Supplementary methods:

Liver Magnetic Resonance Spectroscopy (MRS)

A 20 x 20 x 20 mm voxel was placed in the right lobe of the liver, avoiding major vasculature, bile ducts, liver edges, and artifacts. After shimming during free breathing, five single average STEAM spectra (mixing time 5 ms) were acquired consecutively at progressively longer echo times of 10, 15, 20, 25 and 30 ms in a single 21-second breathhold[1]. The echo time range and mixing time were chosen to minimize J coupling effects while allowing T2 correction[2]. A TR of 3,500 ms was chosen to minimize T1 effects. No water or spatial saturation was applied. An anatomic image illustrating the placement of the MRS voxel was saved and spectra were transferred offline for analysis.

The spectra from the individual channels were combined using a singular value decomposition[3]. A single blinded experienced observer analyzed the spectra using the Advanced Method for Accurate, Robust and Efficient Spectral (AMARES) algorithm, included in the MRUI software package[4]. As described previously[1], the T2-corrected areas of the water (4-6 ppm) and the fat (0-3 ppm) were estimated. The contribution to the water peak from neighboring fat peaks (4.2 and 5.2 ppm) was corrected using a previously derived fat spectrum post T2 correction, which reassigned these fat peaks from water to the fat signal and the PDFF was calculated from these corrected peak areas.

Polymerase chain reaction (PCR) amplification of the 16S rRNA gene

Primers used for PCR amplification the V1-V3 region of the 16S rRNA gene were 27F (5'-AATGATACGGCGACCACCGAGATCTACACNNNNNNNNACACTCTTTCCCTACACGA-3') and 534R (5'-CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATTACCGCGGCTGCTGG-3'), where N's represent a unique 8bp index for each sample. The PCR mixture contained 1 µM of each forward and reverse primer, 4 ng template DNA, 0.75 U AccuPrime Taq High Fidelity DNA Polymerase (Life Technologies), and 2 µl AccuPrime Buffer II (Life Technologies) in a final volume of 20 µl. Thermal cycling consisted of an initial denaturation step at 95°C for 2 min, followed by 30 cycles of denaturation at 95°C for 20s, annealing at 56°C

for 30 s and extension at 72 °C for 60s, with a final extension step at 72°C for 5 min. PCR product purification and sequencing were carried out as described in the main text.

Bioinformatic processing of 16S rRNA gene sequence data

Following sequencing, reads were trimmed to remove adapters and low quality 3' sequence using Trimmomatic v0.321[5], with the commands HEADCROP:20, TRAILING:10, MINLEN:100. Assembly of paired-end reads into single contiguous sequences (contigs) covering the V1-V3 region was done using PEAR v0.9.102[6]. To remove spurious assemblies contigs containing ambiguous bases (N's) were discarded, as were contigs shorter than 448bp or longer than 529bp (these values approximated the 1st and 99th percentile of the contig length distribution observed across all samples). Any contig showing similarity to the PhiX genome using Blastn (e<1x10-5) was also removed[7].

Contigs passing filtering were pooled, and OTUs generated following the UPARSE pipeline[8] (usearch v8.0.1517). Clustering of a unique set of contigs was performed using the UPARSE-OTU algorithm and an additional chimera removal step was carried out using UCHIME in conjunction with the ChimeraSlayer reference database. Contigs from all samples were then assigned to a single OTU using the USEARCH global algorithm at a 97% similarity threshold[9]. Any contig that did not match an OTU at this threshold was assumed to be a sequencing or assembly artifact and was discarded.

Relative abundance estimates were calculated based on the number of contigs assigned to each OTU. Abundance estimates were adjusted to account for differences in sequencing effort by normalizing total counts for each sample to the median sequencing depth. Normalized data were filtered to remove any OTU that was detected in less than 5% of samples and whose maximum relative abundance constituted less than 5% of the counts detected in any one sample. Normalization and filtering were performed using the R package Phyloseq[10].

Taxonomic classification of OTUs was performed using the Ribosomal Database Project (RDP) classifier v2.2[11] in conjunction with the RDP reference database and using an 80% confidence threshold.

Classifying individuals into discrete groups based on *Prevotella copri* relative abundance

Preliminary investigation of the abundance of the genus Prevotella across individuals indicated a bimodal log-normal distribution. Comparable distributions were not seen for any other taxon. The Prevotella genus has previously been identified as a major determinant of enterotypes[12]. However, this community-based approach to portioning microbiomes has recently come under criticism[13]. Multiple OTUs in this study were identified as belonging to the Prevotella genus, but those responsible for driving the bimodal trend observed in this study (Fig. 1C Supplementary Fig. 5) were identified as *P. copri* (see methods below). Therefore, in light of recent criticism of enterotypes, we chose to group individuals into discrete categories based on *P. copri* abundance alone.

The RDP classifier only provides taxonomic classification to genus level. To infer species-level classification for OTUs belonging to the genera *Bacteroides* and *Prevotella* all 16S gene sequences for these two genera were downloaded from the NCBI collection of complete bacterial genomes. Downloaded sequences were randomly subsampled to select up to three representative, full-length, 16S gene sequences for all available *Bacteroides* and *Prevotella* species. Selected sequences were then aligned separately for each genus using MUSCLE[14], and a phylogenetic tree constructed using FastTree10[15]. Representative sequences for each Bacteroides and Prevotella OTU (spanning the V1-V3 region of the 16S) were then positioned within their respective phylogeny using the Quantitative Insights Into Microbial Ecology (QIIME)[16] command make_phylogeny.py, and the proximity of each OTU to full-length gene sequences of known taxonomic origin was used to infer species (Supplementary Figure 4).

Multiple *Prevotella* OTUs appeared to be most closely related to two *P. copri* 16S gene reference sequences extracted from a single *P. copri* National Center for Biotechnology Information (NCBI) reference genome assembly (Supplementary Figure 4). Therefore, in order to estimate species-level taxonomic quantification, the abundance of all OTUs in this arm of the tree were summed and used to indicate *P. copri* relative abundance.

A Gaussian finite mixture model was fitted for the natural log of *P. copri* abundance using the Expectation Maximization (EM) algorithm in the R package mclust[17]. Three components were identified by the model. Based on cutoff points, data were therefore categorized into low, medium, and high *P. copri* groups. Mean + 3 SD of component 1 was used to obtain the cutoff

point between low and medium while mean - 3 SD of component 3 for medium and high (Supplementary Figure 5).

Bioinformatic processing of metagenomic whole genome shotgun (mWGS) sequence data

Following sequencing, mWGS data were processed using the MOCAT[18] pipeline to remove sequencing adapters and screen reads for host contamination (using the hg19 reference genome assembly) This resulted in an average of 16,195,254 read pairs per sample (minimum 9,875,881, maximum 27,365,397). Cleaned sequences were mapped to the UniRef90 reference gene database provided by HUMAnN2[19] and normalized to counts per kilobase per million reads mapped (cpm) using the script humann2_renorm_table.py.

The HUMAnN2 software package provides an internal reference for mapping UniRef90[20] gene families to their respective KO groups, which is based on database cross- references provided by the Universal Protein Resource (UniProt). Preliminary data investigation identified deficiencies in UniProt database cross-reference annotations that were causing artifactual results during the analysis of KOs. For example, the Prevotella gene for S-adenosylmethionine synthase (R6XFI0_9BACT) was not annotated as mapping to its corresponding KO (K00789, see www.uniprot.org/uniprot/R6XFI0.txt), in spite of the fact that the sequence belongs within this orthologous group (https://www.ncbi.nlm.nih.gov/protein/EFB35257.1). To avoid such database artifacts influencing functional analysis results, UniRef90 centroid sequences were mapped directly to the KEGG[21] protein database.

Representative protein sequences for all detected UniRef90 gene families were extracted from the HUMAnN2 UniRef90 diamond[22] indexed database (v0.11.0) and mapped to the KEGG protein sequence database (release 2015-08-31) using the USEARCH -ublast command with the parameters –evalue 1e-9 –accel 0.5 –maxhits 1. Of the UniRef90 gene families detected in this study 82% matched an entry in the KEGG gene database and 52% of detected KEGG genes could be assigned to a KO. The relative abundance of KOs was calculated by summing the normalized counts of each contributing UniRef gene family.

Effect of *P. copri* abundance on reference gene database mapping success: No significant differences were observed when comparing the proportion of reads that could be mapped by HUMAnN2 between cases and controls. However, significant differences in the proportion of

reads mapped by HUMAnN2 were observed when comparing individuals classified as belonging to the low, intermediate, or high *P. copri* groups (Supplementary Figure 11E). This difference in mapping success indicated the possibility that reference gene databases were biased, and contained fewer gene sequences representative of the high *P. copri* metagenome.

In more detail – both UniRef90 gene families and KOs are clusters of orthologous sequences. Representational bias could therefore mean that a single orthologous sequence cluster contains more reference sequences representative of a low *P. copri* microbiome than a high *P. copri* microbiome. Differential abundance in the number of reads mapping to such an orthologous cluster would not then represent differential abundance in the gene function represented by the cluster, rather it would represent a failure of reads from high *P. copri* samples to map to gene sequences that are absent from the database.

To avoid potential artefacts due to gene reference database bias, functional comparisons based on Gene Set Enrichment Analysis (GSEA were carried out separately for high and low *P. copri* groups when the conditions under analysis (for example, absent/mild vs. moderate/severe fibrosis) was conflated with differences in the distribution of high and low *P. copri* individuals.

Gene set enrichment analysis: KOs are clusters of genes that share a similar function, while KEGG Pathways are lists of KOs that contribute to a single, well-studied metabolic pathway. To investigate alteration in microbial metabolic pathways with NAFLD, all KOs detected in the intestinal metagenome were ranked based on the significance of their change in abundance between conditions (cases vs. controls, NASH vs. NAFLD, absent/mild vs. moderate/severe fibrosis). KEGG pathways were downloaded as part of the KEGG database (release 2015-08-31) and filtered to retain 148 pathways for which ≥5 contributing KOs were detected in our mWGS dataset. Gene set enrichment analysis was then performed as described in the main text[23].

