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

Fatty liver disease is the most common liver disease in the world. Its connection with the gut microbiome has been known for at least 80 years, but 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 of 6,269 Finnish participants. We based our models on biometric covariates and gut microbiome compositions from shallow metagenome sequencing. Our classification models could discriminate between individuals with a high FLI (≥ 60, indicates likely liver steatosis) and low FLI (< 60) in internal cross-region validation, consisting of 30% of the data not used in model training, with an average AUC of 0.75 and AUPRC of 0.56 (baseline at 0.30). In addition to age and sex, our models included differences in 11 microbial groups from class *Clostridia*, mostly belonging to orders *Lachnospirales* and *Oscillospirales*. Our models were also predictive of the high FLI group in a different Finnish cohort, consisting of 258 participants, with an average AUC of 0.77 and AUPRC of 0.51 (baseline at 0.21). Pathway analysis of representative genomes of the positively FLI-associated taxa in (NCBI) *Clostridium* subclusters IV and XIVa indicated the presence of *e.g.,* ethanol fermentation pathways. These results support several findings from smaller case-control studies, such as the role of endogenous ethanol producers in the development of fatty liver.

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

Introduction

Fatty liver disease affects roughly a quarter of the world's population.[1] It is characterized by accumulation of fat in the liver cells and is intimately linked with pathophysiology of metabolic syndrome.[2–4] 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.[5] 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.[6]

The human gut harbors up to 10^{12} microbes per gram of content,[7] 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.[8] Numerous studies over the past 80 years have reported associations between gut microbial composition and liver disease.[9] For example, gut permeability and overgrowth of bacteria in the small intestine,[10] changes in *Gammaproteobacteria* and *Erysipelotrichi* abundance during choline deficiency,[11] elevated abundance of ethanol-producing bacteria,[12,13] metagenomic signatures of specific bacterial species,[14,15] 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.[16] For example, oral taxa

appear to invade the gut in liver cirrhosis,[17] and this phenotype can accurately be detected by analyzing the fecal microbiome composition (AUC = 0.87 in a validation cohort).[8] 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),[18] showing that the signatures of serious future liver disease are easy to detect.

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.[19] 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.[19] Recent studies have also suggested that endogenous ethanol production by gut bacteria could lead to an increase in gut membrane permeability.[13] 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.[20]

Liver biopsy assessment is the current gold standard for diagnosis of fatty liver disease and its severity,[21] 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.[22,23]

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.[24] Because fatty liver disease is generally underdiagnosed in the general population,[25] 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.[26] In parallel, we used shallow shotgun sequencing to analyze gut microbiome composition,[27] 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 replicable in an external Finnish cohort, and thus generalizable in the Finnish population.

Results

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

In our main analyses, we classified our reads against the Genome Taxonomy Database (GTDB)[28]. This study mainly follows the GTDB taxonomy, unless otherwise noted. The Centrifuge/GTDB microbiome data used in our main analyses was based on archaeal and bacterial phylogenetic "balances". This method was used to associate larger groups or clades of related organisms with fatty liver disease, and to avoid grouping of taxa on strict hierarchical taxonomic ranks featuring varying ranges of evolutionary divergence.[28] Here, we used the PhILR

transform, where each balance represents a single internal node in a phylogenetic tree, and its value is a log-ratio of the abundances of the two descending clades (for details, see methods and ref. [29]). Positive values of the balance signify that the clade in the numerator is more abundant, and negative values that the clade in the denominator is more abundant. Thus, each association of a balance with the target variable necessarily includes both microbial clades descending from the node, one of them positively and the other negatively associated with the target variable. The clades in the numerator and denominator can be also freely switched by changing the sign of the balance value to retain the equivalence. Notably, we used this feature to show all balance-FLI associations in the positive direction to facilitate the comparison of their effect sizes (in **Figures S4**, **S7**, and **S9**).

Because the combined approach of using the GTDB taxonomy and the recently introduced PhILR -phylogenetic transform complicates the comparison of our results to previous studies, we also conducted more traditional statistical analyses with NCBI-annotated data to anchor our results to previous findings on the associations between fatty liver disease and gut microbiome composition. Overall, the Centrifuge/GTDB classification assigned 5.3 billion reads in the 6,269 samples (after exclusions in FINRISK 2002) to 23,457 bacterial and 1,248 archaeal taxa, and the SHOGUN/NCBI classification assigned 5.5 billion reads to 5,024 bacterial and 261 archaeal taxa. Starting from high level descriptions of the microbial communities in the high and low FLI groups (< 60 or ≥ 60 FLI; **Figure 1A**), the phylum-level distributions of bacterial and archaeal taxa appeared to be highly similar between the groups (**Figures S1, S2**). However, the proportion of taxa assigned to *Firmicutes* in Centrifuge/GTDB appeared to be slightly higher than in the SHOGUN/NCBI data. Furthermore, only 58% of the number of reads assigned in

Centrifuge/GTDB to 6 main archaeal phyla were assigned to a single main archaeal phylum in SHOGUN/NCBI. Alpha diversity (as Shannon diversity) was significantly lower in the high FLI group, in both the SHOGUN/NCBI data (14.7% lower; AIC = 6685; all $P < 1\times10^{-6}$) and the Centrifuge/GTDB (13.4% lower; AIC = 6607; all $P < 1\times10^{-4}$) data, while adjusting for age, sex, and self-reported alcohol use in both models.

To further examine the high-level associations between FLI (as a proxy of fatty liver disease) and microbial community composition in FINRISK 2002, we fit a linear regression model on the three first principal component (PC) axes of the fecal bacterial beta diversity (between individuals), sex, age, and alcohol. Log_{10}(FLI) significantly correlated with all three bacterial PC axes, sex, age, and alcohol use in Centrifuge/GTDB data (adjusted $R^2 = 0.29$; all $P < 1\times10^{-6}$), and PC1, PC3, sex, age, and alcohol use in SHOGUN/NCBI data (adjusted $R^2 = 0.27$; all $P < 1\times10^{-4}$). Correlations between FLI and archaeal PC axes were not significant in Centrifuge/GTDB data (at the chosen significance level, $P > 0.001$), and between FLI and bacterial PC2 in SHOGUN/NCBI data ($P > 0.001$). In Centrifuge/GTDB data, 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 in Centrifuge/GTDB data are visualized for each of the three components in **Figure 1C**. A comparison of these relationships in Centrifuge/GTDB and SHOGUN/NCBI is included in the SI (**Figure S3**).

We also further assessed the phylogenetic balances contributing to the PC axes in the Centrifuge/GTDB data. Bacterial clades associated with higher FLI values, on the positive side

of the balances contributing to PC1, included members of orders *Lachnospirales* and *Oscillospirales*, class *Bacilli*, and the *Ruminococcaceae, Bacteroidaceae* and *Lachnospiraceae* families (**Figure S4**). Several clades had a negative association with FLI, on the negative side of the balances contributing to PC1, such as order *Christensenellales* and genus *Faecalibacterium*. In addition, genus *Bifidobacterium* in PC2, and family *Bifidobacteriaceae* in PC3 had negative associations with continuous FLI.

Several bacterial taxa are differently abundant between the low and high FLI groups

We also assessed significant differences in abundances of individual taxa between the high and low FLI groups in FINRISK 2002. In Centrifuge/GTDB data, we identified 244 taxa (1% of total) with an increased abundance, and 437 taxa (1.9%) with a decreased abundance in the high FLI group (all Q values < 0.001; **Table S7**). In SHOGUN/NCBI data, 80 taxa (1.6%) had an increased abundance, and 44 (0.9%) had a decreased abundance in the high FLI group (all Q values < 0.001). While the number of associated taxa was higher in the Centrifuge/GTDB data than SHOGUN/NCBI data, the proportion of significantly associated taxa was similar between the two methods. In both data sets, family *Lachnospiraceae* comprised over 40% of taxa positively associated with the high FLI group and *Bacterioidaceae* were in the top 3 most common families. The negatively associated taxa were much more diverse, but *Ruminococcaceae* and *Oscillospiraceae* were among the top 3 most common families in both data sets (at least > 6% of all negatively associated taxa).

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

Continuous FLI and differences between FLI groups in the FINRISK 2002 cohort (FLI < 60, N = 4,359 and FLI ≥ 60, N = 1,910; see **Figures 1A, 1B, Table S1**) were modeled with gradient boosting regression or classification using Leave-One-Group-Out Cross-Validation (LOGOCV) between participants from different regions. Only the bacterial PhILR transformed Centrifuge/GTDB data were used here, to find robust associations between phylogenetically related bacterial clades and fatty liver disease (instead of single taxa).

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 classification models (age, sex, and 11 bacterial balances as features; each trained on the data from 5/6 regions) averaged AUC = 0.75 (**Table S2**) and AUPRC = 0.56 (baseline at 0.30; **Table S3**) on (unseen) test data from the omitted region. Models trained using only the covariates averaged AUC = 0.71 (AUPRC = 0.47) and using only the 11 bacterial balances they averaged AUC = 0.66 (AUPRC = 0.47) on test data. Alternative models were constructed by excluding participants with FLI between 30 and 60 (N = 1,583) and discerning between groups of FLI < 30 (N = 2,776) and FLI ≥ 60 (N = 1,910). These models averaged AUC = 0.80 (AUPRC = 0.75, baseline at 0.41) on their respective test data (**Tables S2, S3**). They averaged AUC = 0.76 (AUPRC = 0.68) when using only the covariates, and AUC = 0.70 (AUPRC = 0.63) when using only the 20 bacterial balances.

