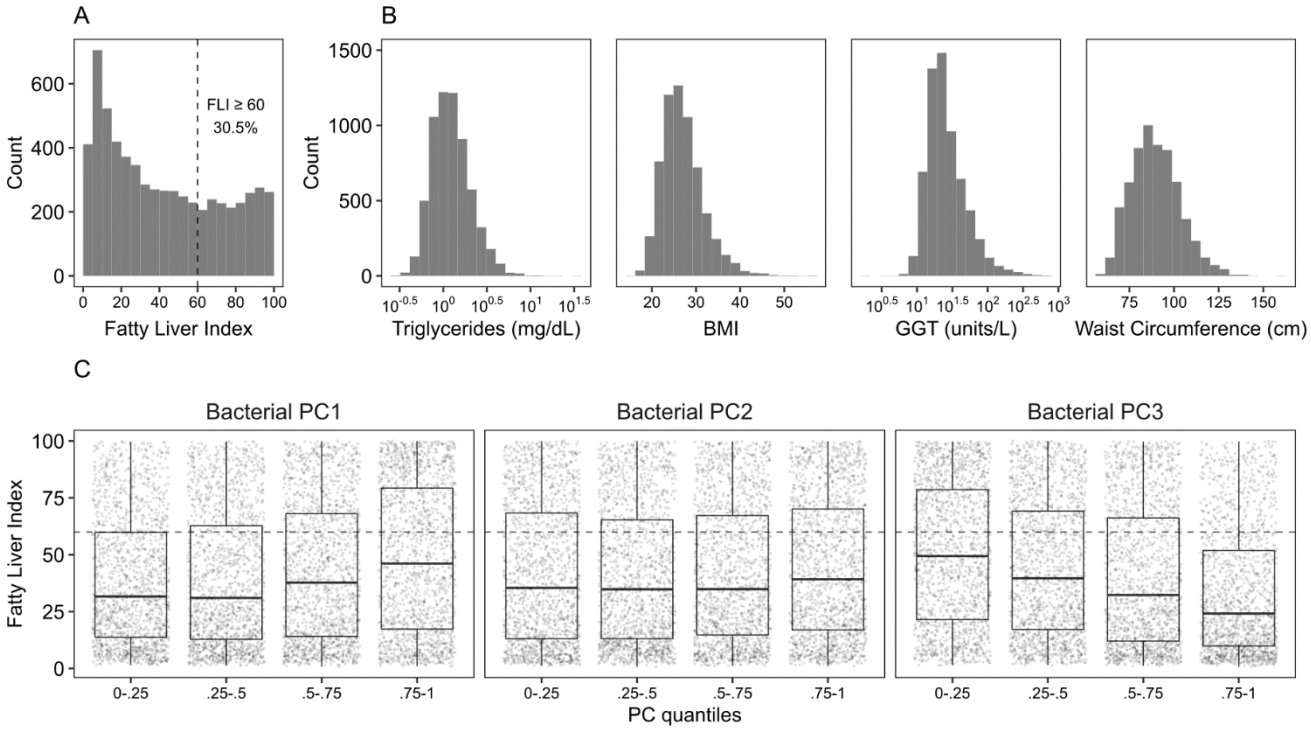
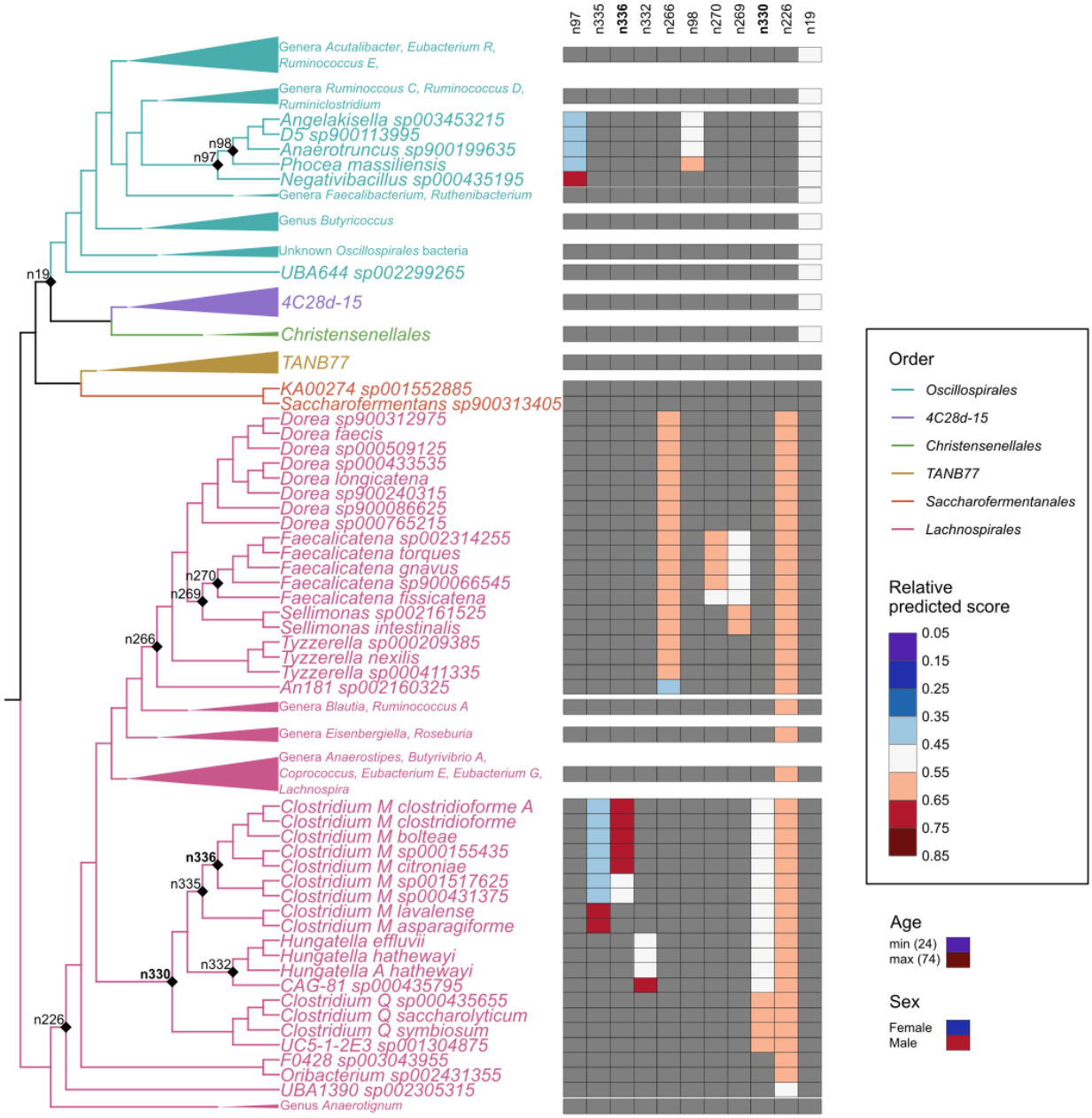