Alcohol metabolism and other gene pathways of *a priori* interest: To explicitly address existing hypotheses relating to the gut microbiome and NAFLD, two online databases – KEGG and the Gene Ontology database[24] (www.geneontology.org) – were manually searched. Gene pathways reflecting Iron(III) transport system (M00190), Secondary bile acid biosynthesis (map00121), ethanol metabolic processes (GO:0006067), choline biosynthetic processes (GO:0042425), and short-chain fatty acid metabolic processes (GO:0046459) were selected.

When read-coverage allowed, the relative abundance of the UniRef90 gene clusters/KOs within each pathway was assessed using principal coordinates analysis, and differences between cases vs. controls, NAFLD vs. NASH, and absent-to-mild vs. moderate-to-severe fibrosis tested using permutational multivariate analysis of variance. No signal metagenomic signal was detected for pathways reflecting choline biosynthetic processes and ethanol metabolic processes all other pathways were non-significant in pairwise contrasts ($p>0.05$).

To further address the possibility that microbial alcohol metabolism contributes to NAFLD, serum ethanol levels were measured using an Ethanol Colorimetric Assay Kit (BioVision Inc., Milpitas, CA). 10µl duplicate aliquots of 1:10 diluted serum in PBS were assayed following kit directions. Results showed no significant variation in serum alcohol levels in comparisons of cases vs. controls, NASH vs. NAFLD, or moderate/severe vs absent/mild fibrosis.

Contribution of bacterial genera to pro-inflammatory pathways: To identify the likely taxonomic origin of genes contributing to the lipopolysaccharide biosynthesis (ko00540) and flagellar assembly (ko02040) pathways, genus-level count information was extracted from HUMAnN2 output and gene abundance estimates for each genus collapsed. The percent contribution of different genera to the total read coverage for every UniRef90 gene family/KO was then calculated. The contribution of each genus to KOs in the lipopolysaccharide biosynthesis pathway is shown in Supplementary Figures 12-13. The contribution of each genus to KOs in the flagellar assembly pathway is shown in Supplementary Figures 14-15.

Indicator values[25] were calculated in order to summarize the extent to which each bacterial taxon contributed to either cases vs. controls, NAFLD vs. NASH, or absent/mild vs. moderate/severe fibrosis. Pathways were initially filtered to remove KOs found in <4 genera, and genera were filtered to remove any genus contributing <10 KOs in a pathway (the former threshold was based on preliminary exploration of data, which indicated a small subset of KOs in both pathways represented by only one or few taxa). Following filtering, the contribution of each genus to an entire pathway was summarized as the 75^{th} percentile of its contribution to each of the remaining KOs. Indicator values were then calculated using the R package labdsv[26], including 1000 permutations to estimate the probability of encountering a higher indicator value by chance.

Supplementary references

1. Hamilton G, Yokoo T, Bydder M, et al. In vivo characterization of the liver fat (1)H MR spectrum. NMR Biomed 2011;24:784-90.
2. Hamilton G, Middleton MS, Bydder M, et al. Effect of PRESS and STEAM sequences on magnetic resonance spectroscopic liver fat quantification. J Magn Reson Imaging 2009;30:145-52.
3. Bydder M, Hamilton G, Yokoo T, et al. Optimal phased-array combination for spectroscopy. Magn Reson Imaging 2008;26:847-50.
4. Naressi A, Couturier C, Castang I, et al. Java-based graphical user interface for MRUI, a software package for quantitation of in vivo/medical magnetic resonance spectroscopy signals. Comput Biol Med 2001;31:269-86.
5. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114-20.
6. Zhang J, Kobert K, Flouri T, et al. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 2014;30:614-20.
7. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;215:403-10.
8. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 2013;10:996-8.
9. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26:2460-1.
10. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 2013;8:e61217.
11. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 2007;73:5261-7.
12. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature 2011;473:174-80.

13. Gorvitovskaia A, Holmes SP, Huse SM. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome 2016;4:15.
14. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792-7.
15. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 2010;5:e9490.
16. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7:335-6.
17. Fraley C, Raftery AE, Murphy TB, et al. mclust Vetsion 4 for R: Normal mixture modelling for model-based clusering, classification and density estimation. University of Washington 2012:1-50.
18. Kultima JR, Coelho LP, Forslund K, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics 2016;32:2520-3.
19. Abubucker S, Segata N, Goll J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 2012;8:e1002358.
20. Suzek BE, Wang Y, Huang H, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31:926-32.
21. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27-30.
22. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;12:59-60.
23. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
24. The Gene Ontology C. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47:D330-D338.
25. Dufrene M, Legendre P. Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecological Monographs 1997;67:345-366.
26. Roberts DW. labdsv: Ordination and multivariate analysis for ecology, 2016.

Supplementary tables:

Supplementary table 1: Results of statistical tests to compare microbial taxonomic abundance between NAFLD cases and obese controls.

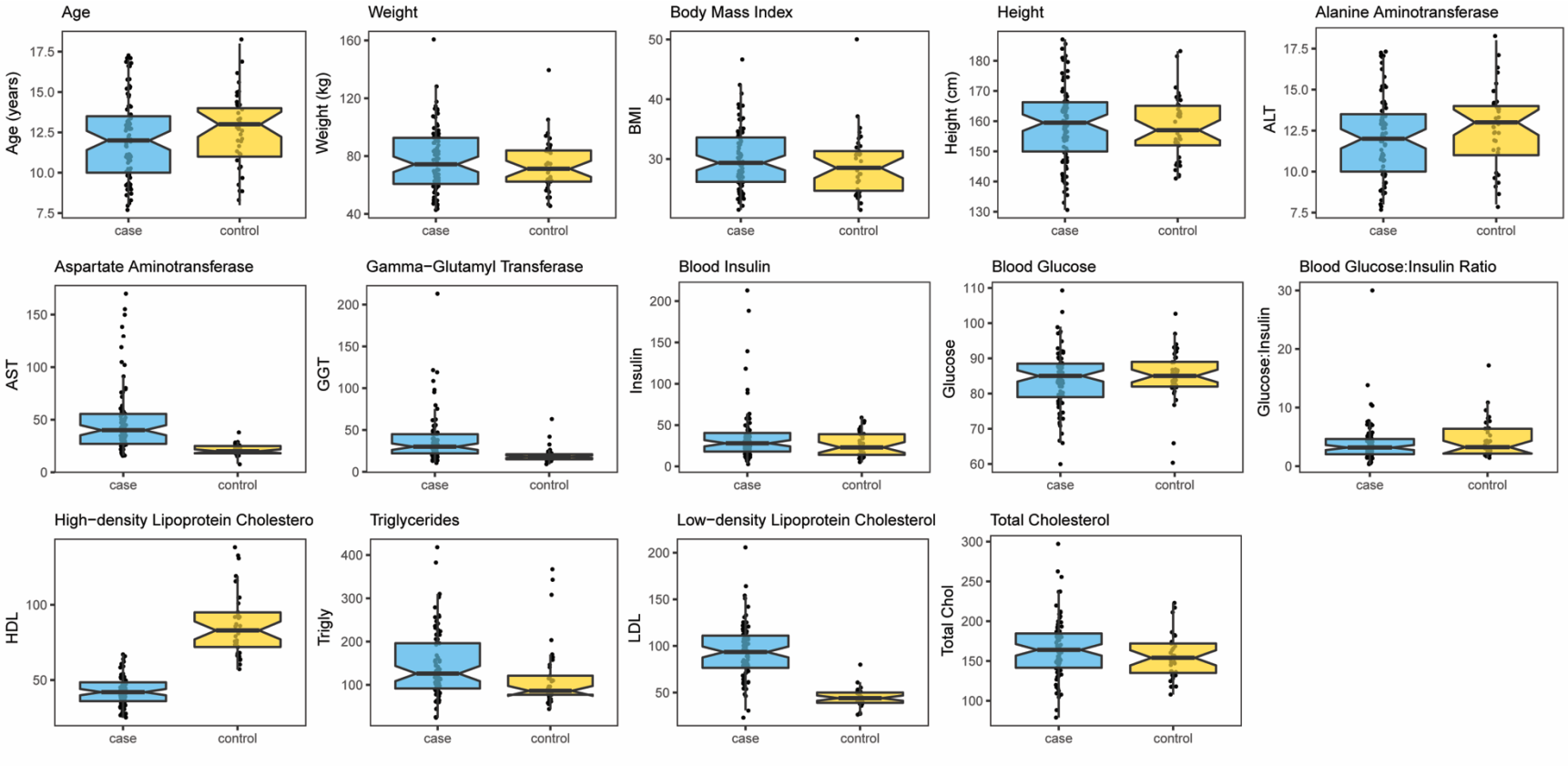
[Due to its size, this table is supplied as an external Microsoft Excel file.]

Supplementary table 2: Patient demographic and clinical characteristics by *P. copri* group. [a]Chi-Square or Fisher's exact test; [b]Kruskall-Wallis test; [c]Calculated basted on the 2000 Centers for Disease Control and revention (CDC) growth charts.