Because training data from all 6 regions was used to prevent overfitting in the selection of core features for all of the models, and similarly in searching for common hyperparameters, participants from the validation region of each model (in the training partition) partly influenced these parameters. Thus, we also constructed classification models discerning between the FLI < 60 and FLI ≥ 60 groups, where data of the validation region was completely excluded in the feature selection and hyperparameter optimization of each LOGOCV model. These models, using their individual feature sets and hyperparameters, averaged AUC = 0.75 and AUPRC = 0.57 (baseline at 0.30) on test data from their respective validation regions (**Table S4**). Using only covariates, they averaged AUC = 0.71 (AUPRC = 0.47), and AUC = 0.67 (AUPRC = 0.48) with only the bacterial balances.

Our external validation data consisted of 258 participants after exclusion of pregnant participants or those on antibiotics in the past 6 months, in the FINRISK 2007 population cohort[30] (**Table S1**, **Figure S5**). The participants originate from North Karelia and Helsinki/Vantaa regions in Finland, and their samples were processed with the same methodology as was used for FINRISK 2002 (with Centrifuge/GTDB approach and PhILR). In this external validation, the 6 full models trained with covariates and the 11 bacterial balances in FINRISK 2002 averaged AUC = 0.77 (AUPRC = 0.51, baseline at 0.21; **Table S5**). The covariate-only models averaged AUC = 0.72 (AUPRC = 0.40) and the balance-only models averaged AUC = 0.69 (AUPRC = 0.44). The receiver operating characteristic and precision-recall curves based on the averaged predictions of the models, tested on this external validation data, also display good predictive ability (AUC = 0.78, AUPRC = 0.51 with baseline at 0.51; **Figure S6**)

241

242 To facilitate interpretability of the results, we continued examining the main classification
243 models using a common set of core features. In these models, the median effect sizes of the
244 features on the model predictions at their minimum and maximum values were highest for age,
245 followed by sex, and the 11 balances in the phylogenetic tree (**Figures S7**, **S8**). All 11 associated
246 balances were in phylum *Firmicutes*, class *Clostridia*, and largely in the NCBI *Clostridium*
247 subclusters IV and XIVa (**Figure 2).** The specific taxa represented standardized GTDB genera
248 (NCBI in brackets) *Negativibacillus* (*Clostridium*), *Clostridium M* (*Lachnoclostridium* /
249 *Clostridium*), *CAG-81* (*Clostridium*), *Dorea* (*Merdimonas* / *Mordavella* / *Dorea* / *Clostridium* /
250 *Eubacterium*), *Faecalicatena* (*Blautia* / *Ruminococcus* / *Clostridium*), *Blautia* (*Blautia*),
251 *Sellimonas* (*Sellimonas* / *Drancourtella*), *Clostridium Q* (*Lachnoclostridium* [*Clostridium*]) and
252 *Tyzzerella* (*Tyzzerella* / *Coprococcus*). Notably, all but one of the features in the main
253 classification models (n226) were identified in the feature selection for the alternative models
254 (constructed otherwise identically, but FLI < 30 was compared against FLI ≥ 60 in different data
255 partitions), together with 10 additional balances (**Figure S9**). Only one of the balances in the
256 alternative models was outside phylum *Firmicutes* (n1712 in *Bacteroidota*), and in addition, 4
257 balances were outside class *Clostridia* (n481 in *Negativicutes*; n826, n1009 and n918 in *Bacilli*).
258 Also, negative associations with the high FLI group were seen for *An181 sp002160325* in the
259 balance n266, where it is compared against the clade including *Dorea*, *Faecalicatena*,
260 *Sellimonas* and *Tyzzerella* species (**Figures 2**, **S8**). A higher abundance of the clade including
261 *Angelakisella*, *D5*, *Anaerotruncus* and *Phocea* species (against *Negativibacillus sp00435195* in
262 balance n97) was also negatively associated with high FLI.

263

In addition to blood test results, FLI is based on two anthropometric markers linked to metabolic syndrome, waist circumference and BMI. This prompted us to dissect the Fatty Liver 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 most specific to hepatic steatosis and liver damage.[31] 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*. On the opposite, negatively associated side of the balances were, among others, (NCBI species) *Hungatella effluvii*, *H. hathewayi*, and two new GTDB-defined species *Clostridium M sp001517625* and *C. M sp000431375*.

Ethanol and acetate production pathways are identified in representative bacterial genomes from taxa linked to high FLI

The values of predictive balances in the phylogenetic tree cannot be summarized for individual taxa, which means that only a qualitative investigation of the associations between their metabolism and fatty liver was possible in this study. We identified genetic pathways predicted to encode for SCFA (acetate, propanoate, butanoate) and ethanol production, BA metabolism, and choline degradation to trimethylamine (TMA) in representative genomes from the taxa we

identified to be linked to liver function (**Figure S8**). These processes were chosen because they have been previously identified to have a mechanistic link to NAFLD (see *e.g.,* ref. [19]).

Acetate and ethanol production pathways appeared to be more common in the representative genomes of the taxa which had a positive association with FLI. In the liver function specific clades, n336 and n330, MetaCyc pathways for pyruvate fermentation to ethanol III (PWY-6587) and L-glutamate degradation V (via hydroxyglutarate; P162-PWY; produces acetate and butanoate) were present only in genomes positively associated with FLI. In balance n336, also heterolactic fermentation (P122-PWY; produces ethanol and lactate) was more often encoded in the clade positively associated with the high FLI group (3/5) than the opposing negatively associated clade (1/2). In representative genomes from the liver-specific balance n336, potential ethanol producers (PWY-6587) were seen in the positively associated clade (*Clostridium M clostridoforme A* and *Clostridium M sp000155435*), and not in the negatively associated clade (*Clostridium M sp001517625* and *Clostridium M sp000431375*). However, for most balances such trends were not clear in the qualitative analysis. Furthermore, we did not detect any of these pathways in the representative genomes of two individual taxa positively associated with FLI, *Negativibacillus sp000435195* and *Phocea massiliensis* (**Figure S8**).

Discussion

The pathophysiology of fatty liver disease in general, and NAFLD in particular, is complex and its clinical diagnosis can be difficult.[32] In this study, we utilized metagenomic data from a large population cohort (FINRISK 2002[30]) 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 positively and negatively associated with the high FLI group (**Figure 2**). It should be noted, that 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 diabetes.[16] Links between these diseases and gut microbiome composition are well documented in previous studies.[33] However, fatty liver disease is increasingly thought to be a component of the metabolic syndrome,[4,34] and while diabetes prevalence is higher in the high FLI group in FINRISK 2002, affected participants still consist only 11% of this group (**Table S1**). Furthermore, we would like to emphasize that our results are not suitable for current clinical application, and should be validated by further, preferably mechanistic studies. We also do not know if our results generalize outside the Finnish population, as all participants in this study were exclusively from Finnish cohorts.

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 the Finnish population,[23] our models were able to reasonably predict the FLI group with AUC = 0.75 (AUPRC = 0.56, baseline at 0.30), in our internal cross-region validation. Furthermore, the performance of our predictive models was highly similar in an external, Finnish validation cohort (AUC = 0.77, AUPRC = 0.51, baseline at 0.21).

Our additional analyses support these main results. While a thorough method comparison is beyond the scope of the current study, the results from the two taxa assignments were very similar despite their differences, such as the fourfold higher number of taxa in the Centrifuge/GTDB data. In the machine learning models (performed only with

Centrifuge/GTDB data), excluding participants with intermediate FLI (between 30 – 60) increased the accuracy slightly in the internal cross-validation (to AUC = 0.8 and AUPRC = 0.75, baseline at 0.41). However, discerning between participants with probable fatty liver disease (FLI ≥ 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 internal cross-validation (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 (see *e.g.,* ref. [35]).

When interpreting our results, several 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,[23,26] we can conservatively contextualize our findings with previous associative work that used these diagnoses or clinical manifestations, only. The cutoff used in our study at FLI ≥ 60 has been used to rule in liver steatosis in a Caucasian cohort comparable to ours,[26] but also a cutoff at FLI ≥ 48 has been found appropriate for simple steatosis in a Portuguese cohort.[36] Much lower cutoffs (FLI ≥ 20 to 30) have been

used in Asian cohorts.[37–39] Thus, it is likely that our high FLI groups include most participants with liver steatosis or fibrosis in both FINRISK cohorts, but the low FLI group also likely includes participants with low grade steatosis.