Figure 1. Distribution of FLI (A), its components (B), and FLI in quantiles of the first three PC components of the fecal bacterial composition of the participants (C). The cutoff at FLI = 60 used to divide the participants is indicated with a dashed line in panels A and C.



654 **Figure 2.** Relative effects of predictive balances and covariates on the FLI < 60 and FLI ≥ 60
655 classification model (AUC = 0.75) predictions. Nodes of the balances are indicated in the cladogram
656 and the relative effect sizes of their clades (opposite sides of each balance) are shown in the associated
657 heatmap. The relative effect sizes of the covariates (age and sex) are shown below the legend with a
658 heatmap on the same scale as was used for the balances. The two liver-specific balances associated
659 with triglyceride and GGT levels are indicated with bold font. Clades with redundant information have
660 been collapsed but their major genera are indicated. The complete tree is included in **Figure S3**.

References

Alexander, M., Loomis, A. K., Fairburn-Beech, J., van der Lei, J., Duarte-Salles, T., Prieto-Alhambra, D., et al. (2018). Real-world data reveal a diagnostic gap in non-alcoholic fatty liver disease. *BMC Medicine* 16, 130. doi:10.1186/s12916-018-1103-x.

Aron-Wisnewsky, J., Vigliotti, C., Witjes, J., Le, P., Holleboom, A. G., Verheij, J., et al. (2020). Gut microbiota and human NAFLD: disentangling microbial signatures from metabolic disorders. *Nature Reviews Gastroenterology & Hepatology* 17, 279–297. doi:10.1038/s41575-020-0269-9.