	Low P. copri (n=73)	Medium P. copri (n=16)	High P. copri (n=35)	P value
Patients				
Cases	49(67)	11(69)	27(77)	0.56
Gender				
Male	44(60)	11(69)	24(69)	0.64[a]
Ethnic				
Hispanic	57(78)	14(88)	34(97)	0.025[a]
Race				
White	33(45)	7(44)	12(34)	0.55[a]
Age	12(11,14)	13(11,15)	12(10,14)	0.92[b]
Height (cm)	159(147,166)	162(155,169)	160(154,165)	0.56[b]
Weight (kg)	69(61,86)	83(60,93)	77(65,93)	0.30[b]
BMI	28(25,31)	30(25,34)	31(26,34)	0.093[b]
BMI percentile[c]	98(96,99)	98(97,99)	99(98,99)	0.048[b]
BMI Z score[c]	2(2,2)	2(2,2)	2(2,2)	0.048[b]
ALT	35(19,69)	33(17,64)	42(22,88)	0.39[b]
AST	27(21,43)	34(21,45)	33(24,52)	0.53[b]
GGT	22(17,32)	30(16,34)	32(22,45)	0.075[b]
Insulin	22(15,36)	31(19,41)	31(22,44)	0.058[b]
Glucose	85(80,90)	86(79,90)	84(80,88)	0.29[b]
MRS Liver Fat	10(4,21)	16(5,20)	14(5,23)	0.81[b]
Body Fat	41(37,46)	41(36,47)	45(41,47)	0.10[b]
Triglyceride	115(82,161)	110(80,154)	115(94,196)	0.73[b]
Total cholesterol	160(139,183)	158(137,178)	161(138,180)	0.96[b]
HDL cholesterol	44(40,50)	38(33,44)	40(36,48)	0.027[b]
LDL cholesterol	89(72,109)	99(72,119)	88(78,105)	0.70[b]

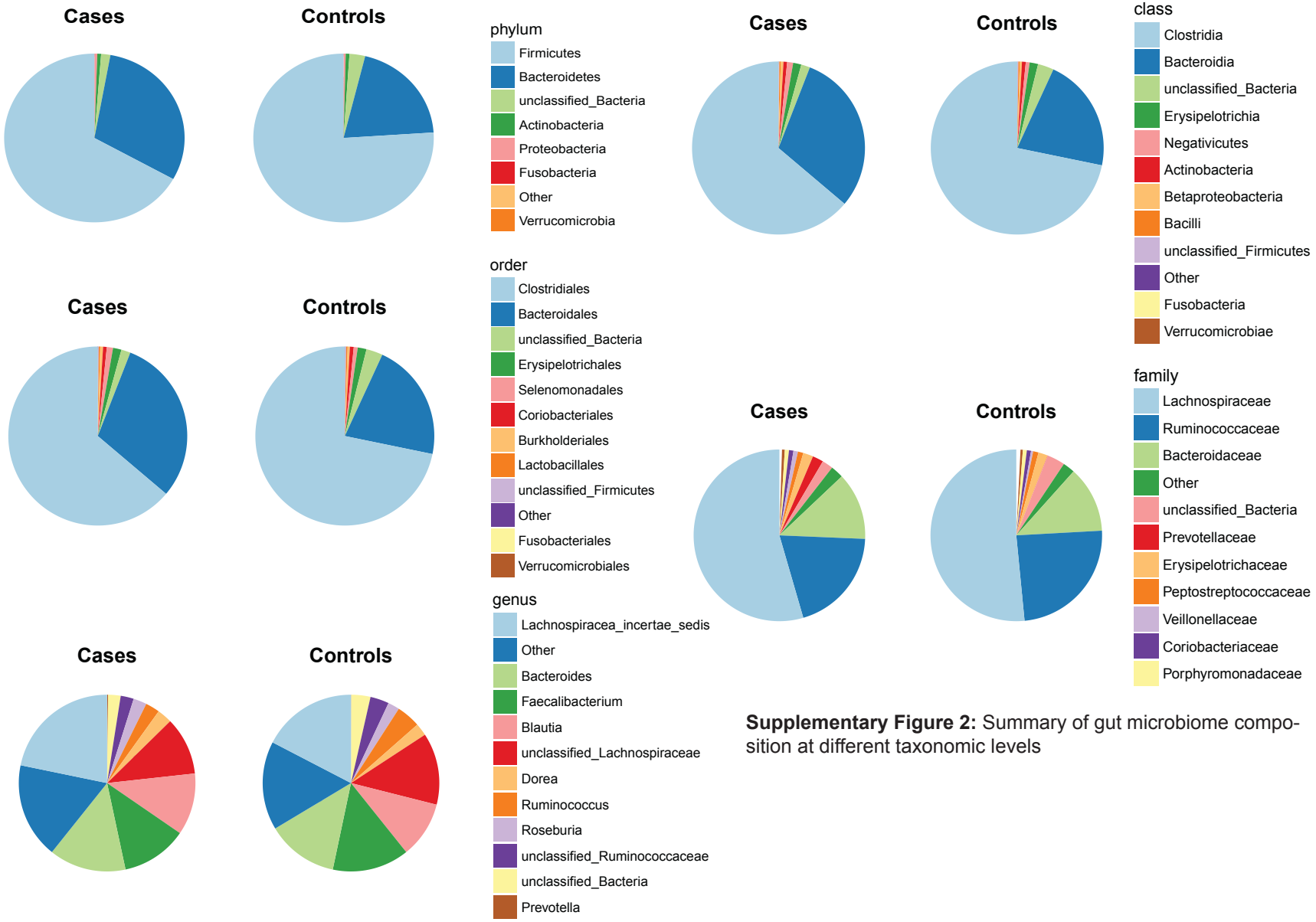
*Data are presented as median (interquartile range) or n (%).

Supplementary figures:

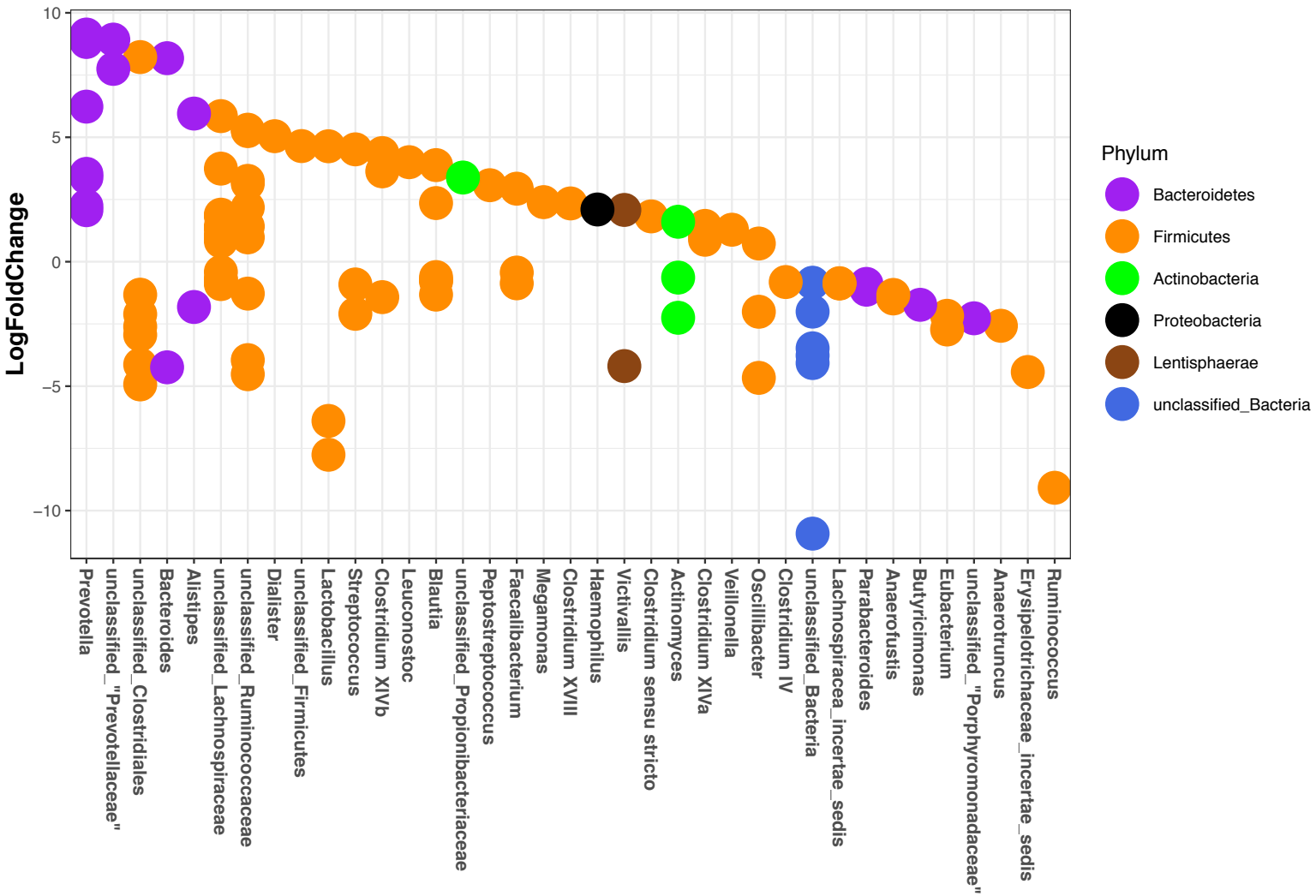


Supplementary figure 1: Descriptive statistics of the study population

Boxplots showing the distribution of demographic and clinical features in NAFLD cases and controls.