Traditional statistical analyses replicate previous findings on gut microbiome composition and fatty liver disease when using FLI as a risk index

Among the significant high level FLI-associated differences in the gut microbiomes of the participants in FINRISK 2002, we found a 14.7% lower Shannon alpha diversity in the high FLI group with SHOGUN/NCBI taxa assignments and 13.4% lower diversity with Centrifuge/GTDB assignments. These results are in good accordance with previous results of decreased gut bacterial diversity in patients with biopsy-proven non-alcoholic steatohepatitis (NASH), the most serious form of NAFLD.[40] In this case-control study, the Shannon diversity of gut microbiomes in NASH patients without liver cirrhosis was on average 7% lower compared to controls, and in patients with cirrhosis, 14% lower. A significantly decreased gut microbiome alpha diversity of similar magnitude was also seen in cohort participants with persistent NAFLD compared to controls.[41]

In both the SHOGUN/NCBI and Centrifuge/GTDB data, we found significant linear correlations between FLI and beta diversity, or two or three main bacterial PC-axes of the samples, respectively (**Figures 1C, S3**). The model fit was slightly better with Centrifuge/GTDB data, which might be due to the higher number of identified taxa, and thus increased taxonomic resolution (although including putative species in GTDB). Our results support previous observations of differences in beta diversity in relation to persistent

NAFLD,[41] and along the NAFLD-cirrhosis spectrum.[8] Through the loadings of the phylogenetic balances on the PC axes in the Centrifuge/GTDB data, we detected several previously known connections between microbial clades and FLI (**Figure S4**). Among others, we observed a positive association between high FLI and family *Lachnospiraceae* and negative associations for order *Christensenellales*, genus *Faecalibacterium*, and genus *Bifidobacterium*. The positive association is supported by previous findings of their connection with obesity,[42] and the negative associations by connections to lean individuals and healthy gut microbiome composition.[43–45]

Our differential abundance analysis also detected a high number of taxa with significantly increased or decreased abundance in the high FLI group. All following results were observed both in the Centrifuge/GTDB and SHOGUN/NCBI data sets, unless otherwise noted. Majority of the taxa with increased abundance in the high FLI group were from family *Lachnospiraceae*, which supports their positive association with NAFLD reported previously in a number of studies,[46] but also with obesity (**Table S7**).[42] The increased abundance of genus *Roseburia* has also been highlighted as a characteristic change in gut microbiome related to NAFLD.[46,47] In the current study, two members of genus *Roseburia* were in the top 10 taxa most strongly associated with high FLI. Furthermore, our results support previous findings on the positive associations of, for example, *Collinsella*,[40] *Prevotella copri*,[48] *Dorea*,[47] with NAFLD. We also detected increases in *Sutterella* and *Streptococcus*, previously associated with cirrhosis.[49] However, we did not find increases in families *Kiloniellaceae* and *Pasteurellaceae*, previously associated with NAFLD.[46] Among the individual taxa negatively associated with high FLI, families *Ruminococcaceae* and *Oscillospiraceae* (such as genus *Oscillibacter*) were common,

401 which supports previous findings on their connections with NAFLD.[12,41,46] A high number of
402 putative (GTDB) species were negatively associated with FLI in the Centrifuge/GTDB data,
403 which were understandably not present in the SHOGUN/NCBI data. Many of these were
404 classified in the recently described order *Christensenellales*,[28] including families such as CAG-
405 74, associated with healthy participants,[50] and *Christensenellaceae*, which are widespread,
406 highly heritable, and associated with health.[44,51]
407
408 While our results from common statistical experiments mainly supported previous findings, we
409 chose to leverage the phylogenetic information included in the GTDB data to find robust
410 associations between larger bacterial clades and fatty liver disease in the Finnish population.
411 This was accomplished by constructing predictive models to classify participants in the FLI
412 groups based on the phylogenetic balances and covariates, subjected to feature selection and
413 geographical cross-validation.
414
415 ***Predictive modeling of FLI reveals consistent associations between gram-positive***
416 ***Clostridia and fatty liver disease***
417 Strikingly, the strongest associations with FLI in our machine learning models were all inside
418 the *Firmicutes* phylum. A possible reason for this might be the higher relative abundance of
419 phylum *Firmicutes* at high latitudes,[52] where Finland is. Among the associations we identified,
420 *Faecalicatena gnavus* (NCBI: *Ruminococcus gnavus*) was positively linked with FLI as part of
421 3 predictive balances, and associated in previous studies with liver cirrhosis.[17] In their study,
422 oral *Firmicutes*, such as *Veillonella*, were suggested to invade the gut. While our balance-based
423 approach did not detect these taxa, *Megasphaera elsdenii* were positively associated with the

high FLI group in our differential abundance analyses (**Table S7**). This might be due to the strict feature selection employed prior to the predictive modeling.

Two individual taxa, *Negativibacillus sp000435195* and *Phocea massiliensis*, both had strong positive associations with the high FLI group (**Figure 2**), but the balances including these species were not predictive of the liver function-specific components (triglycerides and GGT). Positive 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.[53] While comparisons between these studies are difficult due to differences in taxa annotations, 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. Thus, further study on the connections of these bacteria with gut health and diet is warranted.

Among the taxa negatively associated with high FLI, *Hungatella* (see balance n332, **Figure 2**) have been previously shown to correlate negatively with the obesity phenotype in mice[54] and *H. hathewayi* was found to be a common commensal in the gut of healthy volunteers.[55] However, genus *Hungatella* has also been positively associated with concentrations of trimethylamine-N-oxide (TMAO),[56] a metabolite associated with cardiovascular disease and NAFLD. In our study, on the positively associated side (of balance n332) opposite to genus

447 *Hungatella* was a novel GTDB species, *CAG-81 sp000435795*, previously included in NCBI
448 genus *Clostridium*. The *CAG-81* genus was recently positively associated with TMAO levels in
449 urine in a study using the GTDB classification.[57] While we did not find the pathway for TMA
450 (precursor to TMAO) production in its genome, this would explain the positive association of
451 the *CAG-81* species with high FLI. Furthermore, the previous contradictive results among these
452 taxa could be explained by grouping of putatively TMA producing taxa in *CAG-81* together
453 with the closely related genus *Hungatella*.
454
455 Most taxa in our study with a positive association with FLI belonged to the broadly defined
456 *Clostridium* NCBI genus, which supports several previous observations.[14,46,58] However,
457 taxonomic standardization according to GTDB has identified the *Clostridium* genus as the most
458 phylogenetically inconsistent of all bacterial genera in the NCBI taxonomy, and divides it into
459 a total of 121 monophyletic genera in 29 distinct families.[28] The GTDB reassignment
460 complicates comparisons to previous studies, but it is phylogenetically and biologically
461 sensible, and can thus provide new insights into the microbiomes. Our results also strongly
462 suggest that despite its higher cost compared to metabarcoding, the increased resolution of
463 (shallow) shotgun metagenomic sequencing is highly useful in identifying specific taxon-
464 disease associations (see *e.g.,* refs. [27,59]).
465
466 ***Bacterial taxa positively associated with high FLI have a genetic potential to***
467 ***exacerbate the development of fatty liver disease***
468 We identified several plausible new associations between individual taxa and clades of bacteria
469 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 high 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,[60] and several of the taxa associated with the high FLI group have also been experimentally shown to produce ethanol, such as *C. M asparagiforme*, *C. M bolteae*, *C. M clostridioforme / C. M clostridioforme A* [61], and *C. Q Saccharolyticum*.[62] The relative abundances of these putatively ethanol-producing taxa were also 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 harbored genes predicted to encode for the L-glutamate fermentation V (P162-PWY; **Figure S8**) 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.[63] 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.[64] Interestingly, *C. Q saccharolyticum* (in our study, a taxa positively associated with high FLI deriving from balance n330), was experimentally shown to ferment various carbohydrates, including fructose, to acetate, hydrogen, carbon dioxide, and ethanol.[62] Furthermore, while our own pathway analysis did not detect BA modification pathways in the reference genome of *C. Q saccharolyticum*, a strain of this

493 species has been highlighted as a probable contributor to NAFLD development through the
494 synthesis of secondary BA.[15] The links between dietary intake and gene regulation, combined
495 with microbial fermentation in the gut warrant further mechanistic experiments to elucidate
496 their links with fatty liver, and likely other metabolic diseases.
497
498 NAFLD-associated ethanol producing bacteria in previous cohort studies have all been gram-
499 negatives, such as (NCBI-defined) *Klebsiella pneumoniae,*[13] and *Escherichia coli.*[12] In our
500 population sample, instead of gram-negatives, bacteria from the *C. M bolteae*, *C. M*
501 *clostridioforme / C. M clostridioforme A* and *C. M citroniae* species (positively associated with
502 high FLI in balance n336) have been described as opportunistic pathogens,[65] and are
503 hypothesized to exacerbate fatty liver development similarly through endogenous ethanol
504 production. This result suggests that geographical,[35] and ethnic variability,[66] might also
505 strongly affect gut microbiome composition and its associations with disease. In addition to
506 putative endogenous ethanol producers, we identified other taxa positively associated with high
507 FLI in balance n330, for which reference genomes harbored a genetic pathway predicted to
508 encode for the ability to ferment L-lysine to acetate and butyrate. While the production of these
509 SCFAs is often considered beneficial for gut health, other metabolism of proteolytic bacteria
510 might negatively contribute to fatty liver disease.[67]
511
512 Through modeling a previously validated index for fatty liver, FLI, we found replicable
513 associations with specific microbial taxa and likely liver disease of the participants. In addition,
514 sex and age of participants were also strongly predictive of the FLI group in our models
515 (**Figures 2, S7**). Their similar positive associations with fatty liver disease are known from

516 previous studies.[68,69] The associated microbial balances could be used to improve the
517 predictions above the baseline of these covariates on 5/6 regions in Finland in the main cohort.
518 For example, in the model cross-validated with Lapland the balances were more predictive of
519 FLI group than the covariates by themselves, and their combination increased the AUC further.
520 Yet, when testing the model where Turku/Loimaa region was used for internal cross-validation,
521 the microbial balances were slightly predictive of FLI group but failed to improve the AUC
522 over the covariates (**Table S2**). This pattern might stem from the cultural and genetic west-east
523 division in Finland,[70,71] with a closer proximity of the Helsinki/Vantaa region to eastern regions
524 than Turku/Loimaa, in both terms. It is thus likely that further incorporation and investigation
525 on the use of spatial information in microbiome modeling would elucidate these geographical
526 patterns in taxa-disease associations.
527
528 Our models were also able to accurately predict the FLI group of participants in the external
529 validation cohort, which were from the North Karelia and Helsinki/Vantaa regions. The
530 observed difficulty to geographically extrapolate taxa-disease associations[35] might mean that
531 associations reported in our study are specific to Finland and nearby regions. Notably, many of
532 the positive associations between specific taxa and fatty liver disease have not been reported
533 previously, but the functional potential of these taxa inferred from genomic data is similar to
534 taxa positively associated with NAFLD in previous studies. Thus, the geographical limits of
535 taxa-disease associations reported in studies such as ours warrant further study. Unfortunately,
536 generalization of our own results outside of Finland also remains to be addressed.
537

