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. It is characterized by a build-up of excess fat in the liver that can lead to cirrhosis and liver failure. The link between fatty liver disease and gut microbiome has been known for at least 80 years. However, this association remains mostly unstudied in the general population because of underdiagnosis and small sample sizes. To address this knowledge gap, we studied the link between the Fatty Liver Index (FLI), a well-established proxy for fatty liver disease, and gut microbiome composition in a representative, ethnically homogeneous population sample of 6,269 participants in Finland. 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. Our results provide with replicable high-resolution associations between gut microbiota composition and fatty liver disease risk in a large representative population cohort and support 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 appear to be generalizable at the population level.

Results

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

In addition to known covariates, all microbiome data used in our analyses was either directly or indirectly based on archaeal and bacterial phylogenetic "balances". 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. [28]). 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 S2**, **S5**, and **S7**).

To investigate the link between fatty liver disease (using FLI as a proxy; **Figures 1A, 1B**) and gut microbial composition (**Figure S1**), we used linear regression (adjusted $R^2 = 0.29$) on the three first principal component (PC) axes of the fecal bacterial beta diversity (between individuals), sex, age, and alcohol to model FLI. $\log_{10}$(FLI) significantly correlated with all three bacterial PC axes, sex, age, and alcohol use (all $P < 1\times10^{-6}$). Correlations between FLI and archaeal PC axes were not significant ($P > 0.05$). The effect size estimate on $\log_{10}$(FLI) was a magnitude larger for PC1 (0.11 ± 0.008) than for PC2 (0.04 ± 0.008) and PC3 (-0.06 ± 0.008). The relationships between FLI and the bacterial PC components representing their beta diversity are visualized for each of the three components in **Figure 1C**. In our analyses, we classified our reads against the Genome Taxonomy Database (GTDB),[29] and thus the taxonomy discussed in this study follows the standardized GTDB taxonomy, unless otherwise noted.

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 S2**). Several clades had a negative association with FLI, on the negative side of the balances

154 contributing to PC1, such as order *Christensenellales* and genus *Faecalibacterium*. In addition,
155 genus *Bifidobacterium* in PC2, and family *Bifidobacteriaceae* in PC3 had negative associations
156 with continuous FLI.
157
158 ***Bacterial lineages within the NCBI Clostridium subclusters IV and XIVa associate***
159 ***with FLI***
160 In our study, the continuous FLI and differences between FLI groups in the FINRISK 2002
161 cohort (FLI < 60, N = 4,359 and FLI ≥ 60, N = 1,910; see **Figures 1A, 1B, Table S1**) were
162 modeled with gradient boosting regression or classification using Leave-One-Group-Out Cross-
163 Validation (LOGOCV) between participants from different regions.
164
165 After feature selection and Bayesian hyperparameter optimization, the correlation between the
166 predictions of the final regression models (age, sex, self-reported alcohol use, and 18 bacterial
167 balances as features; each trained on the data from 5/6 regions) and true values in unseen data
168 from the omitted region averaged R^2 = 0.30 (0.26 – 0.33). After feature selection and
169 optimization, the main classification models (age, sex, and 11 bacterial balances as features; each
170 trained on the data from 5/6 regions) averaged AUC = 0.75 (**Table S2**) and AUPRC = 0.56
171 (baseline at 0.30; **Table S3**) on (unseen) test data from the omitted region. Models trained using
172 only the covariates averaged AUC = 0.71 (AUPRC = 0.47) and using only the 11 bacterial
173 balances they averaged AUC = 0.66 (AUPRC = 0.47) on test data. Alternative models were
174 constructed by excluding participants with FLI between 30 and 60 (N = 1,583) and discerning
175 between groups of FLI < 30 (N = 2,776) and FLI ≥ 60 (N = 1,910). These models averaged AUC
176 = 0.80 (AUPRC = 0.75, baseline at 0.41) on their respective test data (**Tables S2, S3**). They

177 averaged AUC = 0.76 (AUPRC = 0.68) when using only the covariates, and AUC = 0.70
178 (AUPRC = 0.63) when using only the 20 bacterial balances.
179
180 Because training data from all 6 regions was used to prevent overfitting in the selection of core
181 features for all of the models, and similarly in searching for common hyperparameters,
182 participants from the validation region of each model (in the training partition) partly influenced
183 these parameters. Thus, we also constructed classification models discerning between the FLI <
184 60 and FLI ≥ 60 groups, where data of the validation region was completely excluded in the
185 feature selection and hyperparameter optimization of each LOGOCV model. These models,
186 using their individual feature sets and hyperparameters, averaged AUC = 0.75 and AUPRC =
187 0.57 (baseline at 0.30) on test data from their respective validation regions (**Table S4**). Using
188 only covariates, they averaged AUC = 0.71 (AUPRC = 0.47), and AUC = 0.67 (AUPRC = 0.48)
189 with only the bacterial balances.
190
191 Our external validation data consisted of 258 participants after exclusion of pregnant participants
192 or those on antibiotics in the past 6 months, in the FINRISK 2007 population cohort[30] (**Table S1**,
193 **Figure S3**). The participants originate from North Karelia and Helsinki/Vantaa regions in
194 Finland, and their samples were processed with the same methodology as was used for FINRISK
195 2002. In this external validation, the 6 full models trained with covariates and the 11 bacterial
196 balances in FINRISK 2002 averaged AUC = 0.77 (AUPRC = 0.51, baseline at 0.21; **Table S5**).
197 The covariate-only models averaged AUC = 0.72 (AUPRC = 0.40) and the balance-only models
198 averaged AUC = 0.69 (AUPRC = 0.44). The receiver operating characteristic and precision-
199 recall curves based on the averaged predictions of the models, tested on this external validation

data, also display good diagnostic ability (AUC = 0.78, AUPRC = 0.51 with baseline at 0.51; **Figure S4**)

To facilitate interpretability of the results, we continued examining the main classification models using a common set of core features. In these models, the median effect sizes of the features on the model predictions at their minimum and maximum values were highest for age, followed by sex, and the 11 balances in the phylogenetic tree (**Figures S5**, **S6**). All 11 associated balances were in phylum *Firmicutes*, class *Clostridia*, and largely in the NCBI *Clostridium* subclusters IV and XIVa (**Figure 2).** The specific taxa represented standardized GTDB genera (NCBI in brackets) *Negativibacillus* (*Clostridium*), *Clostridium M* (*Lachnoclostridium* / *Clostridium*), *CAG-81* (*Clostridium*), *Dorea* (*Merdimonas* / *Mordavella* / *Dorea* / *Clostridium* / *Eubacterium*), *Faecalicatena* (*Blautia* / *Ruminococcus* / *Clostridium*), *Blautia* (*Blautia*), *Sellimonas* (*Sellimonas* / *Drancourtella*), *Clostridium Q* (*Lachnoclostridium* [*Clostridium*]) and *Tyzzerella* (*Tyzzerella* / *Coprococcus*). Notably, all but one of the features in the main classification models (n226) were identified in the feature selection for the alternative models (constructed otherwise identically, but FLI < 30 was compared against FLI ≥ 60 in different data partitions), together with 10 additional balances (**Figure S7**). Only one of the balances in the alternative models was outside phylum *Firmicutes* (n1712 in *Bacteroidota*), and in addition, 4 balances were outside class *Clostridia* (n481 in *Negativicutes*; n826, n1009 and n918 in *Bacilli*). Also, negative associations with the high FLI group were seen for *An181 sp002160325* in the balance n266, where it is compared against the clade including *Dorea*, *Faecalicatena*, *Sellimonas* and *Tyzzerella* species (**Figures 2**, **S6**). A higher abundance of the clade including