Bedogni, G., Bellentani, S., Miglioli, L., Masutti, F., Passalacqua, M., Castiglione, A., et al. (2006). The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. *BMC Gastroenterol* 6, 33. doi:10.1186/1471-230X-6-33.

Belcour, A., Frioux, C., Aite, M., Bretaudeau, A., and Siegel, A. (2019). Metage2Metabo: metabolic complementarity applied to genomes of large-scale microbiotas for the identification of keystone species. *bioRxiv*, 803056. doi:10.1101/803056.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., et al. (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research* 17, 1–5.

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., and Lang, M. (2018). mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. *arXiv:1703.03373 [stat]*. Available at: http://arxiv.org/abs/1703.03373 [Accessed February 18, 2020].

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., and Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis* 143, 106839. doi:10.1016/j.csda.2019.106839.

Borodulin, K., Tolonen, H., Jousilahti, P., Jula, A., Juolevi, A., Koskinen, S., et al. (2018). Cohort Profile: The National FINRISK Study. *International Journal of Epidemiology* 47, 696–696i. doi:10.1093/ije/dyx239.

Borodulin, K., Vartiainen, E., Peltonen, M., Jousilahti, P., Juolevi, A., Laatikainen, T., et al. (2015). Forty-year trends in cardiovascular risk factors in Finland. *Eur J Public Health* 25, 539–546. doi:10.1093/eurpub/cku174.

Canfora, E. E., Meex, R. C. R., Venema, K., and Blaak, E. E. (2019). Gut microbial metabolites in obesity, NAFLD and T2DM. *Nature Reviews Endocrinology* 15, 261–273. doi:10.1038/s41574-019-0156-z.

Carpino, G., Del Ben, M., Pastori, D., Carnevale, R., Baratta, F., Overi, D., et al. (2019). Increased liver localization of lipopolysaccharides in human and experimental non-alcoholic fatty liver disease. *Hepatology*, hep.31056. doi:10.1002/hep.31056.

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., et al. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* 46, D633–D639. doi:10.1093/nar/gkx935.

Caussy, C., Tripathi, A., Humphrey, G., Bassirian, S., Singh, S., Faulkner, C., et al. (2019). A gut microbiome signature for cirrhosis due to nonalcoholic fatty liver disease. *Nature Communications* 10, 1406. doi:10.1038/s41467-019-09455-9.

Chalasani, N., Younossi, Z., Lavine, J. E., Diehl, A. M., Brunt, E. M., Cusi, K., et al. (2012). The diagnosis and management of non-alcoholic fatty liver disease: Practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. *Hepatology* 55, 2005–2023. doi:10.1002/hep.25762.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 785–794. doi:10.1145/2939672.2939785.

Cheng, H.-Y., Wang, H.-Y., Chang, W.-H., Lin, S.-C., Chu, C.-H., Wang, T.-E., et al. (2013). Nonalcoholic Fatty Liver Disease: Prevalence, Influence on Age and Sex, and Relationship with Metabolic Syndrome and Insulin Resistance. *International Journal of Gerontology* 7, 194–198. doi:10.1016/j.ijge.2013.03.008.

Compare, D., Coccoli, P., Rocco, A., Nardone, O. M., De Maria, S., Cartenì, M., et al. (2012). Gut–liver axis: The impact of gut microbiota on non alcoholic fatty liver disease. *Nutrition, Metabolism and Cardiovascular Diseases* 22, 471–476. doi:10.1016/j.numecd.2012.02.007.

de Faria Ghetti, F., Oliveira, D. G., de Oliveira, J. M., de Castro Ferreira, L. E. V. V., Cesar, D. E., and Moreira, A. P. B. (2018). Influence of gut microbiota on the development and progression of nonalcoholic steatohepatitis. *Eur J Nutr* 57, 861–876. doi:10.1007/s00394-017-1524-x.

Dehoux, P., Marvaud, J. C., Abouelleil, A., Earl, A. M., Lambert, T., and Dauga, C. (2016). Comparative genomics of Clostridium bolteae and Clostridium clostridioforme reveals species-specific genomic properties and numerous putative antibiotic resistance determinants. *BMC Genomics* 17, 819. doi:10.1186/s12864-016-3152-x.