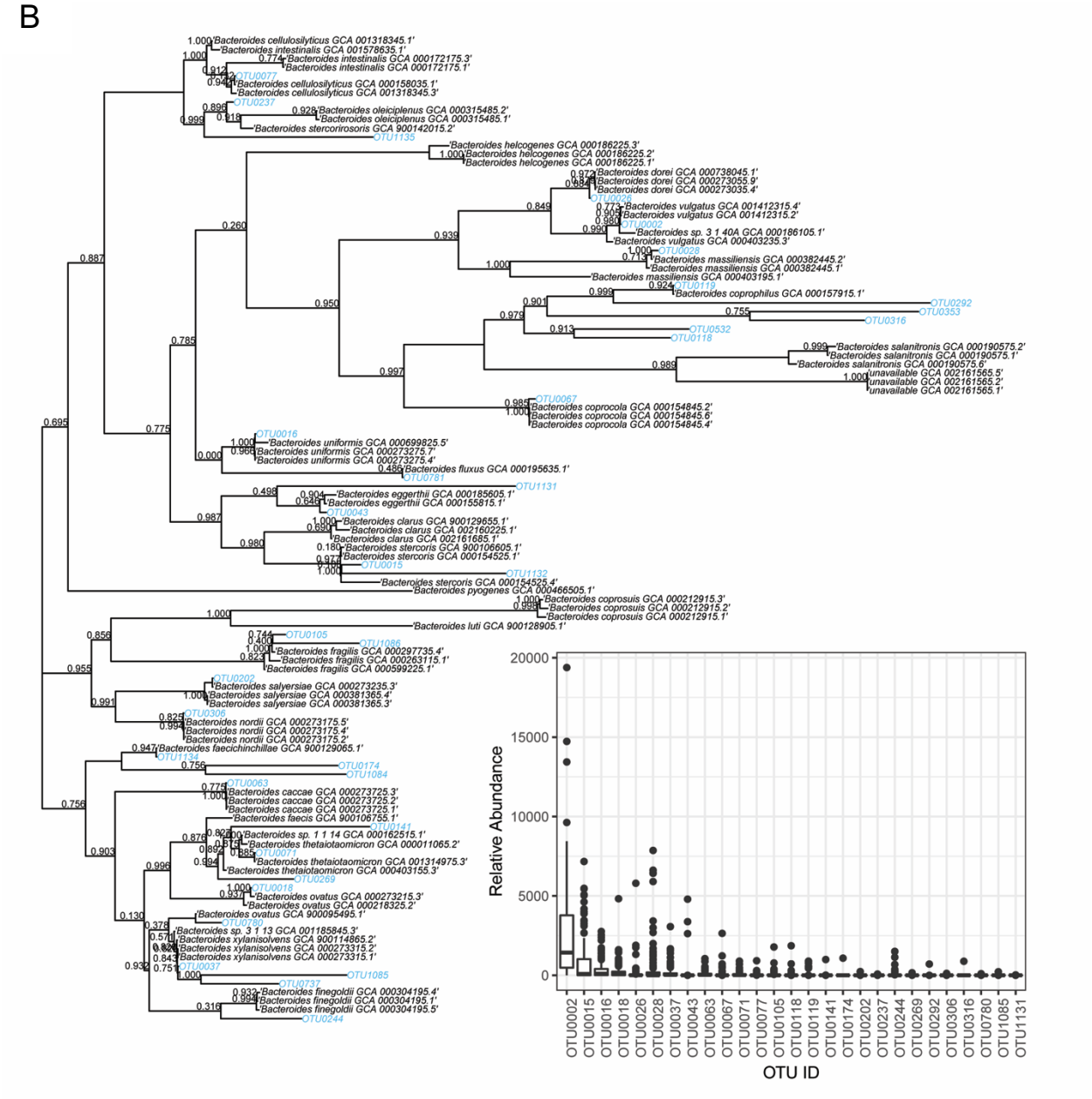
Supplementary Figure 2: Summary of gut microbiome composition at different taxonomic levels



Supplementary figure 3: OTUs whose relative abundance differs between NAFLD cases and controls

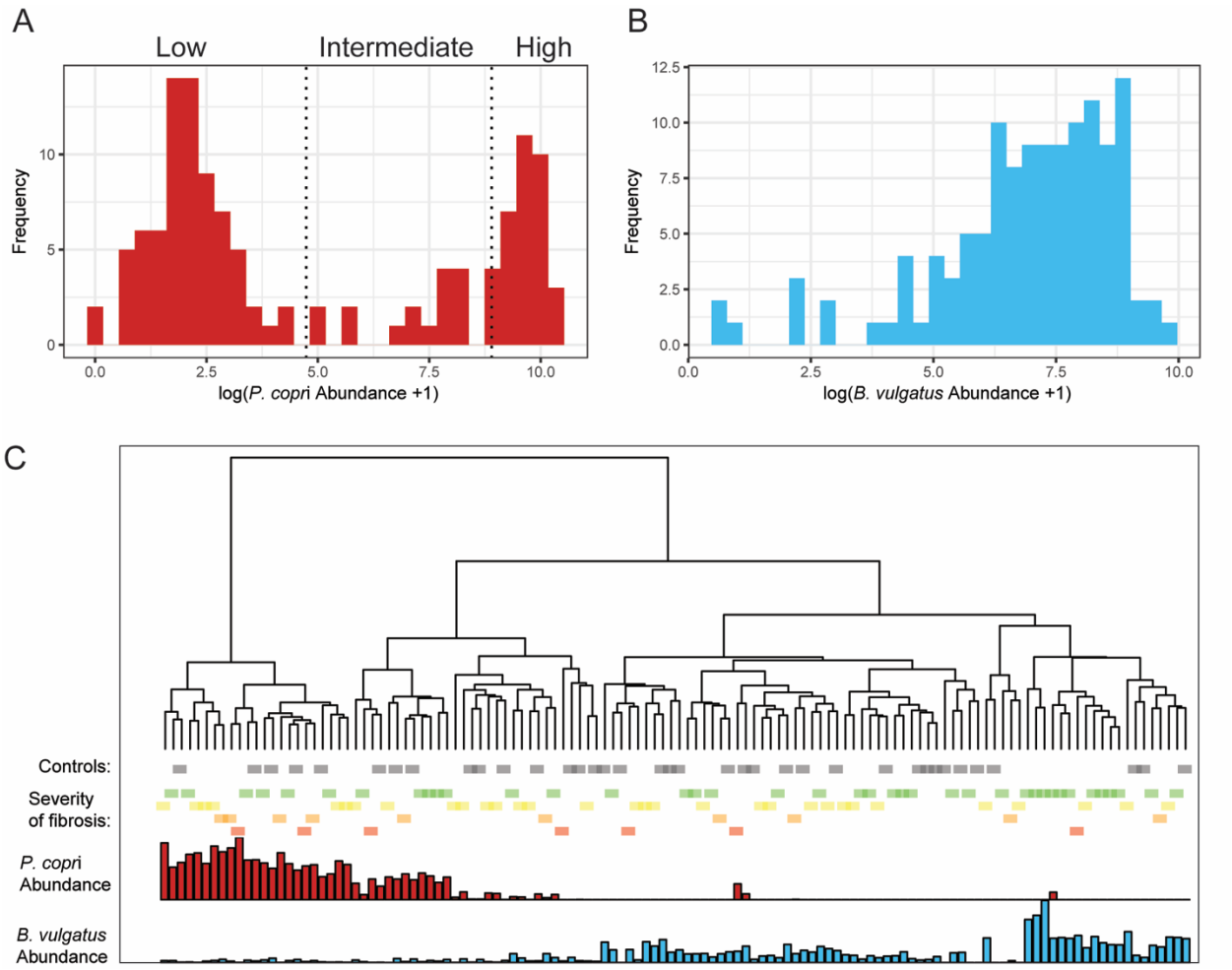
Differential abundance of OTUs was determined using zero-inflated negative binomial/poisson tests (FDR-adjusted p<0.2). Each dot reflects the log2-fold change for a single OTU. OTUs are grouped by genus and coloured by their respective phylum.


'Prevotella micans GCA 000243035.1'
'Prevotella oulorum GCA 000224615.1'
OTU0190
OTU0733
'Prevotella copri GCA 000157935.6'
OTU1127
OTU0870
OTU1129
OTU1130
OTU0086
OTU0004
'Prevotella copri GCA 000157935.7'
OTU0753
OTU1126
OTU0115
OTU0673
OTU1039
OTU1037
OTU0097
OTU0123
OTU0264
'Prevotella bryantii GCA 000179055.5'
'Prevotella bryantii GCA 000179055.10'
'Prevotella bryantii GCA 000179055.3'
'Prevotella bivia GCA 001546565.1'
'Prevotella bivia GCA 000759245.1'
'Prevotella bivia GCA 000262545.1'
OTU1071
'Prevotella amnii GCA 001553225.1'
'Prevotella amnii GCA 000759315.1'
'Prevotella nigrescens GCA 000507825.3'
'Prevotella nigrescens GCA 000507825.2'
'Prevotella nigrescens GCA 000507825.1'
'Prevotella pallens GCA 000220255.1'
'Prevotella intermedia GCA 001953935.4'
'Prevotella intermedia GCA 000261025.4'
'Prevotella intermedia GCA 000261025.1'
OTU1133
'Prevotella disiens GCA 000179675.5'
'Prevotella disiens GCA 000179675.4'
'Prevotella disiens GCA 000179675.1'
'Prevotella sp. HMSC069G02 GCA 001814855.1'
'Prevotella oris GCA 000162915.1'
'Prevotella maculosa GCA 000243015.3'
'Prevotella maculosa GCA 000243015.2'
'Prevotella maculosa GCA 000243015.1'
'Prevotella salivae GCA 000477535.1'
'Prevotella salivae GCA 000185845.1'
'Prevotella sp. DNF00663 GCA 001553265.1'
'Prevotella baroniae GCA 000468635.1'
'Prevotella buccae GCA 000184945.1'
'unavailable GCA 900113305.1'
'Prevotella dentalis GCA 000242335.2'
'Prevotella dentalis GCA 000242335.1'
'Prevotella dentalis GCA 000220215.1'
'Prevotella multisaccharivorax GCA 000218235.4'
'Prevotella multisaccharivorax GCA 000218235.1'
'Prevotella multisaccharivorax GCA 000218235.3'
'Prevotella bergensis GCA 000160535.1'
OTU0734
'Prevotella scopos GCA 900129535.1'
'Prevotella scopos GCA 001683355.3'
'Prevotella scopos GCA 001683355.2'
'Prevotella veroralis GCA 000162935.5'
'Prevotella histicola GCA 000234055.1'
'Prevotella melaninogenica GCA 000759305.1'
'Prevotella melaninogenica GCA 000144405.3'
'Prevotella melaninogenica GCA 000144405.1'
'Prevotella fusca GCA 001262015.3'
'Prevotella fusca GCA 001262015.1'
'Prevotella fusca GCA 001262015.4'
'Prevotella denticola GCA 000193395.2'
'Prevotella denticola GCA 000193395.1'
'Prevotella denticola GCA 000193395.3'
'Prevotella multiformis GCA 000191065.1'
OTU0624
OTU0423
OTU0285
'Prevotella ruminicola GCA 000025925.4'
'Prevotella ruminicola GCA 000025925.2'
'Prevotella ruminicola GCA 900142525.1'
OTU0104
OTU0676
OTU0056
'Prevotella sp. HMSC073D09 GCA 001814145.1'
'Prevotella loescheii GCA 000698925.1'
'Prevotella marshii GCA 000146675.1'
'Prevotella pleuritidis GCA 000468135.1'
'Prevotella enoeca GCA 001444445.3'
'Prevotella enoeca GCA 001444445.2'
'Prevotella enoeca GCA 001444445.1'
'Prevotella oralis GCA 000507905.4'
'Prevotella oralis GCA 000507905.2'
'Prevotella oralis GCA 000185145.2'
'unavailable GCA 001915685.1'
'Prevotella timonensis GCA 000762405.1'
'Prevotella timonensis GCA 000177055.1'
'Prevotella buccalis GCA 000759265.1'
Relative Abundance
25000
20000
15000
10000
5000
0
OTU0004
OTU0056
OTU0086
OTU0097
OTU0104
OTU0115
OTU0123
OTU0190
OTU0264
OTU0673
OTU0676
OTU0733
OTU0734
OTU0870
OTU1037
OTU1039
OTU1127
OTU ID