222 *Angelakisella*, *D5*, *Anaerotruncus* and *Phocea* species (against *Negativibacillus sp00435195* in
223 balance n97) was also negatively associated with high FLI.
224
225 In addition to blood test results, FLI is based on two anthropometric markers linked to metabolic
226 syndrome, waist circumference and BMI. This prompted us to dissect the Fatty Liver Index and
227 identify which of the covariates and associated microbial balances from the phylogenetic tree can
228 be linked to blood GGT and triglycerides measurements (see **Figure 1B**), and therefore would be
229 most specific to hepatic steatosis and liver damage.[31] To do so, we performed feature selection
230 (similarly to continuous FLI) for GGT and triglycerides measurements in subsets of participants
231 grouped by age, sex, and BMI. The feature selection identified two balances within the NCBI
232 *Clostridia* XIVa subcluster (identified as n336 and n330) which were important for both GGT
233 and triglyceride level prediction, and thus likely specific to liver function (**Figure 2**). Bacterial
234 taxa were positively linked to liver function in these balances, and included (NCBI species)
235 *Clostridium clostridioforme*, *C. bolteae*, *C. citroniae*, *C. saccharolyticum* and *C. symbiosum*. On
236 the opposite, negatively associated side of the balances were, among others, (NCBI species)
237 *Hungatella effluvii*, *H. hathewayi*, and two new GTDB-defined species *Clostridium M*
238 *sp001517625* and *C. M sp000431375*.
239
240 ***Ethanol and acetate production pathways are identified in representative bacterial***
241 ***genomes from taxa linked to high FLI***
242 The values of predictive balances in the phylogenetic tree cannot be summarized for individual
243 taxa, which means that only a qualitative investigation of the associations between their
244 metabolism and fatty liver was possible in this study. We identified genetic pathways predicted

245 to encode for SCFA (acetate, propanoate, butanoate) and ethanol production, BA metabolism,
246 and choline degradation to trimethylamine (TMA) in representative genomes from the taxa we
247 identified to be linked to liver function (**Figure S6**). These specific processes were chosen
248 because they have been previously identified to have a mechanistic link to NAFLD (see *e.g.,* ref.
249 [19]).
250
251 Acetate and ethanol production pathways appeared to be more abundant in the representative
252 genomes of the taxa which had a positive association with FLI. In the liver function specific
253 clades, n336 and n330, MetaCyc pathways for pyruvate fermentation to ethanol III (PWY-6587)
254 and L-glutamate degradation V (via hydroxyglutarate; P162-PWY; produces acetate and
255 butanoate) were present only in genomes positively associated with FLI. In balance n336, also
256 heterolactic fermentation (P122-PWY; produces ethanol and lactate) was more often encoded in
257 the clade positively associated with the high FLI group (3/5) than the opposing negatively
258 associated clade (1/2). In representative genomes from the liver-specific balance n336, potential
259 ethanol producers (PWY-6587) were seen in the positively associated clade (*Clostridium M*
260 *clostridoforme A* and *Clostridium M sp000155435*), and not in the negatively associated clade
261 (*Clostridium M sp001517625* and *Clostridium M sp000431375*). However, for most balances
262 such trends were not clear in the qualitative analysis. Furthermore, we did not detect any of these
263 pathways in the representative genomes of two individual taxa positively associated with FLI,
264 *Negativibacillus sp000435195* and *Phocea massiliensis* (**Figure S6**).
265

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**). The focus of the current study was on uncovering the strongest connections between microbiome composition and high risk of fatty liver disease at the population level. We acknowledge that because of the complexity of our analyses, these methods are not yet suited for widespread clinical application. However, our study lays the groundwork for future mechanistic studies, and eventually, the use of this information in disease prevention, diagnosis, and treatment.

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 cohorts investigated in the current study,[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 models was highly similar in an external validation cohort (AUC = 0.77, AUPRC = 0.51, baseline at 0.21).

Our additional analyses support these main results. 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. [33]).

When interpreting results, several different levels of associations can be considered according to types of fatty liver disease and the gut microbiome composition. Because FLI has been mostly validated with simple steatosis and NAFLD,[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.[34] Much lower cutoffs (FLI ≥ 20 to 30) have been used in Asian cohorts.[35–37] 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.

312

313 ***FLI modeling reveals consistent associations between gram-positive Clostridia and***

314 ***fatty liver disease***

315 We found significant linear correlations between the first three bacterial PC-axes of our

316 samples (a measure of beta diversity) and FLI (see results and **Figure 1C**). Previous studies

317 have shown differences in beta diversity in relation to NAFLD.[38] However, FLI used in our

318 study as a proxy for liver disease also includes features such as BMI and waist circumference,

319 which associate with metabolic syndrome and diabetes.[16] Links between these diseases and gut

320 microbiome composition are well documented in previous studies.[39] Several studies have

321 reported highly specific changes in microbial abundances in relation to NAFLD.[12,40–42] In

322 summary, while also conflicting results have been reported, generally increases in

323 *Lactobacillus* and *Escherichia* genera, and a decrease in *Coprococcus* genus have been

324 associated with a NAFLD diagnosis.[43] Furthermore, increased abundance of several gram-

325 positive bacteria belonging to the *Clostridium* genus have been positively linked with

326 NAFLD.[14,44] Differences in unconstrained between-samples (beta) diversity have been also

327 documented for persistent NAFLD,[38] and along the NAFLD-cirrhosis spectrum.[8]

328

329 Through the loadings of the phylogenetic balances on the PC axes, we observed several

330 previously known connections between microbial clades and FLI. Among others, we observed

331 a positive association between high FLI and family *Lachnospiraceae* and negative associations

332 for order *Christensenellales*, genus *Faecalibacterium*, and genus *Bifidobacterium*. The positive

333 association is supported by previous findings of their connection with obesity[45], and the

334 negative associations by connections to lean individuals and healthy gut microbiome