Deschasaux, M., Bouter, K. E., Prodan, A., Levin, E., Groen, A. K., Herrema, H., et al. (2018). Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nature Medicine* 24, 1526–1531. doi:10.1038/s41591-018-0160-1.

Diether, N. E., and Willing, B. P. (2019). Microbial Fermentation of Dietary Protein: An Important Factor in Diet–Microbe–Host Interaction. *Microorganisms* 7. doi:10.3390/microorganisms7010019.

Douglas, G. M., Maffei, V. J., Zaneveld, J., Yurgel, S. N., Brown, J. R., Taylor, C. M., et al. (2019). PICRUSt2: An improved and extensible approach for metagenome inference. *bioRxiv*, 672295. doi:10.1101/672295.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.* 24, 392–400. doi:10.1038/nm.4517.

735 Glenn, T. C., Nilsen, R. A., Kieran, T. J., Sanders, J. G., Bayona-Vásquez, N. J., Finger, J. W., et al.
736 (2019). Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456
737 combinatorially-indexed Illumina libraries (iTru & iNext). *PeerJ* 7, e7755.
738 doi:10.7717/peerj.7755.

739 Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets
740 Are Compositional: And This Is Not Optional. *Front Microbiol* 8.
741 doi:10.3389/fmicb.2017.02224.

742 Greenwell, B. M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*
743 9, 421–436.

744 Haas, J. T., Francque, S., and Staels, B. (2016). Pathophysiology and Mechanisms of Nonalcoholic
745 Fatty Liver Disease. *Annual Review of Physiology* 78, 181–205. doi:10.1146/annurev-physiol-
746 021115-105331.

747 He, Y., Wu, W., Zheng, H.-M., Li, P., McDonald, D., Sheng, H.-F., et al. (2018). Regional variation
748 limits applications of healthy gut microbiome reference ranges and disease models. *Nature*
749 *Medicine* 24, 1532–1535. doi:10.1038/s41591-018-0164-x.

750 Hillmann, B., Al-Ghalith, G. A., Shields-Cutler, R. R., Zhu, Q., Gohl, D. M., Beckman, K. B., et al.
751 (2018). Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems* 3.
752 doi:10.1128/mSystems.00069-18.

753 Hillmann, B., Al-Ghalith, G. A., Shields-Cutler, R. R., Zhu, Q., Knight, R., and Knights, D. (2020).
754 SHOGUN: a modular, accurate, and scalable framework for microbiome quantification.
755 *Bioinformatics*, btaa277. doi:10.1093/bioinformatics/btaa277.

756 Jiang, W., Wu, N., Wang, X., Chi, Y., Zhang, Y., Qiu, X., et al. (2015). Dysbiosis gut microbiota
757 associated with inflammation and impaired mucosal immune function in intestine of humans
758 with non-alcoholic fatty liver disease. *Sci Rep* 5, 8096. doi:10.1038/srep08096.

759 Jiao, N., Wu, D., Yang, Z., Fang, S., Li, X., Yuan, M., et al. (2019). Gut bacteria contributes to
760 NAFLD pathogenesis by promoting secondary bile acids biosynthesis. *The FASEB Journal* 33,
761 126.4-126.4. doi:10.1096/fasebj.2019.33.1_supplement.126.4.

762 Karp, P. D., Midford, P. E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., et al.
763 (2019). Pathway Tools version 23.0 update: software for pathway/genome informatics and
764 systems biology. *Briefings in Bioinformatics*, bbz104. doi:10.1093/bib/bbz104.

765 Kerminen, S., Havulinna, A. S., Hellenthal, G., Martin, A. R., Sarin, A.-P., Perola, M., et al. (2017).
766 Fine-Scale Genetic Structure in Finland. *G3: Genes, Genomes, Genetics* 7, 3459–3468.
767 doi:10.1534/g3.117.300217.

768 Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. (2016). Centrifuge: rapid and sensitive
769 classification of metagenomic sequences. *Genome Res.* doi:10.1101/gr.210641.116.

770 Kim, H.-N., Joo, E.-J., Cheong, H. S., Kim, Y., Kim, H.-L., Shin, H., et al. (2019). Gut Microbiota and
771 Risk of Persistent Nonalcoholic Fatty Liver Diseases. *Journal of Clinical Medicine* 8, 1089.
772 doi:10.3390/jcm8081089.