538 It is likely that not all associations in the current study are related solely to liver steatosis,
539 because FLI is based on measurements related to metabolic syndrome. However, our approach
540 is supported by recent views of NAFLD as the integral liver component of the metabolic
541 syndrome.[34,72] Indeed, the prevalences of diabetes and cardiovascular disease in both FINRISK
542 2002 and 2007 cohorts are elevated in the high FLI group, although the majority of the high
543 FLI participants did not have either of these diagnoses at the time of sampling (**Table S1**). We
544 also dissected the FLI by dividing participants into age/sex/BMI groups and detected microbial
545 groups specific to the blood work measurements of liver damage, triglycerides and GGT. These
546 associated taxa can thus be thought of as most closely associated with liver function, if such a
547 division is deemed practical.
548
549 ***Conclusions***
550 Modeling an established risk index for fatty liver enabled the detection of associations between
551 the disease and gut microbiome composition, to the level of individual taxa. While utilizing
552 FLI as a proxy, NCBI taxa identified with standard statistical methods were supportive of
553 previously reported differences between NAFLD cases and healthy controls. In our machine
554 learning framework, all clades robustly predictive of the FLI group were from the obligately
555 anaerobic gram-positive class *Clostridia*, representing several redefined GTDB genera
556 previously included in the NCBI genus *Clostridium*. Many of the representative genomes of
557 taxa positively associated with high FLI had genomic potential for endogenous ethanol
558 production. Our results support previous findings on the likely contribution of ethanol and
559 increased gut permeability on the induction of hepatic steatosis. Further support was also found
560 for the involvement TMA and SCFAs, especially acetate, in the likely pathophysiology of fatty

liver disease. Our models were able to predict the FLI group of participants in Finland across geographical regions and in an external Finnish cohort, showing that the associations are robust and generalizable in this population. Based on our results, mechanistic connections between specific microbes and fatty liver disease, and the geographical differences in such taxa-disease associations, should be addressed in further studies.

Materials and 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.[30] 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.

Details of the FINRISK cohorts analyzed in this study are included in the supplementary files (**Table S1**). Further details and sampling have also been extensively covered in previous publications (see refs. [24,73]). The Coordinating Ethics Committee of the Helsinki University Hospital District approved the study protocol for FINRISK 2002 (Ref. 558/E3/2001), and all participants have given their written informed consent.

Stool DNA extraction and shallow shotgun metagenome sequencing

DNA extraction was performed according to the Earth Microbiome Project protocols, with the MagAttract PowerSoil DNA kit (Qiagen), as previously described.[74] A miniaturized version of the Kapa HyperPlus Illumina-compatible library prep kit (Kapa Biosystems) was used for library generation, following the previously published protocol.[75] 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,[76] 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.

606 ***Taxonomic matching and phylogenetic transforms***

607 We quality trimmed the sequences and removed the sequencing adapters with Atropos.[77] Host
608 reads were removed by mapping the reads against the human genome assembly GRCh38 with
609 Bowtie2.[78] To improve the taxonomic assignments of our reads, we used a custom index,[79] based
610 on the Genome Taxonomy Database (GTDB) release 89 taxonomic redefinitions,[28,80] for read
611 classification with default parameters in Centrifuge 1.0.4.[81] Viral and eukaryotic sequences were
612 removed in this step, as the database contains only bacterial and archaeal reference genomes.
613 After read classification, all following steps were performed with R version 3.5.2,[82] using
614 phyloseq 1.30.0[83] to manage the data. To reduce the number of spurious read assignments, and to
615 facilitate more accurate phylogenetic transformations, only reads classified at the species level,
616 matching individual GTDB reference genomes, were retained. Samples with less than 50,000
617 reads, from pregnant participants or recorded antibiotic use in the past 6 months were removed,
618 resulting in a final number of 6,269 samples. We first filtered taxa not seen with more than 3
619 counts in at least 1% of samples and those with a coefficient of variation ≤ 3 across all samples,
620 following McMurdie and Holmes[83], with a slight adaption from 20% of samples to 1% of
621 samples, because of our larger sample size. The complete bacterial and archaeal phylogenetic
622 trees of the GTDB release 89 reference genomes, constructed from an alignment of 120 bacterial
623 or 122 archaeal marker genes,[28] were then combined with our taxa tables. The resulting trees
624 were thus subset only to species which were observed in at least one sample in our data. The read
625 counts were transformed to phylogenetic node balances in both trees with PhILR.[29] The default
626 method for PhILR inputs a pseudocount of 1 for taxa absent in an individual sample before the
627 transform.
628

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,[29] 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,[84] 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.[85] 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.

To further validate our approach, assess how the use of the GTDB taxonomic redefinitions and custom database affected our results, and to facilitate comparisons with previous results, we annotated our raw reads in FINRISK 2002 samples also with NCBI taxonomy and performed

several additional analyses. For this comparison data, after quality trimming the FINRISK 2002 reads and removing host sequences as described above, SHOGUN v1.0.5[59] was used for taxonomy assignments against the NCBI RefSeq version 82 (May 8, 2017) database containing complete bacterial, archaeal, and viral genomes. To facilitate comparisons between different annotations, we subset the samples included in the SHOGUN/NCBI annotated data to those included in the Centrifuge/GTDB data (for exclusion criteria, see above).

Covariates

Because fatty liver disease is underdiagnosed at the population level,[25] and our sampling did not have extensive coverage of liver fat measurements, we chose to use the Fatty Liver index as a proxy for fatty liver.[26] Furthermore, the index performs well in cohorts of Caucasian ethnicity, such as ours, to diagnose the presence of NAFLD.[23] We calculated FLI after Bedogni et al.[26]:

$$(e^{0.953*\log_e(\text{triglycerides mg/dL}) + 0.139*\text{BMI} + 0.718*\log_e(\text{GGT}) + 0.053*\text{waist circumference} - 15.745}) / (1 + e^{0.953*\log_e(\text{triglycerides mg/dL}) + 0.139*\text{BMI} + 0.718*\log_e(\text{GGT}) + 0.053*\text{waist circumference} - 15.745}) * 100$$

. We chose the cutoff at FLI $\geq$ 60 to identify participants likely to be diagnosed with hepatic steatosis (positive likelihood ratio = 4.3 and negative likelihood ratio = 0.5, after Bedogni et al.[26]). 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 in FINRISK 2002 (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

675 balances, GGT, triglycerides and BMI were not imputed but observations where any of these
676 were missing were simply removed.
677
678 ***Taxa composition and alpha diversity***
679 The baseline compositions of the microbial communities in the samples were summarized at
680 phylum level in the different FLI groups (< 60 and ≥ 60 FLI) with the Centrifuge/GTDB data in
681 FINRISK 2002 and 2007 (**Figure S1**), and with SHOGUN/NCBI data in FINRISK 2002 (**Figure**
682 **S2**) by total sum scaling and merging taxa at phylum level, separately for bacteria and archaea.
683
684 Bacterial alpha diversity of each individual sample in FINRISK 2002 was estimated through
685 Shannon diversity as the mean of 10 random rarefactions of raw annotated read counts (see ref.
686 [86]), separately in both the Centrifuge/GTDB and SHOGUN/NCBI data sets. Associations
687 between the FLI group (< 60 or ≥ 60 FLI) and Shannon diversity in the data sets was modeled
688 using binomial regression and adjusted for age, sex, and self-reported alcohol use, using "glm" in
689 base R.[82]
690
691 ***Beta diversity and linear modeling of FLI***
692 In the Centrifuge/GTDB data, beta diversity was calculated as Euclidian distance of the PhILR
693 balances through Principal Component Analysis (PCA) on bacterial and archaeal balances
694 separately with "rda" in vegan 2.5.6.[87] To calculate beta diversity with the SHOGUN/GTDB
695 data, raw bacterial taxa counts were centered log-ratio (CLR) transformed with "transform" in
696 microbiome 1.8.0,[88] and their Euclidian distances were obtained similarly through PCA. Linear
697 regression models were constructed for FLI with "lm" in base R[82] with Centrifuge/GTDB data,

and separately with SHOGUN/NCBI. Log_{10}(FLI) was used as the dependent variable and the first three bacterial PCs, sex, age, and self-reported alcohol were used as the independent variables. Archaeal PCs were not included in the models because none of them were significantly correlated with FLI in Centrifuge/GTDB data (all $P > 0.001$). To visualize the association between beta diversity and FLI, the FLI of each participant was plotted against its quantiles along the three bacterial PC axes in Centrifuge/GTDB data (**Figure 1C**). A comparison of the associations with the alternative SHOGUN/NCBI annotated data was also included in the SI (**Figure S3**).

Differential abundance of individual taxa between the FLI groups

To facilitate comparisons to previous studies, we assessed the associations between the FLI group (< 60 or ≥ 60 FLI) of the participants and Centrifuge/GTDB and SHOGUN/NCBI annotated individual taxa present in the samples. With both data sets, differential abundance of the bacterial taxa between the FLI groups was assessed with the ALDEx2 compositional data analysis tool.[89] Briefly, significance of the abundance differences between the groups were estimated with a Welch's t-test, and only taxa with (Benjamini Hochberg) false discovery rate - adjusted P values (or Q values) < 0.001 were retained. The associated taxa were then divided in each data set to those positively or negatively associated with the high FLI group and sorted based on effect sizes estimated from the median CLR differences between the groups.