335 composition.[46–48] However, these associations discovered from linear modeling of FLI based on
336 microbial beta diversity were coarse-grained, and prompted us to construct the cross-validated
337 machine learning models based directly on the phylogenetic balances to increase their
338 resolution.
339
340 In these machine learning models, abundances of bacteria from the *Coprococcus* genus were
341 not specifically associated with FLI, although the genus was nested inside our predictive
342 balances. Strikingly, the strongest associations with FLI in our machine learning models were
343 inside the *Firmicutes* phylum. A possible reason for this might be the higher relative abundance
344 of phylum *Firmicutes* at high latitudes, where Finland is.[49] Among the associations we
345 identified, *Faecalicatena gnavus* (NCBI: *Ruminococcus gnavus*) was positively linked with
346 FLI as part of 3 predictive balances, and associated in previous studies with liver cirrhosis.[17]
347 Interestingly, none of the oral *Firmicutes*, such as *Veillonella*, suggested to invade the gut, were
348 identified in our own analyses. This might be caused by using only the broadly defined high
349 FLI as a proxy, which does not directly target advanced liver disease, such as cirrhosis.
350
351 Two individual taxa, *Negativibacillus sp000435195* and *Phocea massiliensis*, both had strong
352 positive associations with the high FLI group (**Figures 2, S6**), but the balances including these
353 species were not predictive of the liver function-specific components (triglycerides and GGT).
354 Positive associations of these taxa with fatty liver disease have not been documented
355 previously. However, a decreasing abundance of both bacteria, *Negativibacillus sp000435195*
356 (NCBI: *Clostridium* sp. CAG:169) and *Phocea massiliensis* (NCBI: *Phocea massiliensis*), were
357 seen when the intake of meat and refined cereal was reduced isocalorically in favor of fruit,

358 vegetables, wholegrain cereal, legumes, fish and nuts in overweight and obese subjects in
359 Italy.[50] While comparisons between these studies are difficult due to differences in taxa
360 annotations, bacteria such as *Faecalicatena gnavus* (NCBI: *Ruminococcus gnavus*) and
361 *Clostridium Q saccharolyticum* (NCBI: *Clostridium saccharolyticum*) were also found to
362 respond negatively to the Mediterranean diet. Together with their positive associations with
363 FLI in our study, these observations would warrant further study on these species as plausible
364 biomarkers for healthy diet choices.
365
366 Among the taxa negatively associated with high FLI, *Hungatella* (see balance n332, **Figures 1**,
367 **S6**) have been previously shown to correlate negatively with the obesity phenotype in mice[51]
368 and *H. hathewayi* was found to be a common commensal in the gut of healthy volunteers.[52]
369 However, genus *Hungatella* has also been positively associated with concentrations of
370 trimethylamine-N-oxide (TMAO),[53] a metabolite associated with cardiovascular disease and
371 NAFLD. In our study, on the positively associated side (of balance n332) opposite to genus
372 *Hungatella* was a novel GTDB species, *CAG-81 sp000435795*, previously included in NCBI
373 genus *Clostridium*. The *CAG-81* genus was recently positively associated with TMAO levels in
374 urine in a study using the GTDB classification.[54] While we did not find the pathway for TMA
375 (precursor to TMAO) production in its genome, this would explain the positive association of
376 the *CAG-81* species with high FLI. Furthermore, the previous contradictive results among these
377 taxa could be explained by grouping of putatively TMA producing taxa in *CAG-81* together
378 with the closely related genus *Hungatella*.
379

Most taxa in our study with a positive association with FLI belonged to the broadly defined *Clostridium* NCBI genus, which supports several previous observations.[14,44] However, taxonomic standardization according to GTDB has identified the *Clostridium* genus as the most phylogenetically inconsistent of all bacterial genera in the NCBI taxonomy, and divides it into a total of 121 monophyletic genera in 29 distinct families.[29] The GTDB reassignment complicates comparisons to previous studies, but it is phylogenetically and biologically sensible, and can thus provide new insights into the microbiomes. Our results also strongly suggest that despite its higher cost compared to metabarcoding, the increased resolution of (shallow) shotgun metagenomic sequencing is highly useful in identifying specific taxon-disease associations (see *e.g.,* refs. [27,55]).

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

We identified several plausible new associations between individual taxa and clades of bacteria and fatty liver. All taxa were from class *Clostridia*, which are obligate anaerobes. We observed that reference genomes from the bacterial taxa positively associated with 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,[56] 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* [57], and *C. Q Saccharolyticum*.[58] The relative abundances of these putatively ethanol-producing

403 taxa were also predictive of FLI groups in previously unseen data. However, the self-reported
404 alcohol consumption from the participants was not among the best predictors for the FLI
405 groups, as it was excluded in the feature selection step.
406
407 All reference genomes from taxa positively associated with FLI in balance n330 harbored
408 genes predicted to encode for the L-glutamate fermentation V (P162-PWY; **Figure S5**)
409 pathway, which results in the production of acetate and butanoate. Glutamate fermentation
410 could lead to increased microbial protein fermentation in the gut, which has been previously
411 been linked with obesity, diabetes and NAFLD.[59] Recently, the combined intake of fructose
412 and microbial acetate production in the gut was experimentally observed to contribute to
413 lipogenesis in the liver in a mouse model.[60] Interestingly, *C. Q saccharolyticum* (in our study, a
414 taxa positively associated with high FLI deriving from balance n330), was experimentally
415 shown to ferment various carbohydrates, including fructose, to acetate, hydrogen, carbon
416 dioxide, and ethanol.[58] Furthermore, while our own pathway analysis did not detect BA
417 modification pathways in the reference genome of *C. Q saccharolyticum*, a strain of this
418 species has been highlighted as a probable contributor to NAFLD development through the
419 synthesis of secondary BA.[15] The links between dietary intake and gene regulation, combined
420 with microbial fermentation in the gut warrant further mechanistic experiments to elucidate
421 their links with fatty liver, and likely other metabolic diseases.
422
423 NAFLD-associated ethanol producing bacteria in previous cohort studies have all been gram-
424 negatives, such as (NCBI-defined) *Klebsiella pneumoniae,*[13] and *Escherichia coli.*[12] In our
425 population sample, instead of gram-negatives, bacteria from the *C. M bolteae*, *C. M*