773 Koehler, E. M., Schouten, J. N. L., Hansen, B. E., Hofman, A., Stricker, B. H., and Janssen, H. L. A.
774 (2013). External Validation of the Fatty Liver Index for Identifying Nonalcoholic Fatty Liver
775 Disease in a Population-based Study. *Clinical Gastroenterology and Hepatology* 11, 1201–
776 1204. doi:10.1016/j.cgh.2012.12.031.

777 Liu, Y., Meric, G., Havulinna, A. S., Teo, S. M., Ruuskanen, M., Sanders, J., et al. (2020). Early
778 prediction of liver disease using conventional risk factors and gut microbiome-augmented
779 gradient boosting. Genetic and Genomic Medicine doi:10.1101/2020.06.24.20138933.

780 Lonardo, A., Nascimbeni, F., Ballestri, S., Fairweather, D., Win, S., Than, T. A., et al. (2019). Sex
781 Differences in Nonalcoholic Fatty Liver Disease: State of the Art and Identification of Research
782 Gaps. *Hepatology* 70, 1457–1469. doi:10.1002/hep.30626.

783 Loomba, R., Seguritan, V., Li, W., Long, T., Klitgord, N., Bhatt, A., et al. (2017). Gut Microbiome-
784 Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human
785 Nonalcoholic Fatty Liver Disease. *Cell Metabolism* 25, 1054-1062.e5.
786 doi:10.1016/j.cmet.2017.04.001.

787 Louca, S., Polz, M. F., Mazel, F., Albright, M. B. N., Huber, J. A., O'Connor, M. I., et al. (2018).
788 Function and functional redundancy in microbial systems. *Nat Ecol Evol* 2, 936–943.
789 doi:10.1038/s41559-018-0519-1.

790 Marchesini, G., Bugianesi, E., Forlani, G., Cerrelli, F., Lenzi, M., Manini, R., et al. (2003).
791 Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. *Hepatology* 37, 917–923.
792 doi:10.1053/jhep.2003.50161.

793 McMurdie, P. J., and Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis
794 and graphics of microbiome census data. *PLoS ONE* 8, e61217.
795 doi:10.1371/journal.pone.0061217.

796 Méric, G., Wick, R. R., Watts, S. C., Holt, K. E., and Inouye, M. (2019). Correcting index databases
797 improves metagenomic studies. *bioRxiv*, 712166. doi:10.1101/712166.

798 Meslier, V., Laiola, M., Roager, H. M., Filippis, F. D., Roume, H., Quinquis, B., et al. (2020).
799 Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and
800 causes changes in the gut microbiome and metabolome independently of energy intake. *Gut*.
801 doi:10.1136/gutjnl-2019-320438.

802 Miele, L., Valenza, V., Torre, G. L., Montalto, M., Cammarota, G., Ricci, R., et al. (2009). Increased
803 intestinal permeability and tight junction alterations in nonalcoholic fatty liver disease.
804 *Hepatology* 49, 1877–1887. doi:10.1002/hep.22848.

805 Mohan, R., Namsolleck, P., Lawson, P. A., Osterhoff, M., Collins, M. D., Alpert, C.-A., et al. (2006).
806 Clostridium asparagiforme sp. nov., isolated from a human faecal sample. *Systematic and*
807 *Applied Microbiology* 29, 292–299. doi:10.1016/j.syapm.2005.11.001.

Mouzaki, M., Comelli, E. M., Arendt, B. M., Bonengel, J., Fung, S. K., Fischer, S. E., et al. (2013). Intestinal microbiota in patients with nonalcoholic fatty liver disease. *Hepatology* 58, 120–127. doi:10.1002/hep.26319.

Murray, W. D., Khan, A. W., and van den BERG, L. (1982). Clostridium saccharolyticum sp. nov., a Saccharolytic Species from Sewage Sludge. *International Journal of Systematic Bacteriology* 32, 132–135. doi:10.1099/00207713-32-1-132.

Näyhä, S. (1989). Geographical variations in cardiovascular mortality in Finland, 1961-1985. *Scand J Soc Med Suppl* 40, 1–48.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2018). *vegan: Community Ecology Package*. Available at: https://CRAN.R-project.org/package=vegan [Accessed June 4, 2018].

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, 1–8. doi:10.1038/s41587-020-0501-8.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 36, 996–1004. doi:10.1038/nbt.4229.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi:10.1038/nature13568.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: https://www.R-project.org/ [Accessed March 4, 2019].

