Cell Metabolism, Volume *34*

Supplemental information

Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting

Yang Liu, Guillaume Méric, Aki S. Havulinna, Shu Mei Teo, Fredrik Åberg, Matti Ruuskanen, Jon Sanders, Qiyun Zhu, Anupriya Tripathi, Karin Verspoor, Susan Cheng, Mohit Jain, Pekka Jousilahti, Yoshiki Vázquez-Baeza, Rohit Loomba, Leo Lahti, Teemu Niiranen, Veikko Salomaa, Rob Knight, and Michael Inouye

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Baseline gut microbial compositions, related to Table 1. The box plot shows the distribution of relative abundances of the most abundant 20 bacterial taxa at different taxonomic levels across all samples. Samples are represented as dots and colored by follow-up disease status for (a) liver disease and (b) alcoholic liver disease.

Figure S2. Comparison of AUROC of conventional model and species-level gut microbiome models for (a) any liver disease and (b) alcoholic liver disease, related to Figure 2. Difference of AUCs between gradient boosting model and other models were tested using one-sided Wilcoxon-signed rank test. *P<0.05; **P<0.01. Boxplots show the distribution of AUROC scores of the conventional risk factor model, as well as gradient boosting, ridge regression and logistic regression models using species-level gut microbiome. The dots represent data points.

Figure S3. Case-control analysis of NAFLD in UCSD cohort, related to Figure 3. (a-i) Performance of age, sex, BMI, WHR, smoking, TRIG, HDL, LDL and GGT individually in distinguishing NAFDL cases from controls using logistic regression models. (j)Performance of the logistic regression model combining all the above risk factors. (k)Performance of species-level gut microbiome in identifying NAFLD cases using ridge logistic regression. (I)Performance of the logistic regression model incorporating gut microbiome based predictions and conventional risk factors.

Figure S4. Taxonomy tree of gut microbiome signatures, related to Figure 5.

Figure S5. Phylogenetic tree from species-level gut microbial signatures, related to Figure 5. Phylogenetic tree (NJ) from 169 species found to be predictive of liver disease, generated from multiple protein sequence alignments from 120 single copy marker proteins used to define GTDB taxonomy (release 89) in the GTDB project (length= 5040 aa). Species are coloured by their GTDB phylum. The tree scale represents the number of subsitutions per site. Predictive microbial signatures for alcoholic liver disease (ALD) and liver disease (LD) are indicated in red on the right of the tree.