426 *clostridioforme / C. M clostridioforme A* and *C. M citroniae* species (positively associated with
427 high FLI in balance n336) have been described as opportunistic pathogens,[61] and are
428 hypothesized to exacerbate fatty liver development similarly through endogenous ethanol
429 production. This result suggests that geographical,[33] and ethnic variability,[62] might also
430 strongly affect gut microbiome composition and its associations with disease. In addition to
431 putative endogenous ethanol producers, we identified other taxa positively associated with high
432 FLI in balance n330, for which reference genomes harbored a genetic pathway predicted to
433 encode for the ability to ferment L-lysine to acetate and butyrate. While the production of these
434 SCFAs is often considered beneficial for gut health, other metabolism of proteolytic bacteria
435 might negatively contribute to fatty liver disease.[63]
436
437 Through modeling a previously validated index for fatty liver, FLI, we found replicable
438 associations with specific microbial taxa and high risk of liver disease. In addition, sex and age
439 of participants were also strongly predictive of the FLI group in our models (**Figures 2, S5**).
440 Their similar positive associations with fatty liver disease are known from previous studies.[64,65]
441 The associated microbial balances could be used to improve the predictions above the baseline
442 of these covariates on 5/6 regions in Finland in the main cohort. For example, in the model
443 cross-validated with Lapland the balances were more predictive of FLI group than the
444 covariates by themselves, and their combination increased the AUC further. Yet, when testing
445 the model where Turku/Loimaa region was used for internal cross-validation, the microbial
446 balances were slightly predictive of FLI group but failed to improve the AUC over the
447 covariates (**Table S2**). This pattern might stem from the cultural and genetic west-east division
448 in Finland,[66,67] with a closer proximity of the Helsinki/Vantaa region to eastern regions than

449 Turku/Loimaa, in both terms. It is thus likely that further incorporation and investigation on the
450 use of spatial information in microbiome modeling would elucidate these geographical patterns
451 in taxa-disease associations.
452
453 Our models were also able to accurately predict the FLI group of participants in the external
454 validation cohort, which were from the North Karelia and Helsinki/Vantaa regions.
455 Unfortunately, generalization of our results outside of Finland remains to be addressed in future
456 studies. The observed difficulty to geographically extrapolate taxa-disease associations[33] might
457 mean that associations reported in our study are specific to Finland and nearby regions.
458 Notably, many of the positive associations between specific taxa and fatty liver disease have
459 not been reported previously, but the functional potential of these taxa inferred from genomic
460 data is similar to taxa positively associated with NAFLD in previous studies. Thus, the
461 geographical limits of taxa-disease associations reported in studies such as ours warrant further
462 study.
463
464 It is likely that not all associations in the current study are related solely to liver steatosis,
465 because FLI is based on measurements related to metabolic syndrome. However, our approach
466 is supported by recent views of NAFLD as the integral liver component of the metabolic
467 syndrome.[68,69] Indeed, the prevalences of diabetes and cardiovascular disease in both FINRISK
468 2002 and 2007 cohorts are elevated in the high FLI group, although the majority of the high
469 FLI participants did not have either of these diagnoses at the time of sampling (**Table S1**). We
470 also dissected the FLI by dividing participants into age/sex/BMI groups and detected microbial
471 groups specific to the blood work measurements of liver damage, triglycerides and GGT. These

472 associated taxa can thus be thought of as most closely associated with liver function, if such a
473 division is deemed practical.
474
475 ***Conclusions***
476 Modeling an established risk index for fatty liver enabled the detection of associations between
477 the disease and gut microbiome composition, even to the level of individual taxa. The
478 associated clades were all from the obligately anaerobic gram-positive class *Clostridia*,
479 representing several redefined GTDB genera previously included in the polyphyletic NCBI
480 genus *Clostridium*. Many of the representative genomes of taxa positively associated with high
481 FLI had genomic potential for endogenous ethanol production. This suggests a possible
482 mechanistic link to the pathophysiology of fatty liver disease through increased gut
483 permeability and induction of hepatic steatosis. Further mechanistic links with microbial
484 production of TMA and SCFAs, especially acetate, and fatty liver development are also likely.
485 Our models were able to predict the FLI group of participants in Finland across geographical
486 regions and in an external cohort, showing that the associations are robust and generalizable in
487 the sampled population. Our hypothesized mechanistic links and geographical limits of taxa-
488 disease associations should be addressed in future studies to increase our understanding of the
489 pathophysiology of fatty liver disease, and for developing clinical applications aimed at its
490 prevention, diagnosis, and treatment.
491

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. [24,70]). The Coordinating Ethics Committee of the Helsinki University

514 Hospital District approved the study protocol for FINRISK 2002 (Ref. 558/E3/2001), and all
515 participants have given their written informed consent.
516

517 ***Stool DNA extraction and shallow shotgun metagenome sequencing***

518 DNA extraction was performed according to the Earth Microbiome Project protocols, with the
519 MagAttract PowerSoil DNA kit (Qiagen), as previously described.[71] A miniaturized version of
520 the Kapa HyperPlus Illumina-compatible library prep kit (Kapa Biosystems) was used for library
521 generation, following the previously published protocol.[72] DNA extracts were normalized to 5 ng
522 total input per sample in an Echo 550 acoustic liquid handling robot (Labcyte Inc.). A Mosquito
523 HV liquid-handling robot (TTP Labtech Inc.) was used for 1/10 scale enzymatic fragmentation,
524 end-repair, and adapter-ligation reactions. Sequencing adapters were based on the iTru
525 protocol,[73] in which short universal adapter stubs are ligated first and then sample-specific
526 barcoded sequences added in a subsequent PCR step. Amplified and barcoded libraries were then
527 quantified by the PicoGreen assay and pooled in approximately equimolar ratios before being
528 sequenced on an Illumina HiSeq 4000 instrument to an average read count of approximately
529 900,000 reads per sample.
530

531 ***Taxonomic matching and phylogenetic transforms***

532 We quality trimmed the sequences and removed the sequencing adapters with Atropos.[74] Host
533 reads were removed by mapping the reads against the human genome assembly GRCh38 with
534 Bowtie2.[75] To improve the taxonomic assignments of our reads, we used a custom index,[76] based
535 on the Genome Taxonomy Database (GTDB) release 89 taxonomic redefinitions,[29,77] for read
536 classification with default parameters in Centrifuge 1.0.4.[78] Viral and eukaryotic sequences were

removed in this step, as the database contains only bacterial and archaeal reference genomes. After read classification, all following steps were performed with R version 3.5.2.[79] To reduce the number of spurious read assignments, and to facilitate more accurate phylogenetic transformations, only reads classified at the species level, matching individual GTDB reference genomes, were retained. Samples with less than 50,000 reads, from pregnant participants or recorded antibiotic use in the past 6 months were removed, resulting in a final number of 6,269 samples. We first filtered taxa not seen with more than 3 counts in at least 1% of samples and those with a coefficient of variation ≤ 3 across all samples, following McMurdie and Holmes[80], with a slight adaption from 20% of samples to 1% of samples, because of our larger sample size. The complete bacterial and archaeal phylogenetic trees of the GTDB release 89 reference genomes, constructed from an alignment of 120 bacterial or 122 archaeal marker genes,[29] were then combined with our taxa tables. The resulting trees were thus subset only to species which were observed in at least one sample in our data. The read counts were transformed to phylogenetic node balances in both trees with PhILR.[28] The default method for PhILR inputs a pseudocount of 1 for taxa absent in an individual sample before the transform.