Rinella, M., and Charlton, M. (2016). The globalization of nonalcoholic fatty liver disease: Prevalence and impact on world health. *Hepatology* 64, 19–22. doi:10.1002/hep.28524.

Safari, Z., and Gérard, P. (2019). The links between the gut microbiome and non-alcoholic fatty liver disease (NAFLD). *Cell. Mol. Life Sci.* 76, 1541–1558. doi:10.1007/s00018-019-03011-w.

Salosensaari, A., Laitinen, V., Havulinna, A. S., Meric, G., Cheng, S., Perola, M., et al. (2020). Taxonomic Signatures of Long-Term Mortality Risk in Human Gut Microbiota. Epidemiology doi:10.1101/2019.12.30.19015842.

Sanders, J. G., Nurk, S., Salido, R. A., Minich, J., Xu, Z. Z., Zhu, Q., et al. (2019). Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biology* 20, 226. doi:10.1186/s13059-019-1834-9.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi:10.1093/bioinformatics/btu153.

Sharpton, S. R., Ajmera, V., and Loomba, R. (2019). Emerging Role of the Gut Microbiome in Nonalcoholic Fatty Liver Disease: From Composition to Function. *Clinical Gastroenterology and Hepatology* 17, 296–306. doi:10.1016/j.cgh.2018.08.065.

Shen, F., Zheng, R.-D., Sun, X.-Q., Ding, W.-J., Wang, X.-Y., and Fan, J.-G. (2017). Gut microbiota dysbiosis in patients with non-alcoholic fatty liver disease. *Hepatobiliary & Pancreatic Diseases International* 16, 375–381. doi:10.1016/S1499-3872(17)60019-5.

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* 6, e21887. doi:10.7554/eLife.21887.

Spencer, M. D., Hamp, T. J., Reid, R. W., Fischer, L. M., Zeisel, S. H., and Fodor, A. A. (2011). Association Between Composition of the Human Gastrointestinal Microbiome and Development of Fatty Liver With Choline Deficiency. *Gastroenterology* 140, 976–986. doi:10.1053/j.gastro.2010.11.049.

Suzuki, T. A., and Worobey, M. (2014). Geographical variation of human gut microbial composition. *Biology Letters* 10, 20131037. doi:10.1098/rsbl.2013.1037.

Toshikuni, N., Tsutsumi, M., and Arisawa, T. (2014). Clinical differences between alcoholic liver disease and nonalcoholic fatty liver disease. *World J Gastroenterol* 20, 8393–8406. doi:10.3748/wjg.v20.i26.8393.

Vanni, E., and Bugianesi, E. (2015). Editorial: utility and pitfalls of Fatty Liver Index in epidemiologic studies for the diagnosis of NAFLD. *Aliment Pharmacol Ther* 41, 406–407. doi:10.1111/apt.13063.

Wigg, A. J., Roberts-Thomson, I. C., Dymock, R. B., McCarthy, P. J., Grose, R. H., and Cummins, A. G. (2001). The role of small intestinal bacterial overgrowth, intestinal permeability, endotoxaemia, and tumour necrosis factor α in the pathogenesis of non-alcoholic steatohepatitis. *Gut* 48, 206–211. doi:10.1136/gut.48.2.206.

Yki-Järvinen, H. (2014). Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. *Lancet Diabetes Endocrinol* 2, 901–910. doi:10.1016/S2213-8587(14)70032-4.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8, 28–36. doi:10.1111/2041-210X.12628.

Yuan, J., Chen, C., Cui, J., Lu, J., Yan, C., Wei, X., et al. (2019). Fatty Liver Disease Caused by High-Alcohol-Producing Klebsiella pneumoniae. *Cell Metabolism* 30, 675-688.e7. doi:10.1016/j.cmet.2019.08.018.

Zhao, S., Jang, C., Liu, J., Uehara, K., Gilbert, M., Izzo, L., et al. (2020). Dietary fructose feeds hepatic lipogenesis via microbiota-derived acetate. *Nature* 579, 586–591. doi:10.1038/s41586-020-2101-7.

877 Zhu, L., Baker, S. S., Gill, C., Liu, W., Alkhouri, R., Baker, R. D., et al. (2013). Characterization of gut
878 microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between
879 endogenous alcohol and NASH. *Hepatology* 57, 601–609. doi:10.1002/hep.26093.