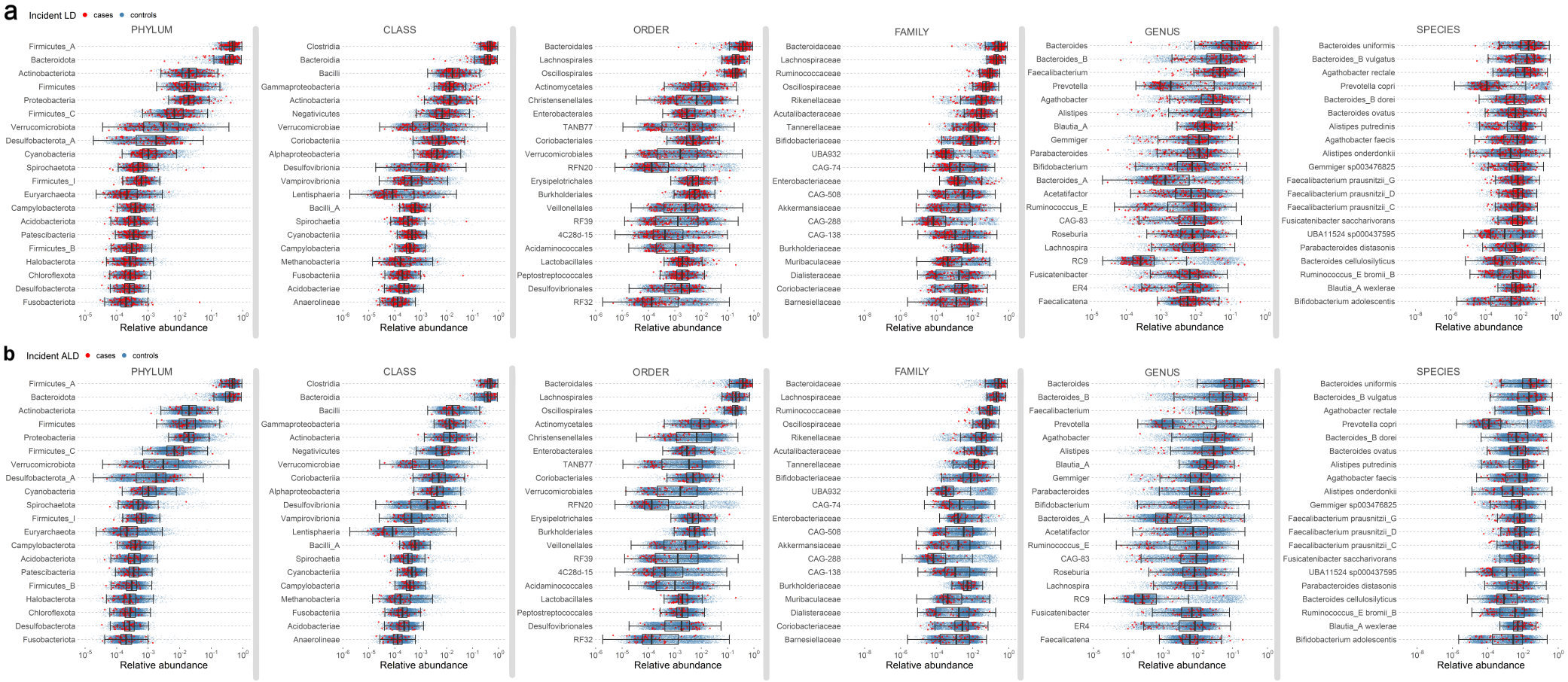

Figure S1 - related to Table 1

Figure S2 - related to Figure 2

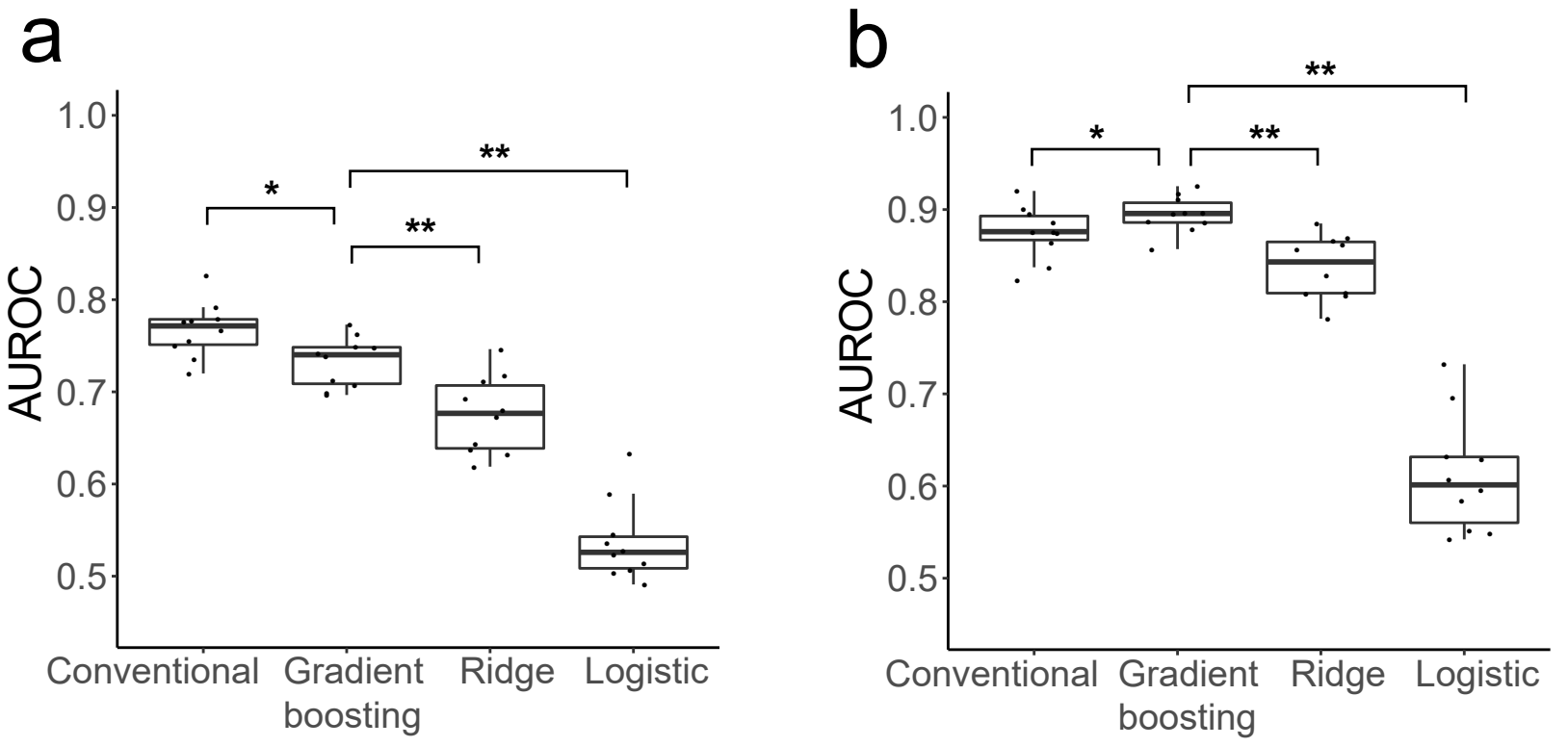

Figure S3 - related to Figure 3

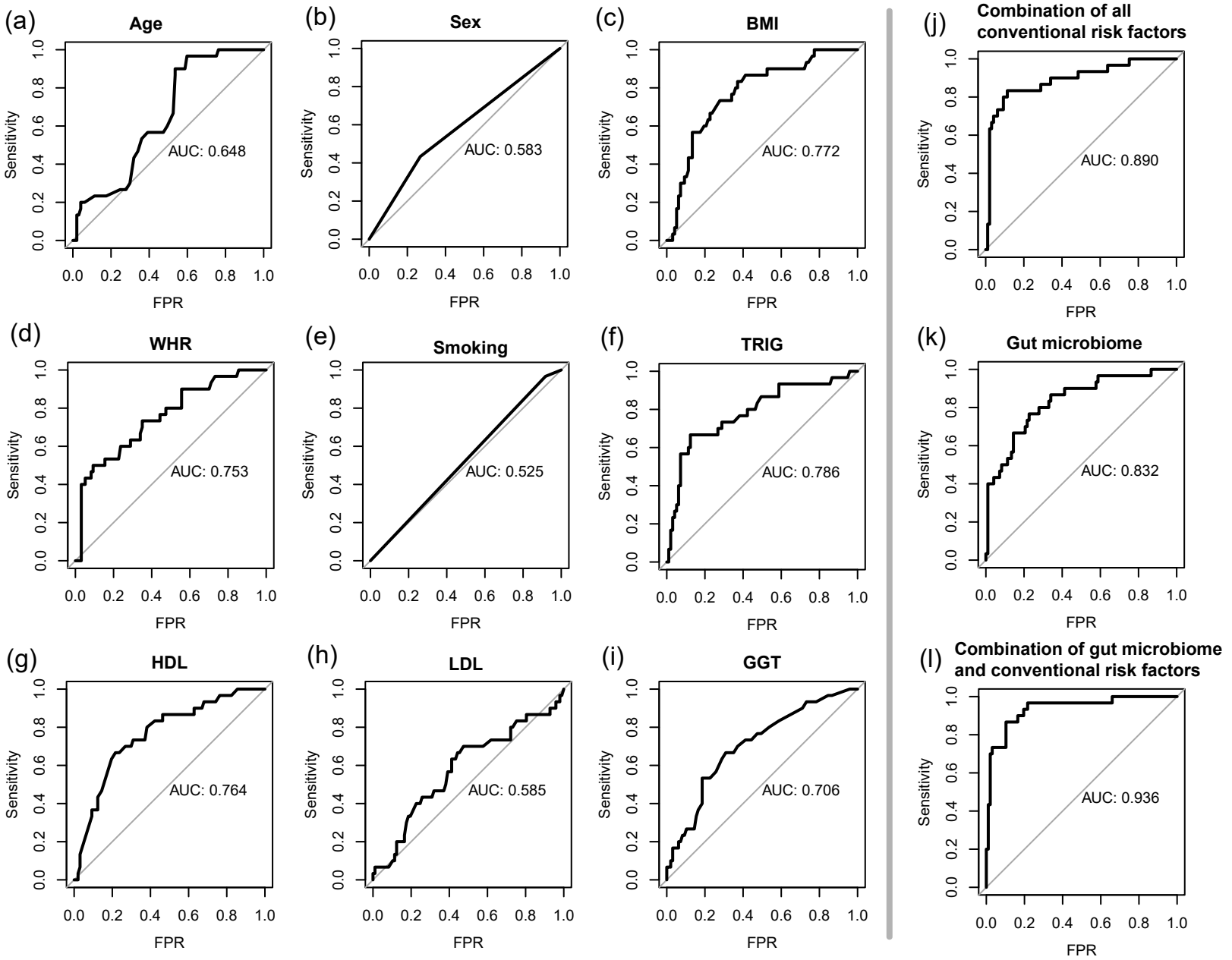

Figure S4 - related to Figure 5

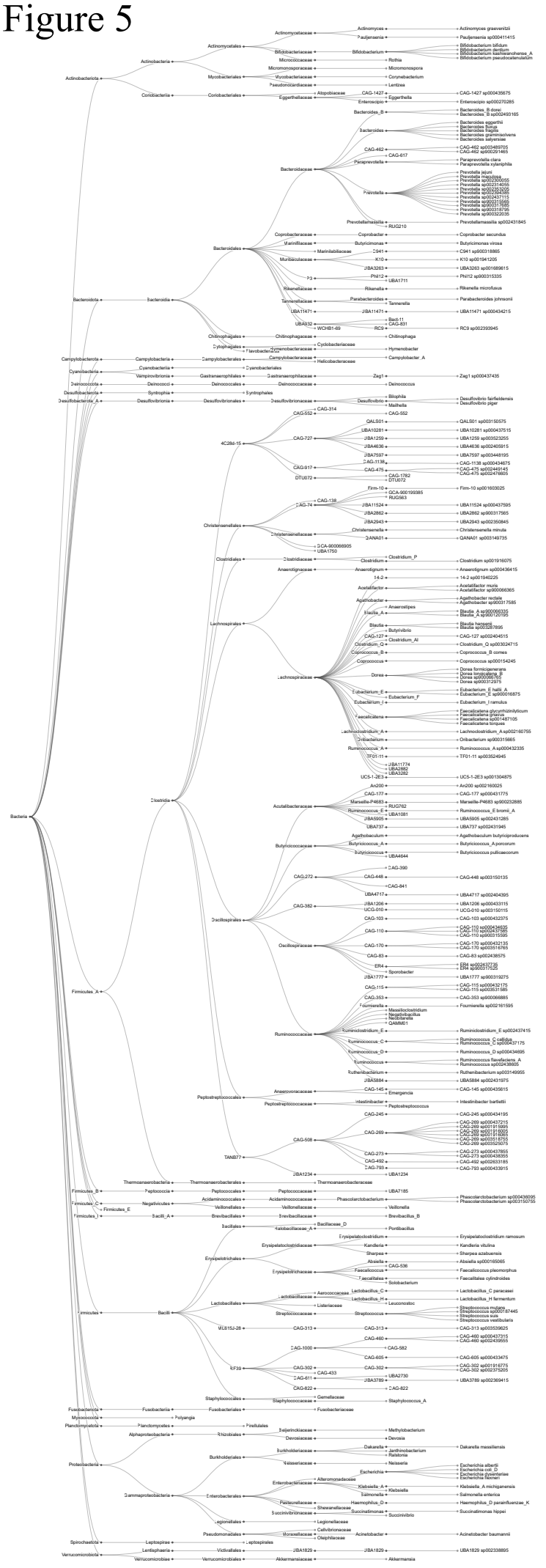

Figure S5 - related to Figure 5

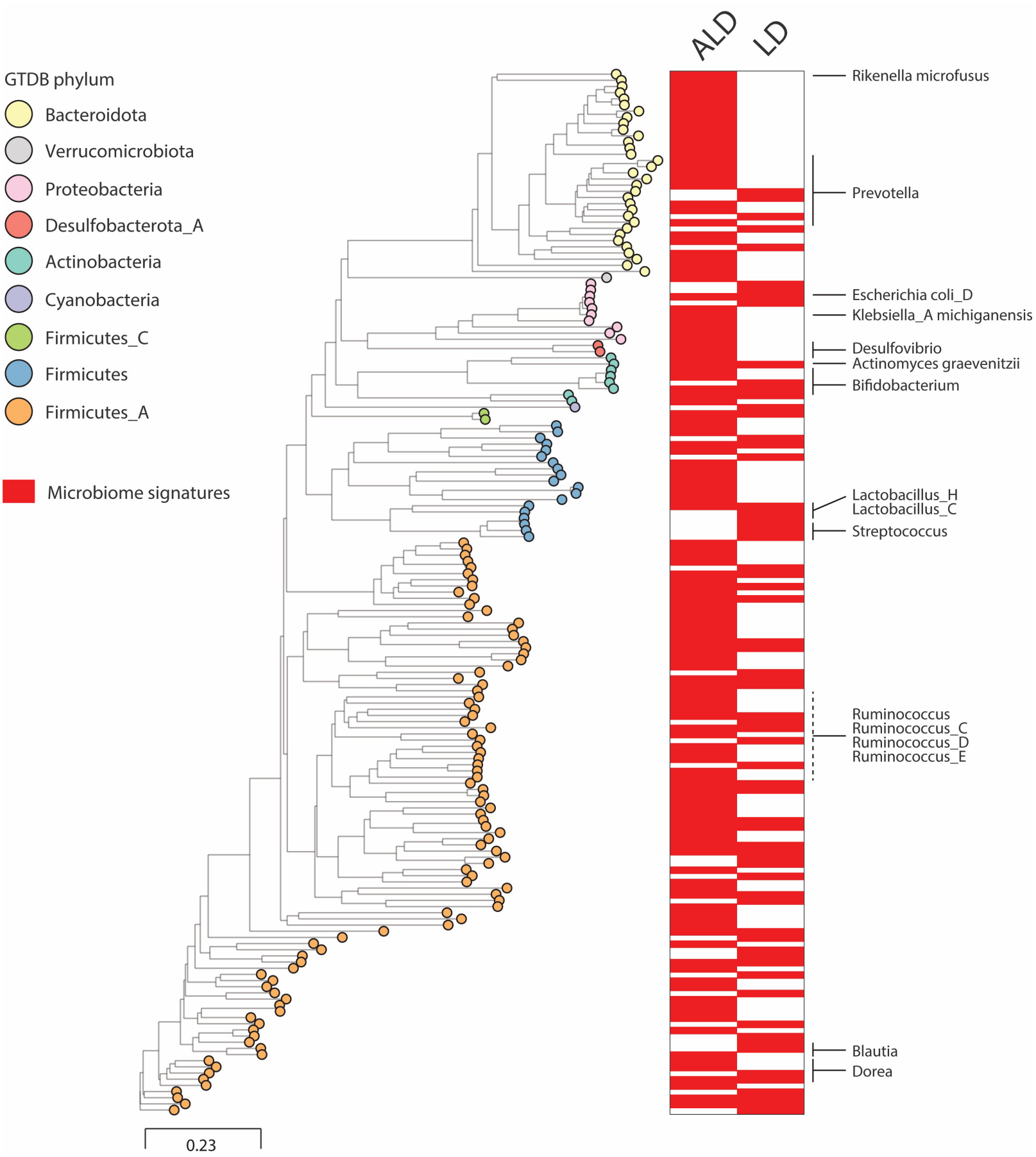