In this study, we did not specifically and solely use relative abundances at various taxonomic levels, as is common practice for microbiome studies. Instead, we applied a PhILR transformation to our microbial composition data,[28] 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,[81] 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.[82] Thus, associations between the microbes and the response variable are likely not best explained by predefined taxonomic levels. In the absence of functional data, concurrently analyzing all clades (partitioned by the nodes in the phylogenetic tree) would likely enable the detection of the associations at the appropriate resolution depending on the function and the local tree topography.

Covariates

Because fatty liver disease is underdiagnosed at the population level,[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]:

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

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 (0.3%) and the self-reported alcohol use was missing for 247 participants (3.9%). In the machine learning framework, missing values for FLI and self-reported alcohol use were mean imputed. However, for the feature selection to identify liver function-specific balances, GGT, triglycerides and BMI were not imputed but observations where any of these were missing were simply removed.

Beta diversity and linear modeling of FLI

Beta diversity was calculated as Euclidian distance of the PhILR balances through Principal Component Analysis (PCA) on bacterial and archaeal balances separately with "rda" in vegan 2.5.6.[83] A linear regression model was constructed for FLI with "lm" in base R,[79] with $\log_{10}$-transformed FLI as the dependent variable and with first three bacterial PCs, sex, age, and self-reported alcohol use as the independent variables. Archaeal PCs were dropped from the model because none of them were significantly correlated with FLI (all $P > 0.05$). Variation of the samples on the top two bacterial PC axes by their effect sizes in the model were plotted together with a unit vector of $\log_{10}$(FLI) to show their correlation.

FLI modeling within a machine learning framework

In the machine learning framework, both regression and categorical models were constructed for FLI. The feature selection, hyperparameter optimization and internal cross-validation methods were identical for both approaches, unless otherwise stated. The continuous or categorical FLI

606 (groups of FLI < 60 and FLI ≥ 60) were modeled with xgboost 0.90.0.2,[84] by using both bacterial
607 and archaeal balances, sex, age, and self-reported alcohol use as preliminary predictor features.
608 We used FLI 60 as the cutoff for ruling in fatty liver (steatosis) for the classification, after
609 Bedogni et al., (2006).[26] The data was first split to 70% train and 30% test sets while preserving
610 sex and region balance. To take into account geographical differences (see *e.g.,* ref. [33]) and to
611 find robust patterns across all 6 sampled regions in Finland between the features and FLI group,
612 we used Leave-One-Group-Out Cross-Validation (LOGOCV) inside the 70% train set to
613 construct 6 separate models in each optimization step. Because of high dimensionality of the data
614 (3,423 predictor features) feature selection by filtering was first performed inside the training
615 data, based on random forest permutation as recommended by Bommert et al.[85] Briefly,
616 permutation importance is based on accuracy, or specifically the difference in accuracy between
617 real and permuted (random) values of the specific variable, averaged in all trees across the whole
618 random forest. The permutation importance in models based on the 6 LOGOCV subsets of the
619 training data were calculated with mlr 2.16.0,[86] and the simple intersect between the top 50
620 features in all LOGOCV subsets were retained as the final set of features. Thus, the feature
621 selection was influenced by the training data from all 6 geographical regions, but this only serves
622 to limit the number of chosen features because of the required simple intersect. This approach
623 was used to obtain a set of core predictive features which would have potential for
624 generalizability across the regions. The number of features included in the models by this
625 approach was deemed appropriate, since the relative effect size of the last included predictor was
626 very small (< 0.1 change in classification probability across its range).
627

Bayesian hyperparameter optimization of the xgboost models was then performed with only the selected features. An optimal set of parameters for the xgboost models were searched over all LOGOCV subsets with "mbo" in mlrMBO 1.1.3,[87] using 30 preliminary rounds with randomized parameters, followed by 100 optimization rounds. Parameters in the xgboost models and their considered ranges were learning rate (eta) [0.001, 0.3], gamma [0.1, 5], maximum depth of a tree [2, 8], minimum child weight [1, 10], fraction of data subsampled per each iteration [0.2, 0.8], fraction of columns subsampled per tree [0.2, 0.9], and maximum number of iterations (nrounds) [50, 5000]. The parameters recommended by these searchers were as following for regression: eta=0.00889; gamma=2.08; max_depth=2; min_child_weight=8; subsample=0.783; colsample_bytree=0.672; nrounds=1,810, and for classification: eta=0.00107; gamma=0.137; max_depth=5; min_child_weight=9; subsample=0.207; colsample_bytree=0.793; nrounds=4,328. We used Root-Mean-Square Error (RMSE) for the regression models and Area Under the ROC Curve (AUC) for the classification models to measure model fit on the left-out data (region) in each LOGOCV subset. Receiver operating characteristic and precision-recall curves for these validation metrics were calculated with "evalmod" in precrec v0.11.2.[88] The final models were trained on the LOGOCV subset training data, the data from one region thus omitted per model, and using the selected features and optimized hyperparameters. Internal validation of these models was conducted against participants only from the region omitted from each model, in the 30% test data which was not used in model training or optimization. Sensitivity analysis was conducted by using only the predictive covariates (sex and age) or balances separately, with the same hyperparameters, data partitions and cross-region internal validation as for the full models.

651 ***Partial dependence interpretation of the FLI classification models***

652 Because the classification models have a more clinically relevant modeling target for the
653 difference between FLI < 60 and FLI ≥ 60, the latter used to rule in fatty liver,[26] we further
654 interpreted the partial dependence of their predictions. Partial dependence of the classification
655 model predictions on individual features was calculated with "partial" in pdp 0.7.0.[89] The partial
656 dependence of the features on the model predictions was also plotted, overlaying the results from
657 each of the 6 models. For each feature, its relative effect on the model prediction was estimated
658 as medians of the minimum and maximum yhat (output probability of the model for the FLI ≥ 60
659 class), calculated at the minimum and maximum values of the feature separately in each of the 6
660 models. The relative effects of the balances were then overlaid as a heatmap on a genome
661 cladogram which covers all balances in the model with ggtree 2.1.1.[90]

662

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

665 To assess robustness of the models and how removing the participants with intermediate FLI
666 (between 30 and 60) affects model performance, we removed this group (N = 1910) and
667 constructed alternative classification models to discern between the FLI < 30 and FLI ≥ 60
668 groups. Other than removing the intermediate FLI participants and resulting new random split to
669 the train (70%) and test (30%) sets, these models were constructed identically to the main
670 models, including LOGOCV design, feature selection, and hyperparameter optimization. The
671 recommended parameters for this classification task were eta=0.00102; gamma=3.7;
672 max_depth=2; min_child_weight=5; subsample=0.49; colsample_bytree=0.631; nrounds=3,119.