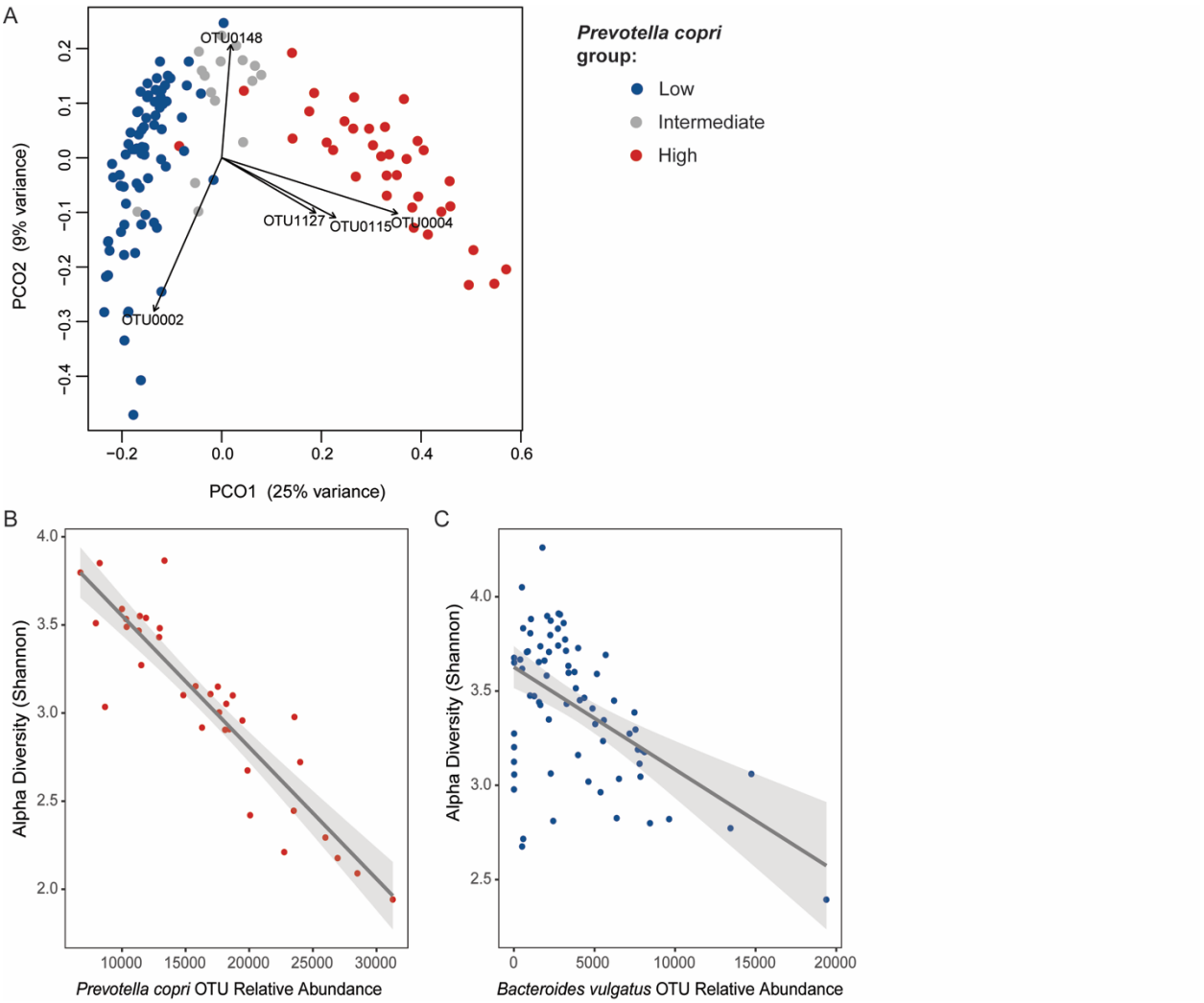
Supplementary figure 4: Taxonomy of OTUs belonging to the genera *Bacteroides* and *Prevotella*

Maximum-likelihood trees showing the phylogenetic position of representative OTU sequences belonging to the genera (A) *Prevotella* and (B) *Bacteroides*. Trees also contain complete representative 16S gene sequences extracted from genome assemblies downloaded from the NCBI RefSeq database. Inset boxplots reflect the distribution in the relative abundance of each OTU across all samples following 16S data normalization. Arrow indicates the node after which all OTUs were assumed to be *P. copri*.



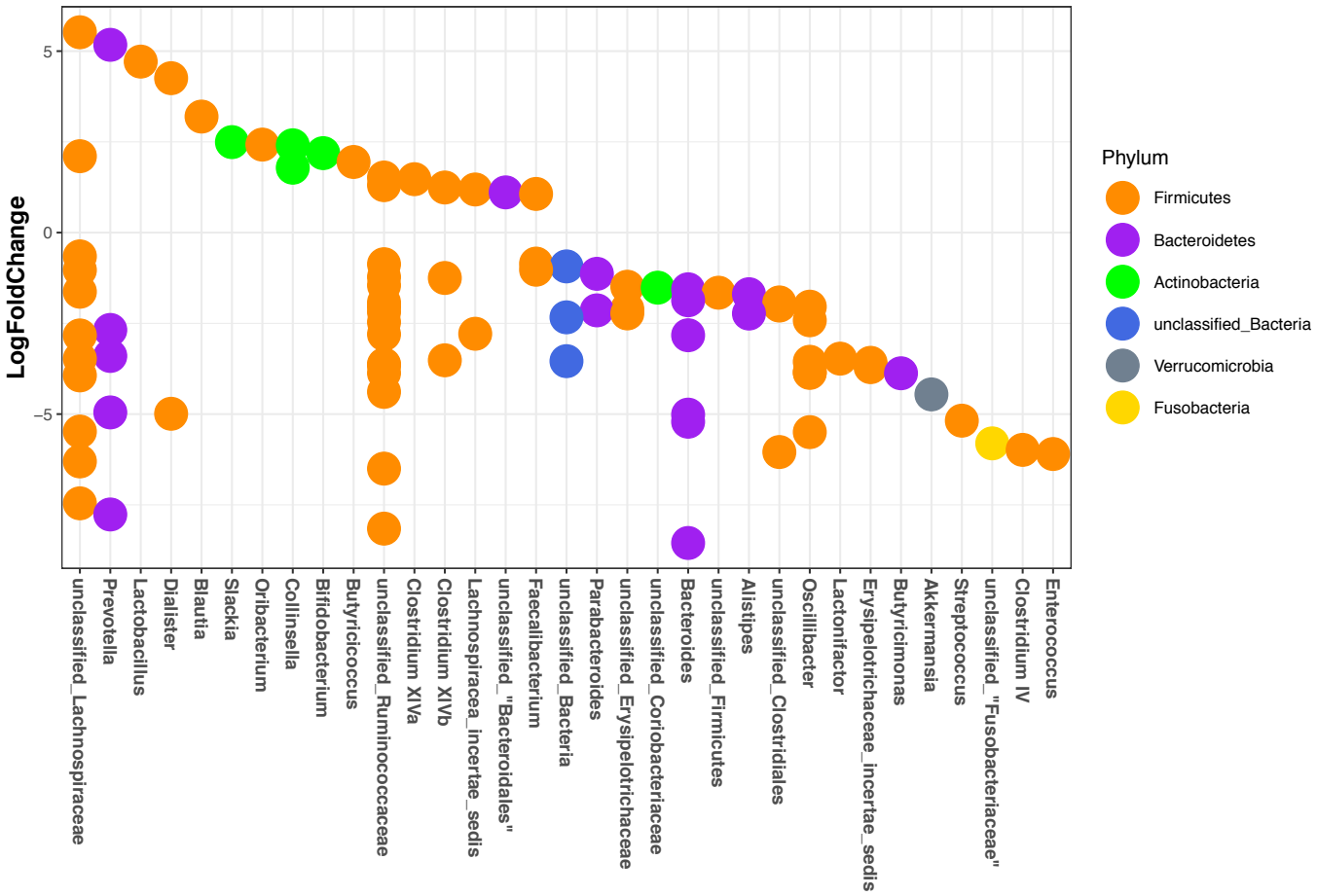
Supplementary figure 5: Classification of individuals based on relative abundance of *P. copri* in their gut microbiome

As an alternative to the use of enterotypes, individuals were assigned to discrete categories (high, indeterminate, low) based on the relative abundance of *P. copri* in their gut microbiome. (A) Histogram showing the distribution of the log-transformed relative abundance of *P. copri* across individuals. Dotted lines denote thresholds at which individuals were classified as either low, intermediate, or high *P. copri*. (B) Histogram showing the distribution of the log-transformed relative abundance of *Bacteroides vulgatus* across individuals. (C) Dendrogram showing hierarchical clustering of individuals based on composition of their gut microbiome. Clustering was based on the Bray-Curtis dissimilarity matrix used in Fig. 1C. Below dendrogram individuals are classified as either controls, or on the basis of severity of fibrosis. The colour scheme denoting severity of fibrosis is the same as that used in Fig. 2B. The relative abundance of *B. vulgatus* and *P. copri* in each individual is also shown.