Table S1. Characteristics of NAFLD case-control study cohort, related to Case-Control Study for Validation in STAR Methods.

Characteristics	NAFLD cases (n=30)	Non-NAFLD control (n=97)
Age	58 [49, 63.5]	51 [25, 63.0]
Female (n%)	57%	73%
BMI (kg/m^2)	31.95 [27.43, 36.08]	24.60 [21.90, 28.80]
Waist-hip ratio	0.95 [0.91, 0.98]	0.90 [0.84, 0.93]
Smoking (n%)	3%	8%
HDL cholesterol (mmol/l)	1.29 [1.06, 1.56]	1.76 [1.40, 2.12]
LDL cholesterol (mmol/l)	2.73 [2.17, 3.26]	2.46 [2.04, 3.00]
Triglycerides (mmol/l)	1.61 [0.93, 1.97]	0.80 [0.63, 1.02]
Gamma-glutamyl transferase (U/L)	29 [18.5, 34.25]	16 [12.0, 25.00]

Median [IQR] for continuous variables

Table S3. Composition and main categories of any liver diseases, related to Table 1.

ICD-10 codes	Diseases of liver	Number of incident cases
K70	Alcoholic liver disease	41
K73	Chronic hepatitis, not elsewhere classified	8
K74	Fibrosis and cirrhosis of liver	15
K75	Other inflammatory liver diseases	8
K76	Other diseases of liver	39

*There are samples that have more than one diagnostic code of liver disease and thus the sum does not refer to the total case number of any liver disease.