673 Interpretation of partial dependence was also performed identically, but only the relative effects
674 of the model features were plotted without a cladogram.
675
676 ***Exclusion of validation region data from feature selection and hyperparameter***
677 ***optimization***
678 Because training data from all 6 regions is used to inform the selection of optimal features and
679 hyperparameters, the validation region data cannot be considered completely independent from
680 the training of the LOGOCV models. Thus, we constructed a set of classification models for the
681 FLI $\geq$ 60 and FLI < 60 groups, where all validation region samples also in the training data were
682 excluded from the simple intercept of top 50 features in each LOGOCV set and from the
683 subsequent hyperparameter optimization. These models with individualized features and
684 hyperparameters were then tested on the validation region samples in the unseen test data to
685 estimate how model performance was affected. The main test (70%) and train (30%) sets were
686 identical to the main models, but additionally 6 randomized 70/30 splits nested inside the test set
687 (excluding the validation region) were used in hyperparameter optimization to reduce overfitting.
688 Average optimal hyperparameters in the 6 models were eta=0.00106; gamma=4.3;
689 max_depth=2; min_child_weight=7; subsample=0.36; colsample_bytree=0.613; nrounds=1,772.
690
691 ***External validation of the models in a separate population cohort***
692 To further validate our models and results, we leveraged the data from a more recent population
693 cohort in Finland, FINRISK 2007 (see **Table S1**). In this cohort, the choice of participants,
694 sample collection, and related methods for the data used in the current study were similar to
695 FINRISK 2002 to facilitate inter-cohort comparisons, and are reported elsewhere.[30] The study

696 protocol of FINRISK 2007 was approved by the Coordinating Ethical Committee of the Hospital
697 District of Helsinki and Uusimaa (Ref. 229/EO/2006). All participants have signed an informed
698 consent.
699
700 Briefly, compared to FINRISK 2002, FINRISK 2007 features a smaller number of participants
701 who donated fecal samples ($N = 258$ after excluding pregnant individuals or antibiotic use in the
702 last 6 months), they were younger on average, and a smaller proportion of them were in the high
703 FLI group. To produce data for the validation, methods and quality control related to DNA
704 extraction, sequencing, taxonomic assignments, and calculation of FLI values were identical to
705 FINRISK 2002 data, as described above. For the phylogenetic transform (performed otherwise
706 identically), only taxa passing the filtering in FINRISK 2002 bacterial data set were retained in
707 FINRISK 2007 and a pseudo-count of 1 was used for taxa unobserved in the new data, to exactly
708 match the node balance names. The FINRISK 2007 data was then subset to the model features of
709 the main classification models (sex, age, and the 11 bacterial balances), and input in each of the 6
710 LOGOCV classification models. The results of these predictions were then compared against the
711 true FLI groups (FLI $\geqslant$ 60 and FLI < 60) of the participants (**Table S5**). Receiver operating
712 characteristic and precision-recall curves for the external validation were calculated similarly to
713 the main models for the AUC and AUPRC metrics and plotted after averaging the predictions of
714 the 6 models to obtain single curves (**Figure S4**).
715
716 ***Identification of predictive features specific to liver function***
717 Because the FLI also incorporates BMI and waist circumference, and they strongly contribute to
718 the index,[26] we deemed it necessary to further investigate which of the identified balances were

719 specific to liver function. The participants were first grouped by age (< 40, 40 – 60, and 60 <),
720 sex (female or male) and BMI (< 25, 25 – 30, and 30 <) into 18 categories (N = 105 ~ 711 per
721 category). We performed feature selection similarly to the FLI models by fitting random forest
722 regressors for GGT and triglycerides with mlr 2.16.0.[86] This was done separately in each of the
723 18 categories, and in each category, we again used LOGOCV with the regions to obtain 6 runs
724 per category. Finally, the features predictive of GGT or triglycerides in each category were
725 selected as the intersect of top 50 features in the 6 LOGOCV iterations by permutation
726 importance. The intersect of features predictive of GGT or triglycerides in any of the categories
727 and the features predictive of categorical FLI were identified as specific to liver function.
728
729 ***Pathway inference for taxa associated with FLI***
730 Our taxonomic matching of the reads is based on the genomes of GTDB (release 89),[29] which are
731 all complete or nearly complete and available in online databases. This enables us to estimate the
732 likely genetic content, and thus, the metabolic potential of the microbes associated with FLI. We
733 use this approach because the sequencing depth of our samples does not allow assembling
734 contigs and (metagenome-assembled) genomes, required for pathway predictions. Because of the
735 compositional phylogenetic transform, among other features of our data, previously developed
736 approaches such as PICRUSt,[91] could not be used here.
737
738 The genomes of all 336 bacteria under at least one of the predictive balances were downloaded
739 from NCBI. 119 of these genomes were originally not annotated, which is a requirement for
740 pathway prediction. Therefore, Prokka v1.14.6,[92] was used to annotate the 119 unannotated
741 genomes as a preliminary step. Pathway predictions were then performed for all 336 genomes

with mpwt v0.5.3 multiprocessing tool,[93] for the PathoLogic pipeline of Pathway Tools 23.0.[94] 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. [95], 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-

765 Ingelheim, Bristol-Myers Squibb, Cirius, Eli Lilly and Company, Galectin Therapeutics, Galmed
766 Pharmaceuticals, GE, Genfit, Gilead, Intercept, Grail, Janssen, Madrigal Pharmaceuticals,
767 Merck, NGM Biopharmaceuticals, NuSirt, Pfizer, pH Pharma, Prometheus, and Siemens. He is
768 also co-founder of Liponexus, Inc.
769

770 ***Funding details***

771 This research was supported in part by grants from the Finnish Foundation for Cardiovascular
772 Research, the Emil Aaltonen Foundation, the Paavo Nurmi Foundation, the Urmas Pekkala
773 Foundation, the Finnish Medical Foundation, the Sigrid Juselius Foundation, the Academy of
774 Finland (#321356 to A.H.; #295741, #307127 to L.L.; #321351 to T.N.) and the National
775 Institutes of Health (R01ES027595 to M.J.). R.L. receives funding support from NIEHS
776 (5P42ES010337), NCATS (5UL1TR001442), NIDDK (U01DK061734, R01DK106419,
777 P30DK120515, R01DK121378, R01DK124318), and DOD PRCRP (W81XWH-18-2-0026).
778 Additional support was provided by Illumina, Inc. and Janssen Pharmaceutica through their
779 sponsorship of the Center for Microbiome Innovation at UCSD.
780

781 ***Authors' contributions***

782 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
783 and R.K. acquired the data. M.R., L.L. and T.N. analyzed the data. M.R. wrote the manuscript in
784 consultation with all authors. M.I., P.J., V.S., R.K., L.L. and T.N. supervised the work. All
785 authors gave final approval of the version to be published.
786

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

29. 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.