FLI modeling within a machine learning framework

In the machine learning framework, both regression and categorical models were constructed for FLI, using only the Centrifuge/GTDB data. The feature selection, hyperparameter optimization

and internal 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,[90] 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).[26] The data was first split to 70% train and 30% test sets while preserving sex and region balance. To take into account geographical differences (see *e.g.,* ref. [35]) 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 (3,423 predictor features) feature selection by filtering was first performed inside the training data, based on random forest permutation as recommended by Bommert et al.[91] 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,[92] 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).

744

745 Bayesian hyperparameter optimization of the xgboost models was then performed with only the
746 selected features. An optimal set of parameters for the xgboost models were searched over all
747 LOGOCV subsets with "mbo" in mlrMBO 1.1.3,[93] using 30 preliminary rounds with randomized
748 parameters, followed by 100 optimization rounds. Parameters in the xgboost models and their
749 considered ranges were learning rate (eta) [0.001, 0.3], gamma [0.1, 5], maximum depth of a tree
750 [2, 8], minimum child weight [1, 10], fraction of data subsampled per each iteration [0.2, 0.8],
751 fraction of columns subsampled per tree [0.2, 0.9], and maximum number of iterations (nrounds)
752 [50, 5000]. The parameters recommended by these searchers were as following for regression:
753 eta=0.00889; gamma=2.08; max_depth=2; min_child_weight=8; subsample=0.783;
754 colsample_bytree=0.672; nrounds=1,810, and for classification: eta=0.00107; gamma=0.137;
755 max_depth=5; min_child_weight=9; subsample=0.207; colsample_bytree=0.793;
756 nrounds=4,328. We used Root-Mean-Square Error (RMSE) for the regression models and Area
757 Under the ROC Curve (AUC) for the classification models to measure model fit on the left-out
758 data (region) in each LOGOCV subset. Receiver operating characteristic and precision-recall
759 curves for these validation metrics were calculated with "evalmod" in precrec v0.11.2.[94] The
760 final models were trained on the LOGOCV subset training data, the data from one region thus
761 omitted per model, and using the selected features and optimized hyperparameters. Internal
762 validation of these models was conducted against participants only from the region omitted from
763 each model, in the 30% test data which was not used in model training or optimization.
764 Sensitivity analysis was conducted by using only the predictive covariates (sex and age) or
765 balances separately, with the same hyperparameters, data partitions and cross-region internal
766 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,[26] 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.[95] 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.[96]

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=3,119.

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 $\geq$ 60 and FLI < 60 groups, where all validation region samples also in the training data were excluded from the simple intercept of top 50 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=1,772.

External validation of the models in a separate population cohort

To further validate our models and results, we leveraged the data from a more recent population cohort in Finland, FINRISK 2007 (see **Table S1**). In this cohort, the choice of participants, sample collection, and related methods for the data used in the current study were similar to FINRISK 2002 to facilitate inter-cohort comparisons, and are reported elsewhere.[30] The study

protocol of FINRISK 2007 was approved by the Coordinating Ethical Committee of the Hospital District of Helsinki and Uusimaa (Ref. 229/EO/2006). All participants have signed an informed consent.

Briefly, compared to FINRISK 2002, FINRISK 2007 features a smaller number of participants who donated fecal samples ($N = 258$ after excluding pregnant individuals or antibiotic use in the last 6 months), they were younger on average, and a smaller proportion of them were in the high FLI group. To produce data for the validation, methods and quality control related to DNA extraction, sequencing, taxonomic assignments, and calculation of FLI values were identical to FINRISK 2002 data, as described above. For the phylogenetic transform (performed otherwise identically), only taxa passing the filtering in FINRISK 2002 bacterial data set were retained in FINRISK 2007 and a pseudo-count of 1 was used for taxa unobserved in the new data, to exactly match the node balance names. The FINRISK 2007 data was then subset to the model features of the main classification models (sex, age, and the 11 bacterial balances), and input in each of the 6 LOGOCV classification models. The results of these predictions were then compared against the true FLI groups (FLI $\geq$ 60 and FLI < 60) of the participants (**Table S5**). Receiver operating characteristic and precision-recall curves for the external validation were calculated similarly to the main models for the AUC and AUPRC metrics and plotted after averaging the predictions of the 6 models to obtain single curves (**Figure S6**).

Identification of predictive features specific to liver function

Because the FLI also incorporates BMI and waist circumference, and they strongly contribute to the index,[26] we deemed it necessary to further investigate which of the identified balances were

836 specific to liver function. The participants were first grouped by age (< 40, 40 – 60, and 60 <),
837 sex (female or male) and BMI (< 25, 25 – 30, and 30 <) into 18 categories (N = 105 ~ 711 per
838 category). We performed feature selection similarly to the FLI models by fitting random forest
839 regressors for GGT and triglycerides with mlr 2.16.0.[92] This was done separately in each of the
840 18 categories, and in each category, we again used LOGOCV with the regions to obtain 6 runs
841 per category. Finally, the features predictive of GGT or triglycerides in each category were
842 selected as the intersect of top 50 features in the 6 LOGOCV iterations by permutation
843 importance. The intersect of features predictive of GGT or triglycerides in any of the categories
844 and the features predictive of categorical FLI were identified as specific to liver function.
845
846 ***Pathway inference for taxa associated with FLI***
847 Our taxonomic matching of the reads is based on the genomes of GTDB (release 89),[28] which are
848 all complete or nearly complete and available in online databases. This enables us to estimate the
849 likely genetic content, and thus, the metabolic potential of the microbes associated with FLI. We
850 use this approach because the sequencing depth of our samples does not allow assembling
851 contigs and (metagenome-assembled) genomes, required for pathway predictions. Because of the
852 compositional phylogenetic transform, among other features of our data, previously developed
853 approaches such as PICRUSt,[97] could not be used here.
854
855 The genomes of all 336 bacteria under at least one of the predictive balances were downloaded
856 from NCBI. 119 of these genomes were originally not annotated, which is a requirement for
857 pathway prediction. Therefore, Prokka v1.14.6,[98] was used to annotate the 119 unannotated
858 genomes as a preliminary step. Pathway predictions were then performed for all 336 genomes

with mpwt v0.5.3 multiprocessing tool,[99] for the PathoLogic pipeline of Pathway Tools 23.0.[100] Pathways for ethanol and short chain fatty acid (acetate, butyrate, propionate) production, bile acid metabolism, and choline degradation to trimethylamine were identified from MetaCyc pathway classifications (see ref. [101], and **Table S4**). The prevalence of these processes was then assessed in the analyzed genomes and summarized per process to consider the possible links of the taxa with fatty liver pathophysiology. Finally, the presence of individual pathways for acetate and ethanol production was also outlined for each genome.

Data availability statement

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

Disclosure of interest

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-

882 Ingelheim, Bristol-Myers Squibb, Cirius, Eli Lilly and Company, Galectin Therapeutics, Galmed
883 Pharmaceuticals, GE, Genfit, Gilead, Intercept, Grail, Janssen, Madrigal Pharmaceuticals,
884 Merck, NGM Biopharmaceuticals, NuSirt, Pfizer, pH Pharma, Prometheus, and Siemens. He is
885 also co-founder of Liponexus, Inc.
886

887 ***Funding details***

888 This research was supported in part by grants from the Finnish Foundation for Cardiovascular
889 Research, the Emil Aaltonen Foundation, the Paavo Nurmi Foundation, the Urmas Pekkala
890 Foundation, the Finnish Medical Foundation, the Sigrid Juselius Foundation, the Academy of
891 Finland (#321356 to A.H.; #295741, #307127 to L.L.; #321351 to T.N.) and the National
892 Institutes of Health (R01ES027595 to M.J.). R.L. receives funding support from NIEHS
893 (5P42ES010337), NCATS (5UL1TR001442), NIDDK (U01DK061734, R01DK106419,
894 P30DK120515, R01DK121378, R01DK124318), and DOD PRCRP (W81XWH-18-2-0026).
895 Additional support was provided by Illumina, Inc. and Janssen Pharmaceutica through their
896 sponsorship of the Center for Microbiome Innovation at UCSD.
897

898 ***Authors' contributions***

899 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
900 and R.K. acquired the data. M.R., L.L. and T.N. analyzed the data. M.R. wrote the manuscript in
901 consultation with all authors. M.I., P.J., V.S., R.K., L.L. and T.N. supervised the work. All
902 authors gave final approval of the version to be published.
903

Acknowledgements

We thank all participants of the FINRISK 2002 and FINRISK 2007 surveys for their contributions to this work, and Tara Schwartz for assistance with laboratory work. We also thank the editor and both anonymous reviewers for their constructive criticism.

References

1. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease—Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology 2016; 64:73–84.

2. Marchesini G, Bugianesi E, Forlani G, Cerrelli F, Lenzi M, Manini R, Natale S, Vanni E, Villanova N, Melchionda N, et al. Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. Hepatology 2003; 37:917–23.

3. Chalasani N, Younossi Z, Lavine JE, Diehl AM, Brunt EM, Cusi K, Charlton M, Sanyal AJ. 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 2012; 55:2005–23.

4. Yki-Järvinen H. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. Lancet Diabetes Endocrinol 2014; 2:901–10.

5. Toshikuni N, Tsutsumi M, Arisawa T. Clinical differences between alcoholic liver disease and nonalcoholic fatty liver disease. World J Gastroenterol 2014; 20:8393–406.