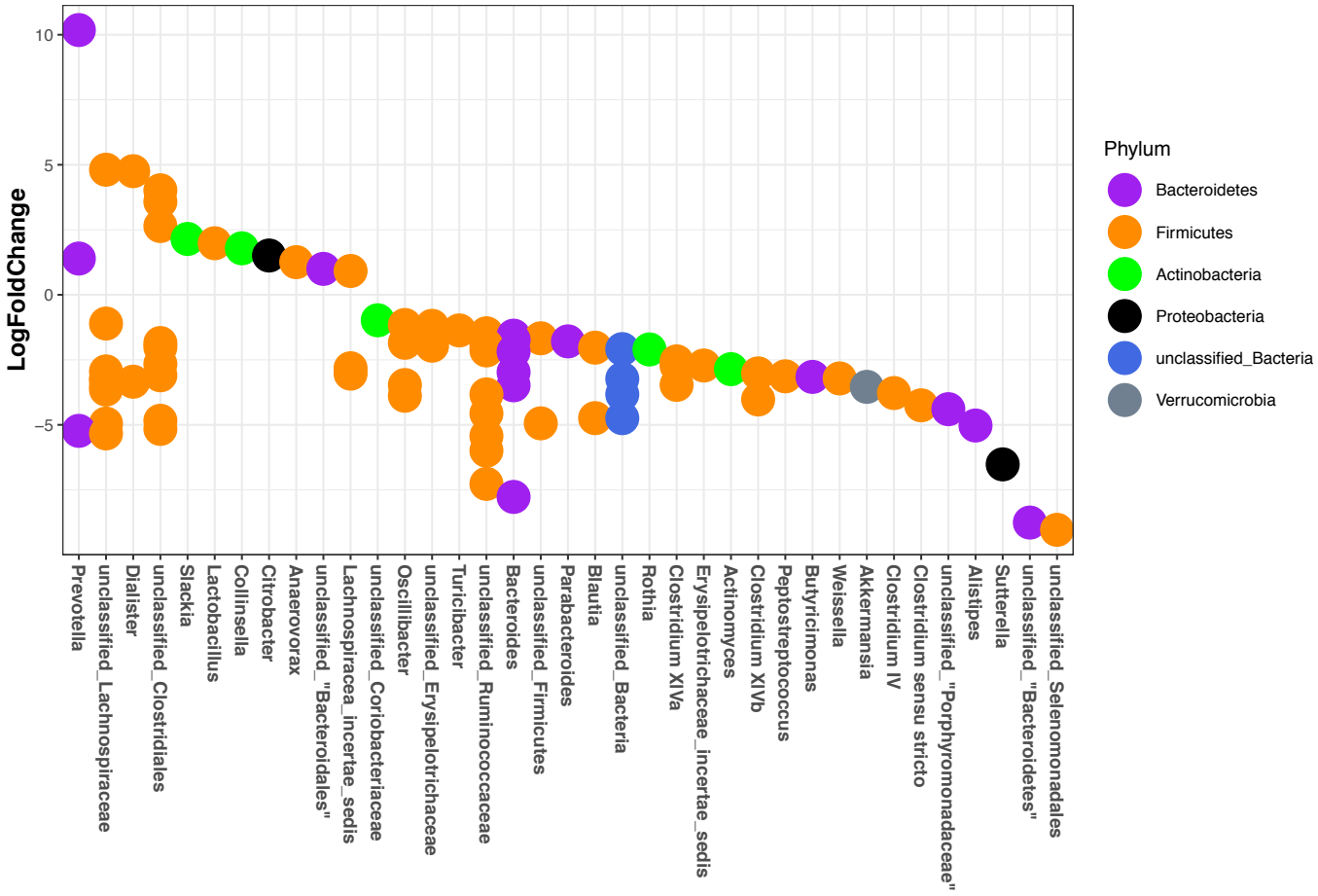
Supplementary figure 6: The relationship between *B. vulgatus* and *P. copri* relative abundance and diversity

(A) Principal coordinates analysis plot showing inter-individual variation in composition of the gut microbiome (beta diversity), as calculated as in Fig. 1C. Individuals are coloured by *P. copri* group (high, indeterminate, low). For the identity of OTUs in overlaid vectors, see Supplementary Fig. 4. (B) The relationship between *P. copri* relative abundance and alpha diversity, shown for individuals classified as high *P. copri*. The linear regression is significant at p<0.0001. (C) The relationship between *B. vulgatus* relative abundance and alpha diversity, shown for individuals classified as low *P. copri*. Linear regression is significant (p<0.0001), and remains significant even if samples with >10000 *B. vulgatus* counts are removed. Grey regions indicate 95% confidence intervals.



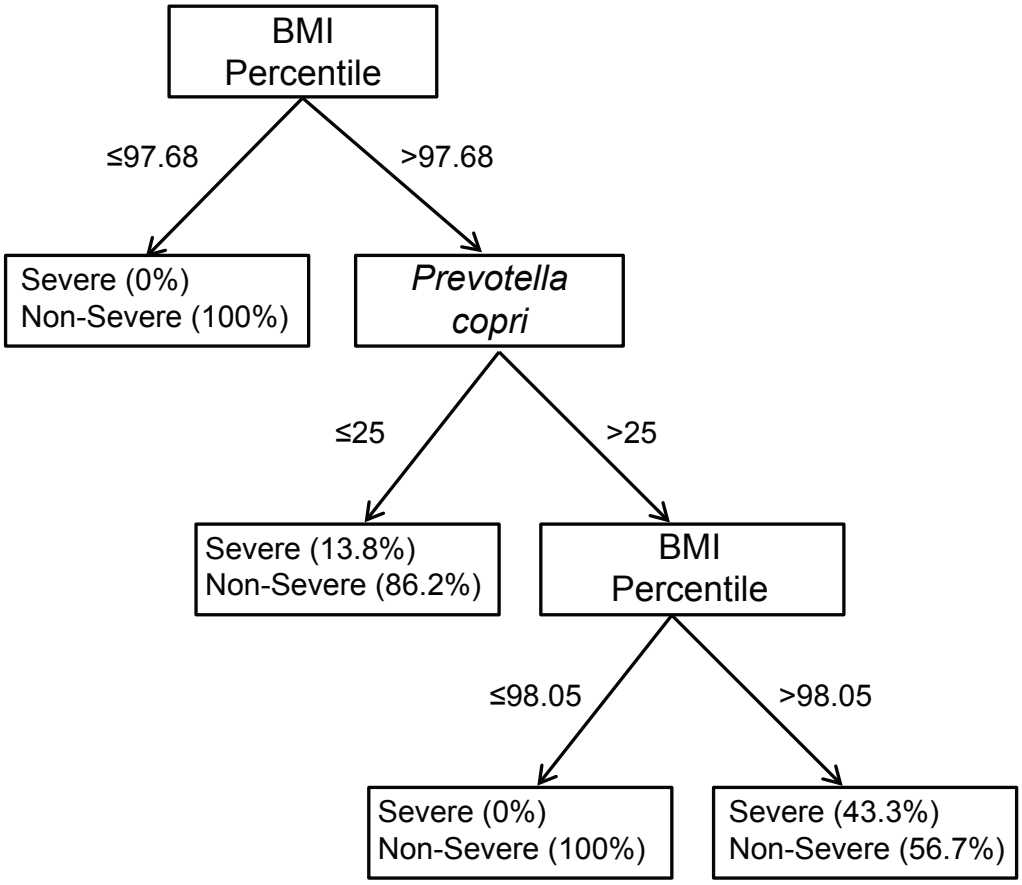
Supplementary figure 7: OTUs whose relative abundance differs between cases with NAFLD not NASH and cases with definite NASH

Differential abundance of OTUs was determined using zero-inflated negative binomial/poisson tests (FDR-adjusted p<0.2). Each dot reflects the log2-fold change for a single OTU. OTUs are grouped by genus and coloured by their respective phylum.



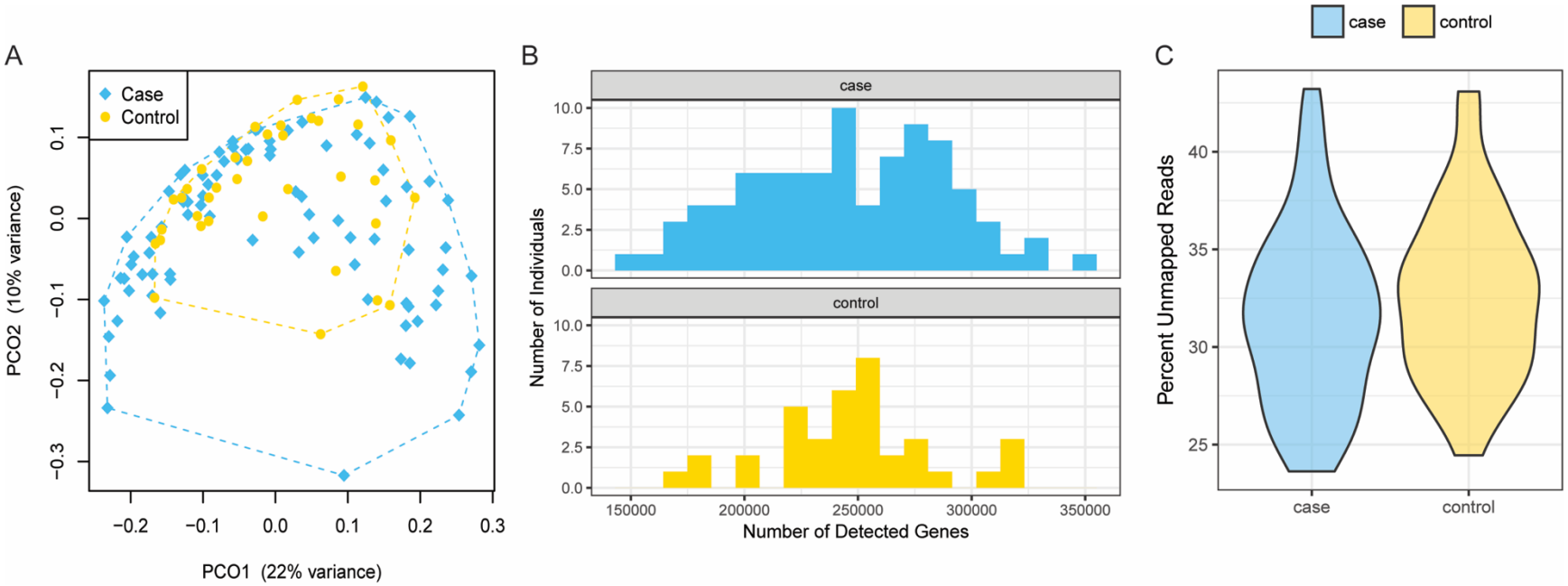
Supplementary figure 8: OTUs whose relative abundance differs between cases with absent-to-mild vs. moderate-to-severe fibrosis

Differential abundance of OTUs was determined using zero-inflated negative binomial/poisson tests (FDR-adjusted p<0.2). Each dot reflects the log2-fold change for a single OTU. OTUs are grouped by genus and coloured by their respective phylum.