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. 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.

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

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

895 36. 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
896 H-C, Lee F-Y, et al. External Validation of Fatty Liver Index for Identifying
897 Ultrasonographic Fatty Liver in a Large-Scale Cross-Sectional Study in Taiwan. PLOS
898 ONE 2015; 10:e0120443.

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

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

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

909 40. Wigg AJ, Roberts-Thomson IC, Dymock RB, McCarthy PJ, Grose RH, Cummins AG. The
910 role of small intestinal bacterial overgrowth, intestinal permeability, endotoxaemia, and
911 tumour necrosis factor α in the pathogenesis of non-alcoholic steatohepatitis. Gut 2001;
912 48:206–11.

913 41. Mouzaki M, Comelli EM, Arendt BM, Bonengel J, Fung SK, Fischer SE, McGilvray ID,
914 Allard JP. Intestinal microbiota in patients with nonalcoholic fatty liver disease. Hepatology
915 2013; 58:120–7.

916 42. Shen F, Zheng R-D, Sun X-Q, Ding W-J, Wang X-Y, Fan J-G. Gut microbiota dysbiosis in
917 patients with non-alcoholic fatty liver disease. Hepatobiliary & Pancreatic Diseases
918 International 2017; 16:375–81.

919 43. Sharpton SR, Ajmera V, Loomba R. Emerging Role of the Gut Microbiome in
920 Nonalcoholic Fatty Liver Disease: From Composition to Function. Clinical
921 Gastroenterology and Hepatology 2019; 17:296–306.

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

45. 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.

46. 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/

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

48. 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.

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

50. 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

51. 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.

52. 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.

53. 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.

54. 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.

55. 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.

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

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

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

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

972 60. Zhao S, Jang C, Liu J, Uehara K, Gilbert M, Izzo L, Zeng X, Trefely S, Fernandez S, Carrer
973 A, et al. Dietary fructose feeds hepatic lipogenesis via microbiota-derived acetate. Nature
974 2020; 579:586–91.

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

979 62. Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, Tremaroli V,
980 Bakker GJ, Attaye I, Pinto-Sietsma S-J, et al. Depicting the composition of gut microbiota
981 in a population with varied ethnic origins but shared geography. Nature Medicine 2018;
982 24:1526–31.

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

985 64. Cheng H-Y, Wang H-Y, Chang W-H, Lin S-C, Chu C-H, Wang T-E, Liu C-C, Shih S-C.
986 Nonalcoholic Fatty Liver Disease: Prevalence, Influence on Age and Sex, and Relationship
987 with Metabolic Syndrome and Insulin Resistance. International Journal of Gerontology
988 2013; 7:194–8.

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

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

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

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

999 69. Eslam M, Sanyal AJ, George J, Sanyal A, Neuschwander-Tetri B, Tiribelli C, Kleiner DE,
1000 Brunt E, Bugianesi E, Yki-Järvinen H, et al. MAFLD: A Consensus-Driven Proposed
1001 Nomenclature for Metabolic Associated Fatty Liver Disease. Gastroenterology 2020;
1002 158:1999-2014.e1.

1003 70. Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T, Männistö S,
1004 Salomaa V, Sundvall J, Puska P. Forty-year trends in cardiovascular risk factors in Finland.
1005 Eur J Public Health 2015; 25:539–46.

1006 71. Marotz L, Schwartz T, Thompson L, Humphrey G, Gogul G, Gaffney J, Amir A, Knight R.
1007 Earth Microbiome Project (EMP) high throughput (HTP) DNA extraction protocol v1
1008 (protocols.io.pdmdi46) [Internet]. 2018 [cited 2020 Nov 10]; Available from:
1009 https://www.protocols.io/view/earth-microbiome-project-emp-high-throughput-htp-d-
1010 pdmdi46

1011 72. Sanders JG, Nurk S, Salido RA, Minich J, Xu ZZ, Zhu Q, Martino C, Fedarko M, Arthur
1012 TD, Chen F, et al. Optimizing sequencing protocols for leaderboard metagenomics by
1013 combining long and short reads. Genome Biology 2019; 20:226.

1014 73. Glenn TC, Nilsen RA, Kieran TJ, Sanders JG, Bayona-Vásquez NJ, Finger JW, Pierson
1015 TW, Bentley KE, Hoffberg SL, Louha S, et al. Adapterama I: universal stubs and primers
1016 for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru &
1017 iNext). PeerJ 2019; 7:e7755.

1018 74. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of
1019 sequencing reads. PeerJ [Internet] 2017 [cited 2020 Nov 10]; 5. Available from:
1020 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5581536/

1021 75. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;
1022 9:357–9.

1023 76. Méric G, Wick RR, Watts SC, Holt KE, Inouye M. Correcting index databases improves
1024 metagenomic studies. bioRxiv 2019; :712166.

1025 77. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete
1026 domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology 2020; :1–8.

1027 78. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification
1028 of metagenomic sequences. Genome Res [Internet] 2016 [cited 2018 May 12]; Available
1029 from: http://genome.cshlp.org/content/early/2016/11/16/gr.210641.116

1030 79. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna,
1031 Austria: R Foundation for Statistical Computing; 2018 [cited 2019 Mar 4]. Available from:
1032 https://www.R-project.org/

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

81. 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/

82. 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.

83. 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

84. 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.

85. 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.

86. 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.