6. Rinella M, Charlton M. The globalization of nonalcoholic fatty liver disease: Prevalence and impact on world health. Hepatology 2016; 64:19–22.

7. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med 2018; 24:392–400.

8. Caussy C, Tripathi A, Humphrey G, Bassirian S, Singh S, Faulkner C, Bettencourt R, Rizo E, Richards L, Xu ZZ, et al. A gut microbiome signature for cirrhosis due to nonalcoholic fatty liver disease. Nature Communications 2019; 10:1406.

9. Compare D, Coccoli P, Rocco A, Nardone OM, De Maria S, Cartenì M, Nardone G. Gut–liver axis: The impact of gut microbiota on non alcoholic fatty liver disease. Nutrition, Metabolism and Cardiovascular Diseases 2012; 22:471–6.

10. Miele L, Valenza V, Torre GL, Montalto M, Cammarota G, Ricci R, Mascianà R, Forgione A, Gabrieli ML, Perotti G, et al. Increased intestinal permeability and tight junction alterations in nonalcoholic fatty liver disease. Hepatology 2009; 49:1877–87.

11. Spencer MD, Hamp TJ, Reid RW, Fischer LM, Zeisel SH, Fodor AA. Association Between Composition of the Human Gastrointestinal Microbiome and Development of Fatty Liver With Choline Deficiency. Gastroenterology 2011; 140:976–86.

12. Zhu L, Baker SS, Gill C, Liu W, Alkhouri R, Baker RD, Gill SR. Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between endogenous alcohol and NASH. Hepatology 2013; 57:601–9.

13. Yuan J, Chen C, Cui J, Lu J, Yan C, Wei X, Zhao X, Li N, Li S, Xue G, et al. Fatty Liver Disease Caused by High-Alcohol-Producing Klebsiella pneumoniae. Cell Metabolism 2019; 30:675-688.e7.

14. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, Dulai PS, Caussy C, Bettencourt R, Highlander SK, et al. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metabolism 2017; 25:1054-1062.e5.

15. Jiao N, Wu D, Yang Z, Fang S, Li X, Yuan M, Zhu R, Zhu L. Gut bacteria contributes to NAFLD pathogenesis by promoting secondary bile acids biosynthesis. The FASEB Journal 2019; 33:126.4-126.4.

16. Aron-Wisnewsky J, Vigliotti C, Witjes J, Le P, Holleboom AG, Verheij J, Nieuwdorp M, Clément K. Gut microbiota and human NAFLD: disentangling microbial signatures from metabolic disorders. Nature Reviews Gastroenterology & Hepatology 2020; 17:279–97.

17. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature 2014; 513:59–64.

18. Liu Y, Meric G, Havulinna AS, Teo SM, Ruuskanen M, Sanders J, Zhu Q, Tripathi A, Verspoor K, Cheng S, et al. Early prediction of liver disease using conventional risk factors and gut microbiome-augmented gradient boosting [Internet]. Genetic and Genomic Medicine; 2020 [cited 2020 Jul 28]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.06.24.20138933

19. Safari Z, Gérard P. The links between the gut microbiome and non-alcoholic fatty liver disease (NAFLD). Cell Mol Life Sci 2019; 76:1541–58.

20. Carpino G, Del Ben M, Pastori D, Carnevale R, Baratta F, Overi D, Francis H, Cardinale V, Onori P, Safarikia S, et al. Increased liver localization of lipopolysaccharides in human and experimental non-alcoholic fatty liver disease. Hepatology 2019; :hep.31056.

21. Li Q, Dhyani M, Grajo JR, Sirlin C, Samir AE. Current status of imaging in nonalcoholic fatty liver disease. World J Hepatol 2018; 10:530–42.

22. Koehler EM, Schouten JNL, Hansen BE, Hofman A, Stricker BH, Janssen HLA. External Validation of the Fatty Liver Index for Identifying Nonalcoholic Fatty Liver Disease in a Population-based Study. Clinical Gastroenterology and Hepatology 2013; 11:1201–4.

23. Vanni E, Bugianesi E. Editorial: utility and pitfalls of Fatty Liver Index in epidemiologic studies for the diagnosis of NAFLD. Aliment Pharmacol Ther 2015; 41:406–7.

24. Salosensaari A, Laitinen V, Havulinna AS, Meric G, Cheng S, Perola M, Valsta L, Alfthan G, Inouye M, Watrous JD, et al. Taxonomic Signatures of Long-Term Mortality Risk in Human Gut Microbiota [Internet]. Epidemiology; 2020 [cited 2020 Jan 4]. Available from: http://medrxiv.org/lookup/doi/10.1101/2019.12.30.19015842

25. Alexander M, Loomis AK, Fairburn-Beech J, van der Lei J, Duarte-Salles T, Prieto-Alhambra D, Ansell D, Pasqua A, Lapi F, Rijnbeek P, et al. Real-world data reveal a diagnostic gap in non-alcoholic fatty liver disease. BMC Medicine 2018; 16:130.

26. Bedogni G, Bellentani S, Miglioli L, Masutti F, Passalacqua M, Castiglione A, Tiribelli C. The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol 2006; 6:33.

27. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, Knight R, Knights D. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems [Internet] 2018 [cited 2020 Apr 9]; 3. Available from: https://msystems.asm.org/content/3/6/e00069-18

28. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 2018; 36:996–1004.

29. Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. eLife 2017; 6:e21887.

30. Borodulin K, Tolonen H, Jousilahti P, Jula A, Juolevi A, Koskinen S, Kuulasmaa K, Laatikainen T, Männistö S, Peltonen M, et al. Cohort Profile: The National FINRISK Study. International Journal of Epidemiology 2018; 47:696–696i.

31. Banderas DZ, Escobedo J, Gonzalez E, Liceaga MG, Ramírez JC, Castro MG. γ-Glutamyl transferase: a marker of nonalcoholic fatty liver disease in patients with the metabolic syndrome. European Journal of Gastroenterology & Hepatology 2012; 24:805–810.

32. Haas JT, Francque S, Staels B. Pathophysiology and Mechanisms of Nonalcoholic Fatty Liver Disease. Annual Review of Physiology 2016; 78:181–205.

33. Castaner O, Goday A, Park Y-M, Lee S-H, Magkos F, Shiow S-ATE, Schröder H. The Gut Microbiome Profile in Obesity: A Systematic Review. Int J Endocrinol [Internet] 2018 [cited 2020 Apr 3]; 2018. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5933040/

34. Eslam M, Sanyal AJ, George J, Sanyal A, Neuschwander-Tetri B, Tiribelli C, Kleiner DE, Brunt E, Bugianesi E, Yki-Järvinen H, et al. MAFLD: A Consensus-Driven Proposed Nomenclature for Metabolic Associated Fatty Liver Disease. Gastroenterology 2020; 158:1999-2014.e1.

35. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, Chen M-X, Chen Z-H, Ji G-Y, Zheng Z-D-X, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nature Medicine 2018; 24:1532–5.

36. Carvalhana S, Leitão J, Alves AC, Bourbon M, Cortez-Pinto H. How good is controlled attenuation parameter and fatty liver index for assessing liver steatosis in general population: correlation with ultrasound. Liver International 2014; 34:e111–7.

37. Dehnavi Z, Razmpour F, Naseri MB, Nematy M, Alamdaran SA, Vatanparast HA, Nezhad MA, Abbasi B, Ganji A. Fatty liver index (FLI) in predicting non-alcoholic fatty liver disease (NAFLD). Hepatitis Monthly 2018; 18.

38. Yang B-L, Wu W-C, Fang K-C, Wang Y-C, Huo T-I, Huang Y-H, Yang H-I, Su C-W, Lin H-C, Lee F-Y, et al. External Validation of Fatty Liver Index for Identifying Ultrasonographic Fatty Liver in a Large-Scale Cross-Sectional Study in Taiwan. PLOS ONE 2015; 10:e0120443.

39. Huang X, Xu M, Chen Y, Peng K, Huang Y, Wang P, Ding L, Lin L, Xu Y, Chen Y, et al. Validation of the Fatty Liver Index for Nonalcoholic Fatty Liver Disease in Middle-Aged and Elderly Chinese. Medicine (Baltimore) 2015; 94:e1682.

40. Astbury S, Atallah E, Vijay A, Aithal GP, Grove JI, Valdes AM. Lower gut microbiome diversity and higher abundance of proinflammatory genus Collinsella are associated with biopsy-proven nonalcoholic steatohepatitis. Gut Microbes 2020; 11:569–80.

41. Kim H-N, Joo E-J, Cheong HS, Kim Y, Kim H-L, Shin H, Chang Y, Ryu S. Gut Microbiota and Risk of Persistent Nonalcoholic Fatty Liver Diseases. Journal of Clinical Medicine 2019; 8:1089.

42. de la Cuesta-Zuluaga J, Corrales-Agudelo V, Velásquez-Mejía EP, Carmona JA, Abad JM, Escobar JS. Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization. Scientific Reports 2018; 8:11356.

43. O'Callaghan A, van Sinderen D. Bifidobacteria and Their Role as Members of the Human Gut Microbiota. Front Microbiol [Internet] 2016 [cited 2020 Nov 9]; 7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4908950/

44. Waters JL, Ley RE. The human gut bacteria Christensenellaceae are widespread, heritable, and associated with health. BMC Biology 2019; 17:83.