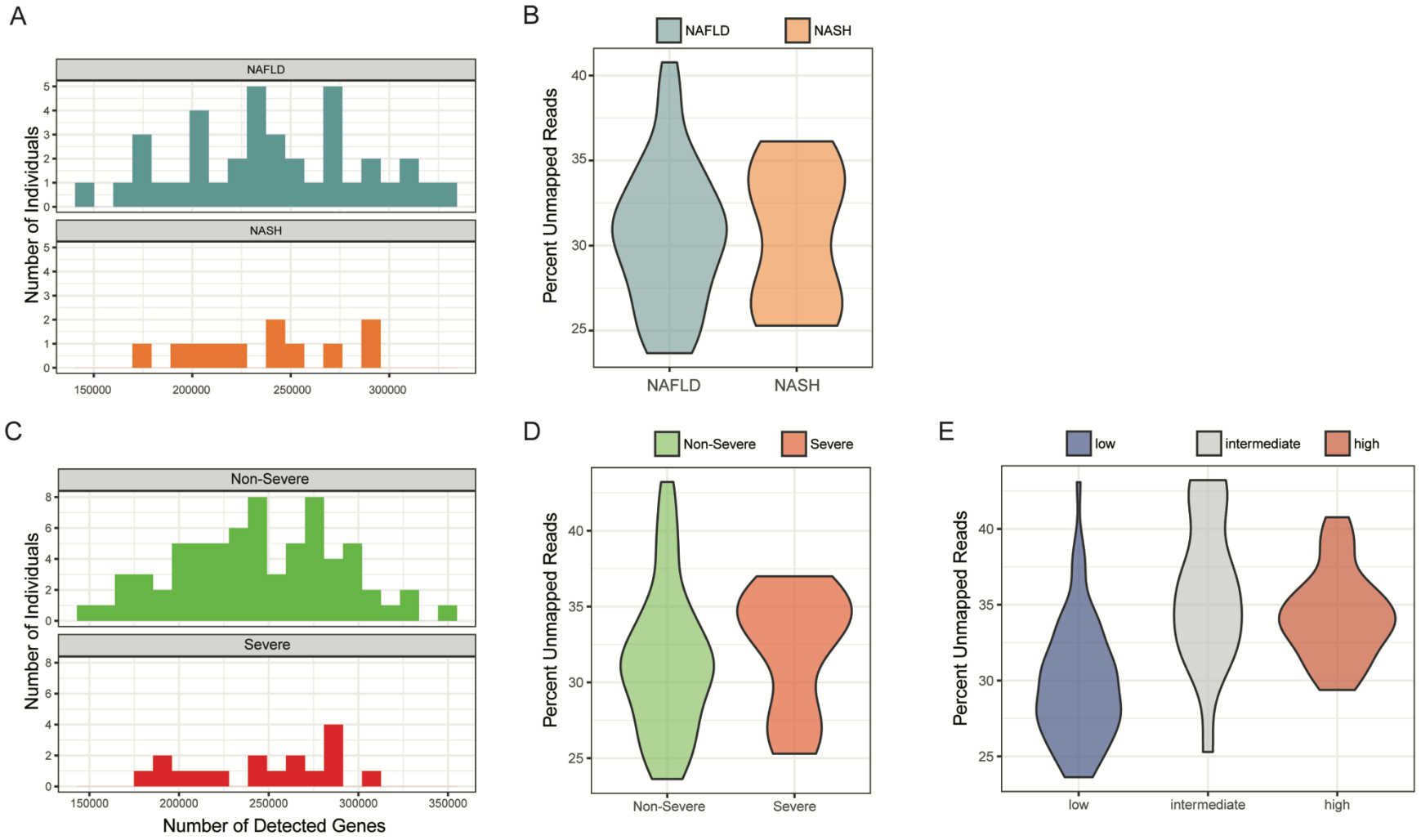
Supplementary Figure 9: CART predicting severity of fibrosis based on *P. copri*, *B. vulgatus* and alpha diversity

Classification and Regression Tree (CART) analysis results showing optimal partitioning of cases with absent/mild fibrosis from cases with moderate/severe fibrosis. Predictor variables included in this analysis were *P. copri* relative abundance, *B. vulgatus* relative abundance, Shannon diversity, age in months, sex, BMI, and ethnicity.



Supplementary Figure 10: Metagenomic signature of the case vs. control gut microbiome

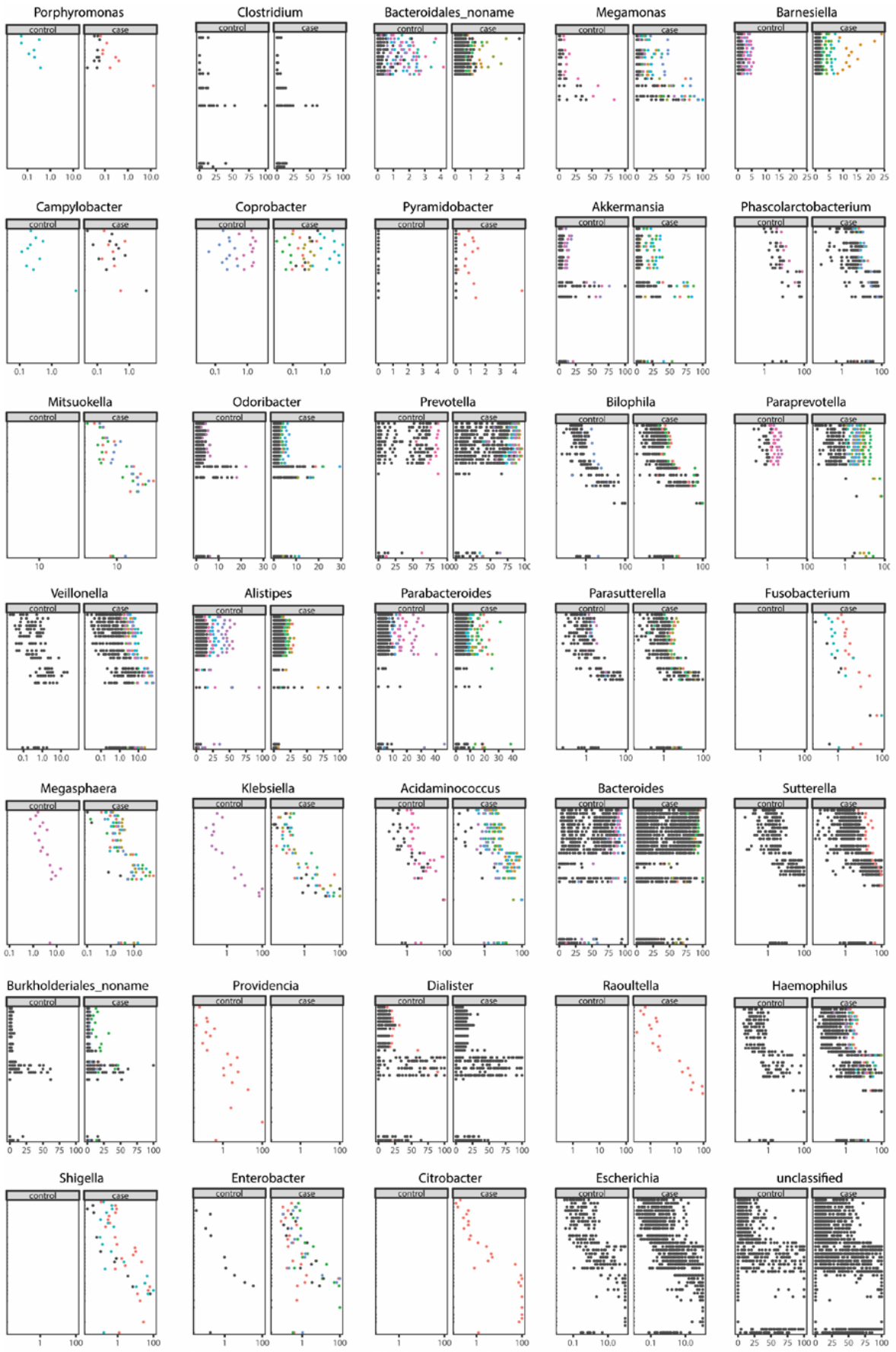
(A) PCO plot showing inter-individual variation in the composition of the gut metagenome. PCO is generated from a Bray-Curtis dissimilarity matrix based on normalized UniRef90 gene cluster abundance. Dotted lines show the outer boundaries for distribution of cases and controls across the first two PC axes. (B) Histogram showing the distribution in the total number of UniRef90 gene clusters detected in cases vs. controls. No significant difference was observed in the number of clusters detected in cases vs. controls (t=-0.30, df=83.2, p=0.76). (C) Violin plot showing distribution in the percentage of mWGS reads that could be successfully mapped to the UniRef90 gene set. No significant difference was observed in the proportion of reads that could be successfully mapped for cases vs. controls (t=-0.83, df=28.7, p=0.41).



Supplementary Figure 11: Metagenomic signature of NASH and fibrosis.