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

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

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

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

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

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

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

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

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

Figures

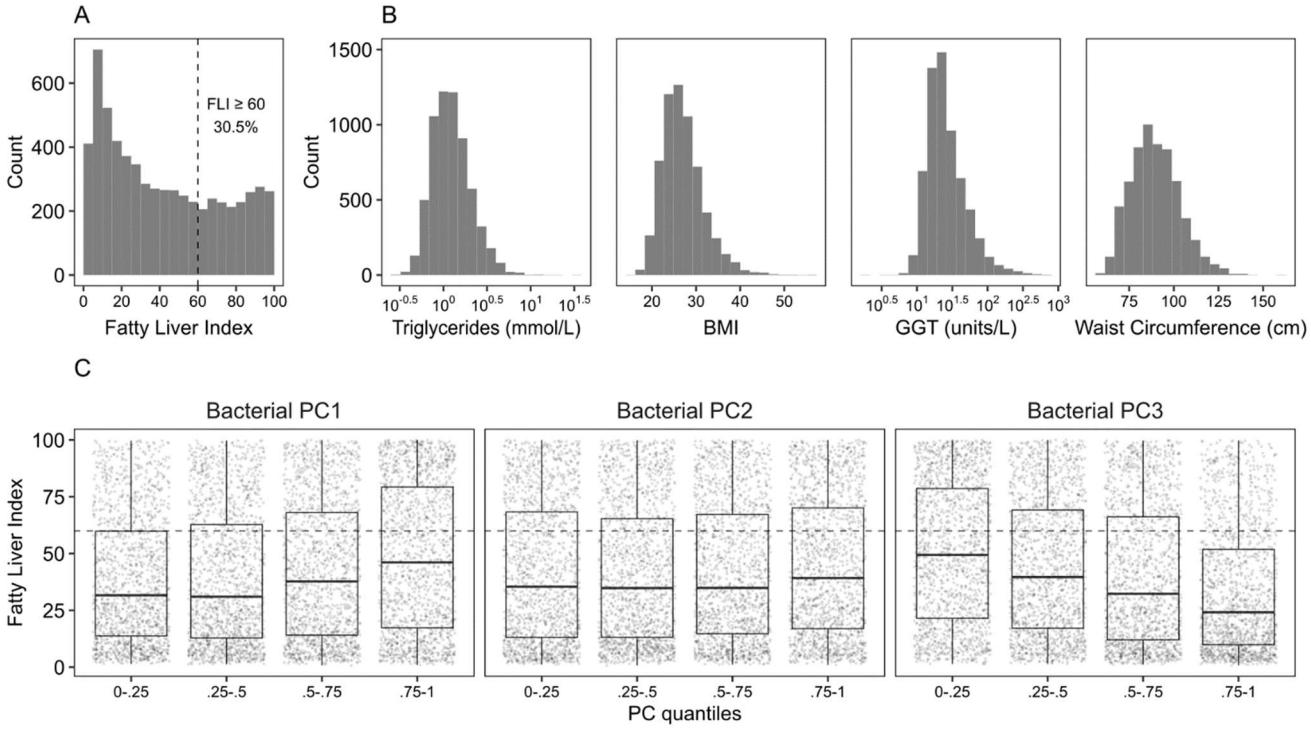
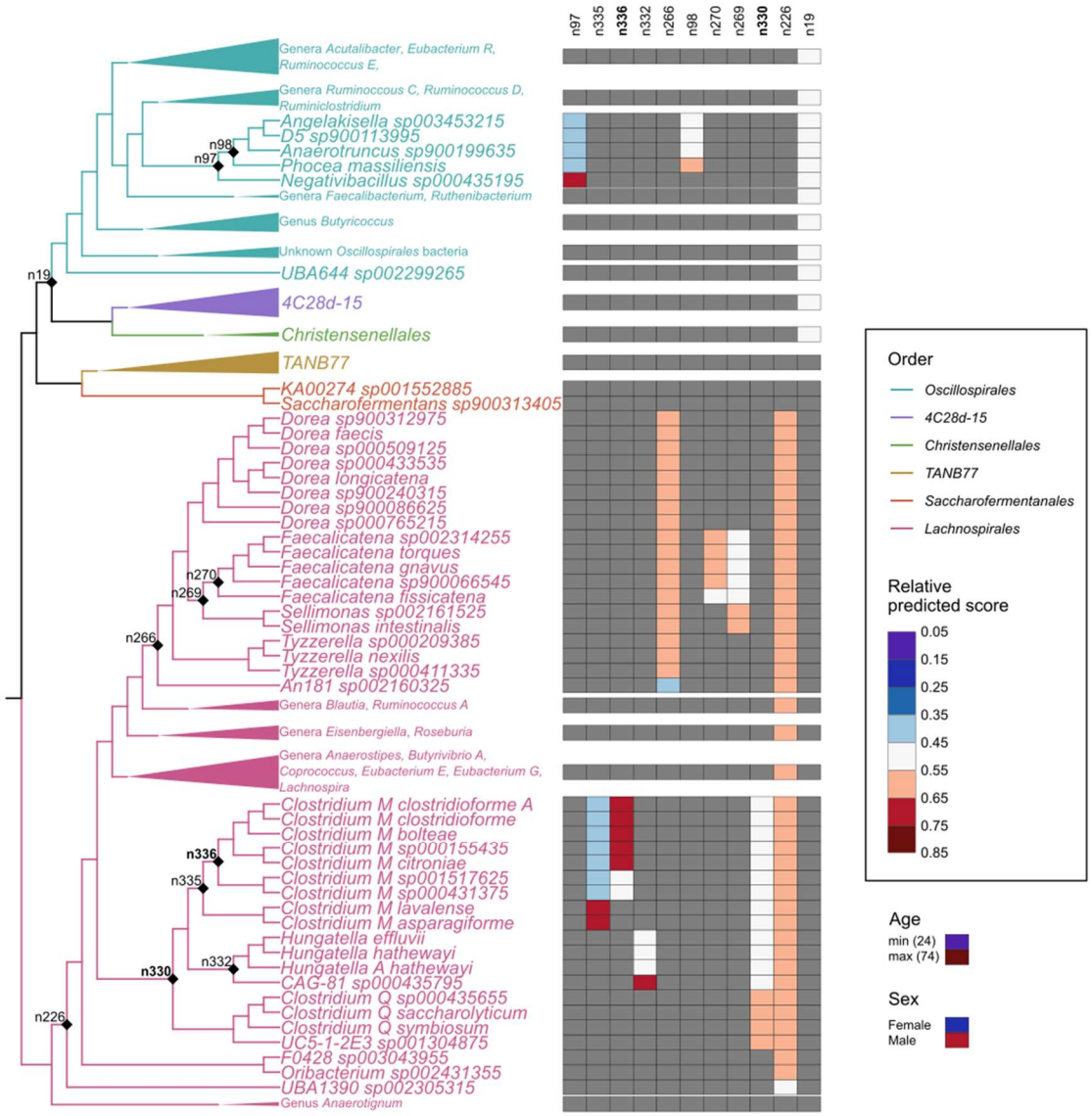


Figure 1. Distribution of FLI (A), its components (B), and FLI in quantiles of the first three PC components of the fecal bacterial composition of the participants (C). The cutoff at FLI = 60 used to divide the participants is indicated with a dashed line in panels A and C.



1083 **Figure 2.** Relative effects of predictive balances and covariates on the FLI < 60 and FLI ≥ 60

1084 classification model (AUC = 0.75) predictions. Nodes of the balances are indicated in the

1085 cladogram and the relative effect sizes of their clades (opposite sides of each balance) are shown

1086 in the associated heatmap. The relative effect sizes of the covariates (age and sex) are shown

1087 below the legend with a heatmap on the same scale as was used for the balances. The two liver-

1088 specific balances associated with triglyceride and GGT levels are indicated with bold font.

1089 Clades with redundant information have been collapsed but their major genera are indicated. The

1090 complete tree is included in **Figure S6**.