45. Ferreira-Halder CV, Faria AV de S, Andrade SS. Action and function of Faecalibacterium prausnitzii in health and disease. Best Pract Res Clin Gastroenterol 2017; 31:643–8.

46. Guohong-Liu, Qingxi-Zhao, Hongyun-Wei. Characteristics of intestinal bacteria with fatty liver diseases and cirrhosis. Annals of Hepatology 2019; 18:796–803.

47. Raman M, Ahmed I, Gillevet PM, Probert CS, Ratcliffe NM, Smith S, Greenwood R, Sikaroodi M, Lam V, Crotty P, et al. Fecal Microbiome and Volatile Organic Compound Metabolome in Obese Humans With Nonalcoholic Fatty Liver Disease. Clinical Gastroenterology and Hepatology 2013; 11:868-875.e3.

48. Dong TS, Katzka W, Lagishetty V, Luu K, Hauer M, Pisegna J, Jacobs JP. A Microbial Signature Identifies Advanced Fibrosis in Patients with Chronic Liver Disease Mainly Due to NAFLD. Scientific Reports 2020; 10:2771.

49. Bajaj JS, Idilman R, Mabudian L, Hood M, Fagan A, Turan D, White MB, Karakaya F, Wang J, Atalay R, et al. Diet affects gut microbiota and modulates hospitalization risk differentially in an international cirrhosis cohort. Hepatology 2018; 68:234–47.

50. Bowerman KL, Rehman SF, Vaughan A, Lachner N, Budden KF, Kim RY, Wood DLA, Gellatly SL, Shukla SD, Wood LG, et al. Disease-associated gut microbiome and metabolome changes in patients with chronic obstructive pulmonary disease. Nature Communications 2020; 11:5886.

51. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, et al. Human Genetics Shape the Gut Microbiome. Cell 2014; 159:789–99.

52. Suzuki TA, Worobey M. Geographical variation of human gut microbial composition. Biology Letters 2014; 10:20131037.

53. Meslier V, Laiola M, Roager HM, Filippis FD, Roume H, Quinquis B, Giacco R, Mennella I, Ferracane R, Pons N, et al. Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake. Gut [Internet] 2020 [cited 2020 Jun 2]; Available from: https://gut.bmj.com/content/early/2020/02/18/gutjnl-2019-320438

54. Cui C, Li Y, Gao H, Zhang H, Han J, Zhang D, Li Y, Zhou J, Lu C, Su X. Modulation of the gut microbiota by the mixture of fish oil and krill oil in high-fat diet-induced obesity mice. PLOS ONE 2017; 12:e0186216.

55. Manzoor SE, McNulty CAM, Nakiboneka-Ssenabulya D, Lecky DM, Hardy KJ, Hawkey PM. Investigation of community carriage rates of Clostridium difficile and Hungatella hathewayi in healthy volunteers from four regions of England. Journal of Hospital Infection 2017; 97:153–5.

56. Genoni A, Christophersen CT, Lo J, Coghlan M, Boyce MC, Bird AR, Lyons-Wall P, Devine A. Long-term Paleolithic diet is associated with lower resistant starch intake, different gut microbiota composition and increased serum TMAO concentrations. Eur J Nutr 2020; 59:1845–58.

57. Burton KJ, Krüger R, Scherz V, Münger LH, Picone G, Vionnet N, Bertelli C, Greub G, Capozzi F, Vergères G. Trimethylamine-N-Oxide Postprandial Response in Plasma and Urine Is Lower After Fermented Compared to Non-Fermented Dairy Consumption in Healthy Adults. Nutrients 2020; 12:234.

58. Jiang W, Wu N, Wang X, Chi Y, Zhang Y, Qiu X, Hu Y, Li J, Liu Y. Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease. Sci Rep 2015; 5:8096.

59. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Knight R, Knights D. SHOGUN: a modular, accurate, and scalable framework for microbiome quantification. Bioinformatics 2020; :btaa277.

60. de Faria Ghetti F, Oliveira DG, de Oliveira JM, de Castro Ferreira LEVV, Cesar DE, Moreira APB. Influence of gut microbiota on the development and progression of nonalcoholic steatohepatitis. Eur J Nutr 2018; 57:861–76.

61. Mohan R, Namsolleck P, Lawson PA, Osterhoff M, Collins MD, Alpert C-A, Blaut M. Clostridium asparagiforme sp. nov., isolated from a human faecal sample. Systematic and Applied Microbiology 2006; 29:292–9.

62. Murray WD, Khan AW, van den BERG L. Clostridium saccharolyticum sp. nov., a Saccharolytic Species from Sewage Sludge. International Journal of Systematic Bacteriology 1982; 32:132–5.

63. Diether NE, Willing BP. Microbial Fermentation of Dietary Protein: An Important Factor in Diet–Microbe–Host Interaction. Microorganisms [Internet] 2019 [cited 2020 Apr 24]; 7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6352118/

64. Zhao S, Jang C, Liu J, Uehara K, Gilbert M, Izzo L, Zeng X, Trefely S, Fernandez S, Carrer A, et al. Dietary fructose feeds hepatic lipogenesis via microbiota-derived acetate. Nature 2020; 579:586–91.

65. Dehoux P, Marvaud JC, Abouelleil A, Earl AM, Lambert T, Dauga C. Comparative genomics of Clostridium bolteae and Clostridium clostridioforme reveals species-specific genomic properties and numerous putative antibiotic resistance determinants. BMC Genomics 2016; 17:819.

66. Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, Tremaroli V, Bakker GJ, Attaye I, Pinto-Sietsma S-J, et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nature Medicine 2018; 24:1526–31.

67. Canfora EE, Meex RCR, Venema K, Blaak EE. Gut microbial metabolites in obesity, NAFLD and T2DM. Nature Reviews Endocrinology 2019; 15:261–73.

68. Cheng H-Y, Wang H-Y, Chang W-H, Lin S-C, Chu C-H, Wang T-E, Liu C-C, Shih S-C. Nonalcoholic Fatty Liver Disease: Prevalence, Influence on Age and Sex, and Relationship

with Metabolic Syndrome and Insulin Resistance. International Journal of Gerontology 2013; 7:194–8.

69. Lonardo A, Nascimbeni F, Ballestri S, Fairweather D, Win S, Than TA, Abdelmalek MF, Suzuki A. Sex Differences in Nonalcoholic Fatty Liver Disease: State of the Art and Identification of Research Gaps. Hepatology 2019; 70:1457–69.

70. Näyhä S. Geographical variations in cardiovascular mortality in Finland, 1961-1985. Scand J Soc Med Suppl 1989; 40:1–48.

71. Kerminen S, Havulinna AS, Hellenthal G, Martin AR, Sarin A-P, Perola M, Palotie A, Salomaa V, Daly MJ, Ripatti S, et al. Fine-Scale Genetic Structure in Finland. G3: Genes, Genomes, Genetics 2017; 7:3459–68.

72. Reccia I, Kumar J, Akladios C, Virdis F, Pai M, Habib N, Spalding D. Non-alcoholic fatty liver disease: A sign of systemic disease. Metabolism 2017; 72:94–108.

73. Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T, Männistö S, Salomaa V, Sundvall J, Puska P. Forty-year trends in cardiovascular risk factors in Finland. Eur J Public Health 2015; 25:539–46.

74. Marotz L, Schwartz T, Thompson L, Humphrey G, Gogul G, Gaffney J, Amir A, Knight R. Earth Microbiome Project (EMP) high throughput (HTP) DNA extraction protocol v1 (protocols.io.pdmdi46) [Internet]. 2018 [cited 2020 Nov 10]; Available from: https://www.protocols.io/view/earth-microbiome-project-emp-high-throughput-htp-d-pdmdi46

75. Sanders JG, Nurk S, Salido RA, Minich J, Xu ZZ, Zhu Q, Martino C, Fedarko M, Arthur TD, Chen F, et al. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. Genome Biology 2019; 20:226.

76. Glenn TC, Nilsen RA, Kieran TJ, Sanders JG, Bayona-Vásquez NJ, Finger JW, Pierson TW, Bentley KE, Hoffberg SL, Louha S, et al. Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). PeerJ 2019; 7:e7755.

77. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ [Internet] 2017 [cited 2020 Nov 10]; 5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5581536/

78. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012; 9:357–9.

79. Méric G, Wick RR, Watts SC, Holt KE, Inouye M. Correcting index databases improves metagenomic studies. bioRxiv 2019; :712166.

80. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology 2020; :1–8.

81. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res [Internet] 2016 [cited 2018 May 12]; Available from: http://genome.cshlp.org/content/early/2016/11/16/gr.210641.116

82. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018 [cited 2019 Mar 4]. Available from: https://www.R-project.org/

83. McMurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 2013; 8:e61217.

84. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol [Internet] 2017 [cited 2020 Jul 20]; 8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5695134/

85. Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, Ackermann M, Hahn AS, Srivastava DS, Crowe SA, et al. Function and functional redundancy in microbial systems. Nat Ecol Evol 2018; 2:936–43.

86. Ruuskanen MO, St. Pierre KA, St. Louis VL, Aris-Brosou S, Poulain AJ. Physicochemical Drivers of Microbial Community Structure in Sediments of Lake Hazen, Nunavut, Canada. Front Microbiol [Internet] 2018 [cited 2018 Jun 6]; 9. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2018.01138/full

87. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, et al. vegan: Community Ecology Package [Internet]. 2018 [cited 2018 Jun 4]. Available from: https://CRAN.R-project.org/package=vegan

88. Lahti L, Shetty S. microbiome R package [Internet]. 2019 [cited 2020 Dec 14]. Available from: http://microbiome.github.io

89. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2014; 2:15.

90. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 2016; :785–94.

91. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis 2020; 143:106839.

92. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones ZM. mlr: Machine Learning in R. Journal of Machine Learning Research 2016; 17:1–5.

1188 93. Bischl B, Richter J, Bossek J, Horn D, Thomas J, Lang M. mlrMBO: A Modular
1189 Framework for Model-Based Optimization of Expensive Black-Box Functions.
1190 arXiv:170303373 [stat] [Internet] 2018 [cited 2020 Feb 18]; Available from:
1191 http://arxiv.org/abs/1703.03373

1192 94. Saito T, Rehmsmeier M. Precrec: fast and accurate precision–recall and ROC curve
1193 calculations in R. Bioinformatics 2017; 33:145–7.

1194 95. Greenwell BM. pdp: An R Package for Constructing Partial Dependence Plots. The R
1195 Journal 2017; 9:421–36.

1196 96. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and
1197 annotation of phylogenetic trees with their covariates and other associated data. Methods in
1198 Ecology and Evolution 2017; 8:28–36.

1199 97. Douglas GM, Maffei VJ, Zaneveld J, Yurgel SN, Brown JR, Taylor CM, Huttenhower C,
1200 Langille MGI. PICRUSt2: An improved and extensible approach for metagenome
1201 inference. bioRxiv 2019; :672295.

1202 98. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014; 30:2068–
1203 9.

1204 99. Belcour A, Frioux C, Aite M, Bretaudeau A, Siegel A. Metage2Metabo: metabolic
1205 complementarity applied to genomes of large-scale microbiotas for the identification of
1206 keystone species. bioRxiv 2019; :803056.

1207 100. Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, Ong
1208 WK, Subhraveti P, Caspi R, Fulcher C, et al. Pathway Tools version 23.0 update: software
1209 for pathway/genome informatics and systems biology. Briefings in Bioinformatics
1210 2019; :bbz104.

1211 101. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse
1212 M, Midford PE, Ong Q, Ong WK, et al. The MetaCyc database of metabolic pathways and
1213 enzymes. Nucleic Acids Res 2018; 46:D633–9.

1214

Figures

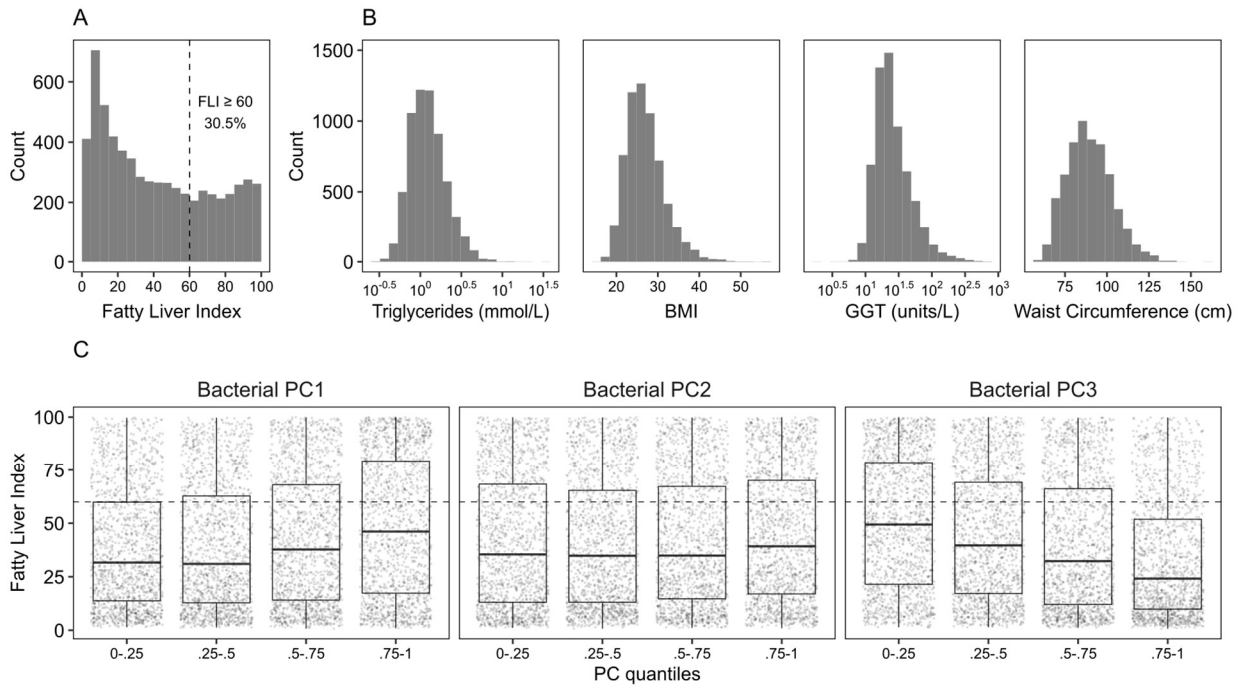

Figure 1. Relative of FLI (A), its components (B), and FLI in quantiles of the first three PC components of the fecal bacterial composition of the participants (C). The cutoff at FLI = 60 used to divide the participants is indicated with a dashed line in panels A and C.

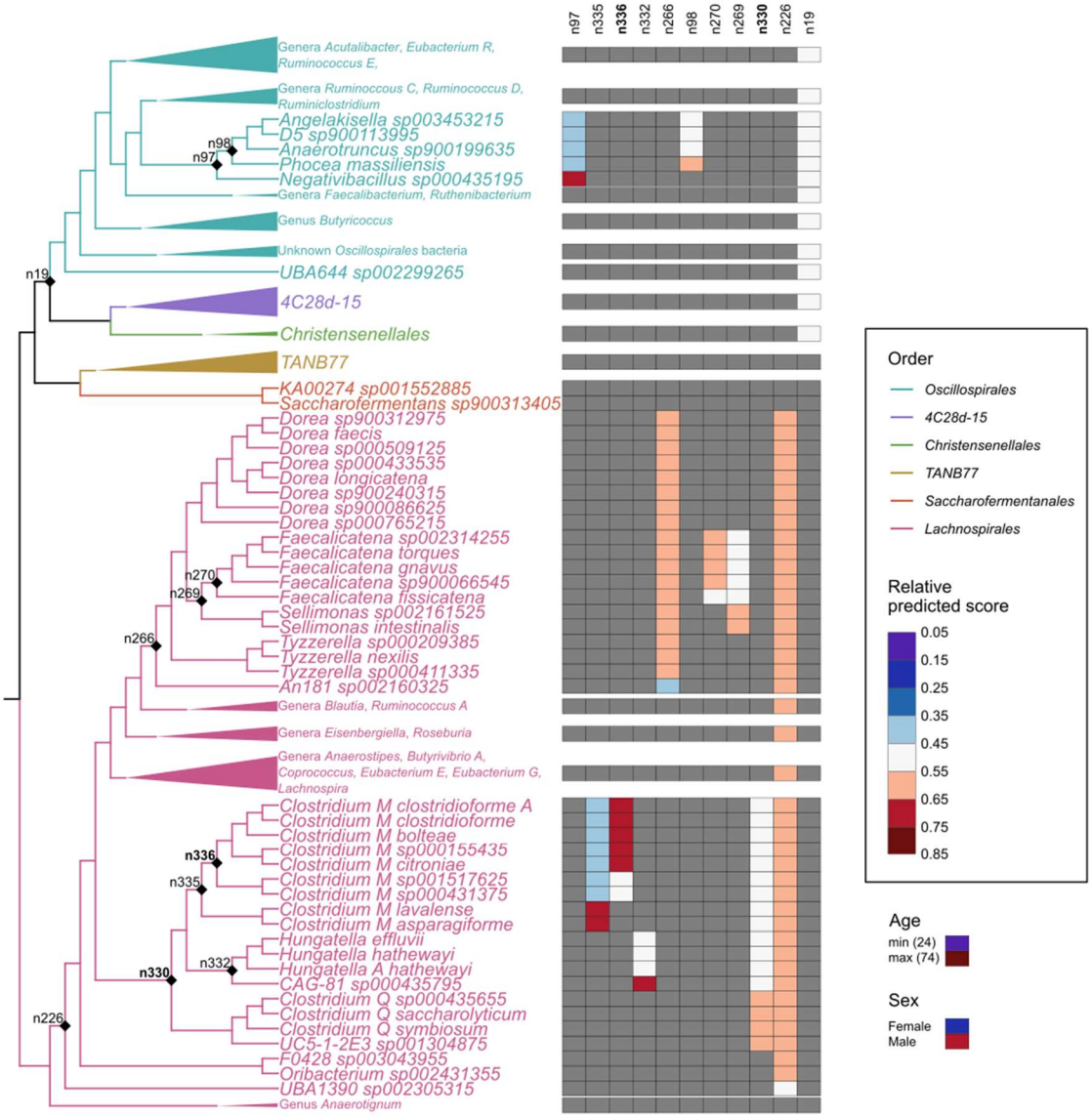

Figure 2. Relative effects of predictive balances and covariates on the FLI < 60 and FLI ≥ 60 classification model (AUC = 0.75) predictions. Nodes of the balances are indicated in the cladogram and the relative effect sizes of their clades (opposite sides of each balance) are shown in the associated heatmap. The relative effect sizes of the covariates (age and sex) are shown below the legend with a heatmap on the same scale as was used for the balances. The two liver-specific balances associated with triglyceride and GGT levels are indicated with bold font. Clades with redundant information have been collapsed but their major genera are indicated. The complete tree is included in **Figure S8**.