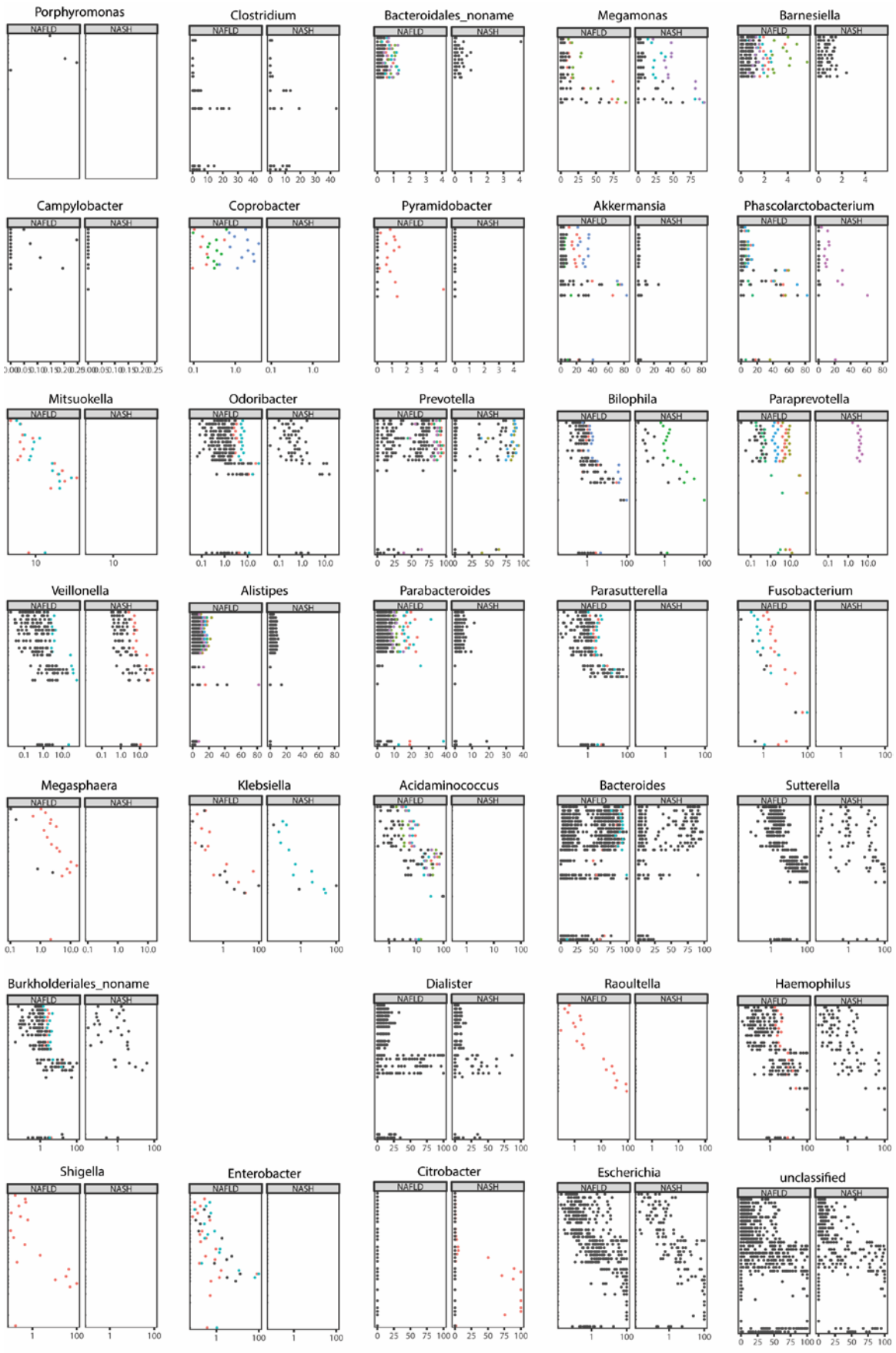
(A) Histogram showing the distribution in the total number of UniRef90 gene clusters detected in patients with NAFLD vs NASH. No significant difference was observed between categories (t=0.32, df=19, p=0.75). (B) Violin plot showing distribution in the percentage

of mWGS reads that could not be successfully mapped to the UniRef90 gene set. No significant difference was observed between categories (t=-0.06, df=17, p=0.96). (C) Histogram showing the distribution in the total number of UniRef90 gene clusters detected in patients with absent-to-mild vs moderate-to-severe fibrosis. No significant difference was observed between categories (t=-0.21, df=27, p=0.84). (D) Violin plot showing distribution in the percentage of mWGS reads that could not be successfully mapped to the UniRef90 gene set. No significant difference was observed between categories (t=-0.83, df=29, p=0.41). (E) Violin plot showing distribution of the percentage of mWGS reads that could not be successfully mapped for individuals in different *P. copri* categories. Significant differences were observed between categories (F=26.86, df=2, p<0.001).


Porphyromonas
Clostridium
Bacteroidales_noname
Megamonas
Barnesiella
Campylobacter
Coprobacter
Pyramidobacter
Akkermansia
Phascolarctobacterium
Mitsuokella
Odoribacter
Prevotella
Bilophila
Paraprevotella
Veillonella
Alistipes
Parabacteroides
Parasutterella
Fusobacterium
Megasphaera
Klebsiella
Acidaminococcus
Bacteroides
Sutterella
Burkholderiales_noname
Providencia
Dialister
Raoultella
Haemophilus
Shigella
Enterobacter
Citrobacter
Escherichia
unclassified
control
case

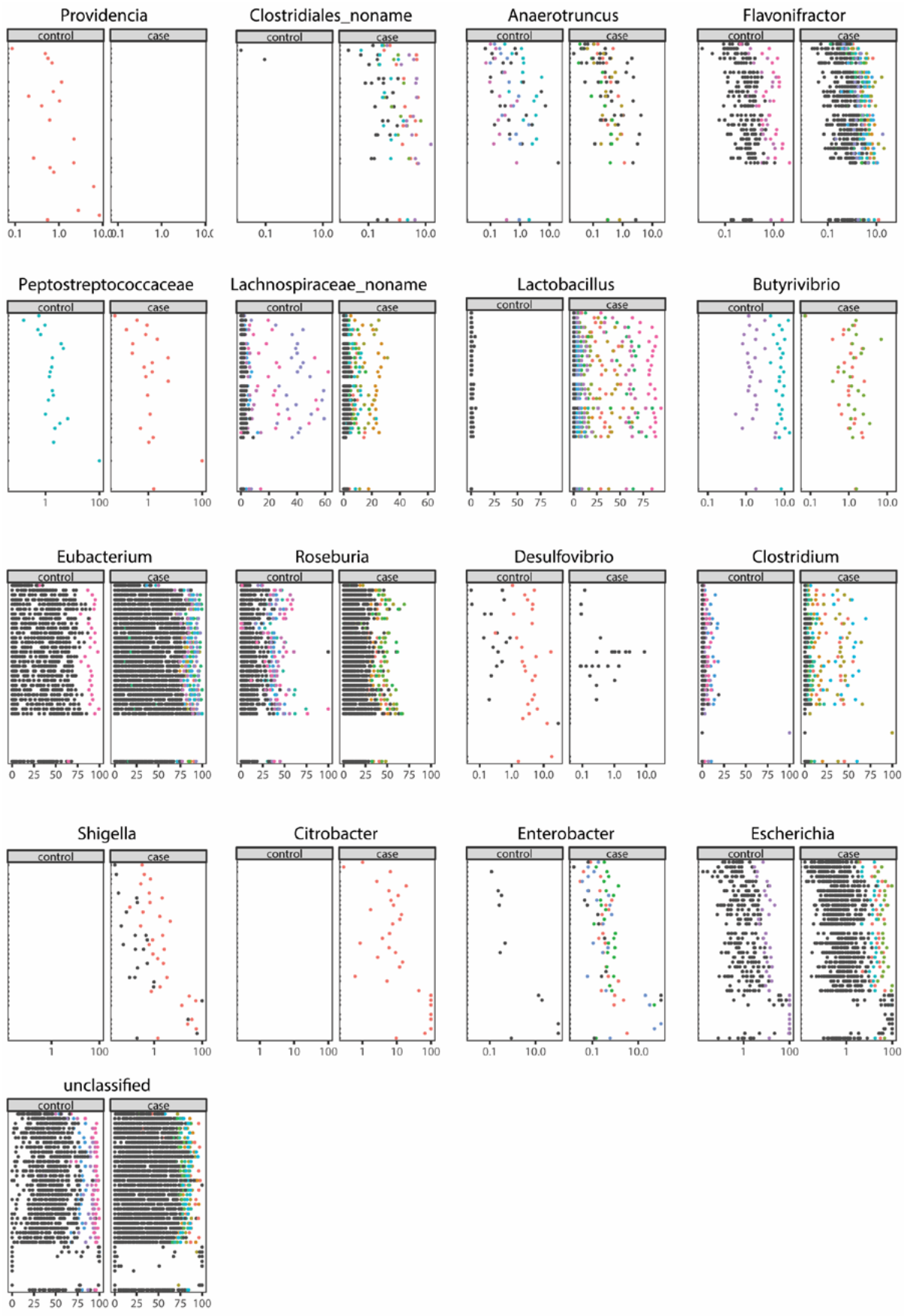
Supplementary Figure 12: Contribution of bacterial genera to lipopolysaccharide biosynthesis gene pathway in individual cases and controls

The proportion of reads mapping to KOs in the lipopolysaccharide biosynthesis pathway that originate from different bacterial genera, shown for individual NAFLD cases and obese controls. X-axis represents the percentage of mWGS reads mapping to a KO that originate from the genus in question. Y-axis represents KOs in the LPS pathway ranked as shown in Fig. 4A. Each data point corresponds to a single patient. Patients for which a bacterial taxon makes a substantial contribution to multiple KOs in the pathway are colored individually (colors are not consistent between panels).


Porphyromonas
Clostridium
Bacteroidales_noname
Megamonas
Barnesiella
Campylobacter
Coprobacter
Pyramidobacter
Akkermansia
Phascolarctobacterium
Mitsuokella
Odoribacter
Prevotella
Bilophila
Paraprevotella
Veillonella
Alistipes
Parabacteroides
Parasutterella
Fusobacterium
Megasphaera
Klebsiella
Acidaminococcus
Bacteroides
Sutterella
Burkholderiales_noname
Dialister
Raoultella
Haemophilus
Shigella
Enterobacter
Citrobacter
Escherichia
unclassified
NAFLD
NASH

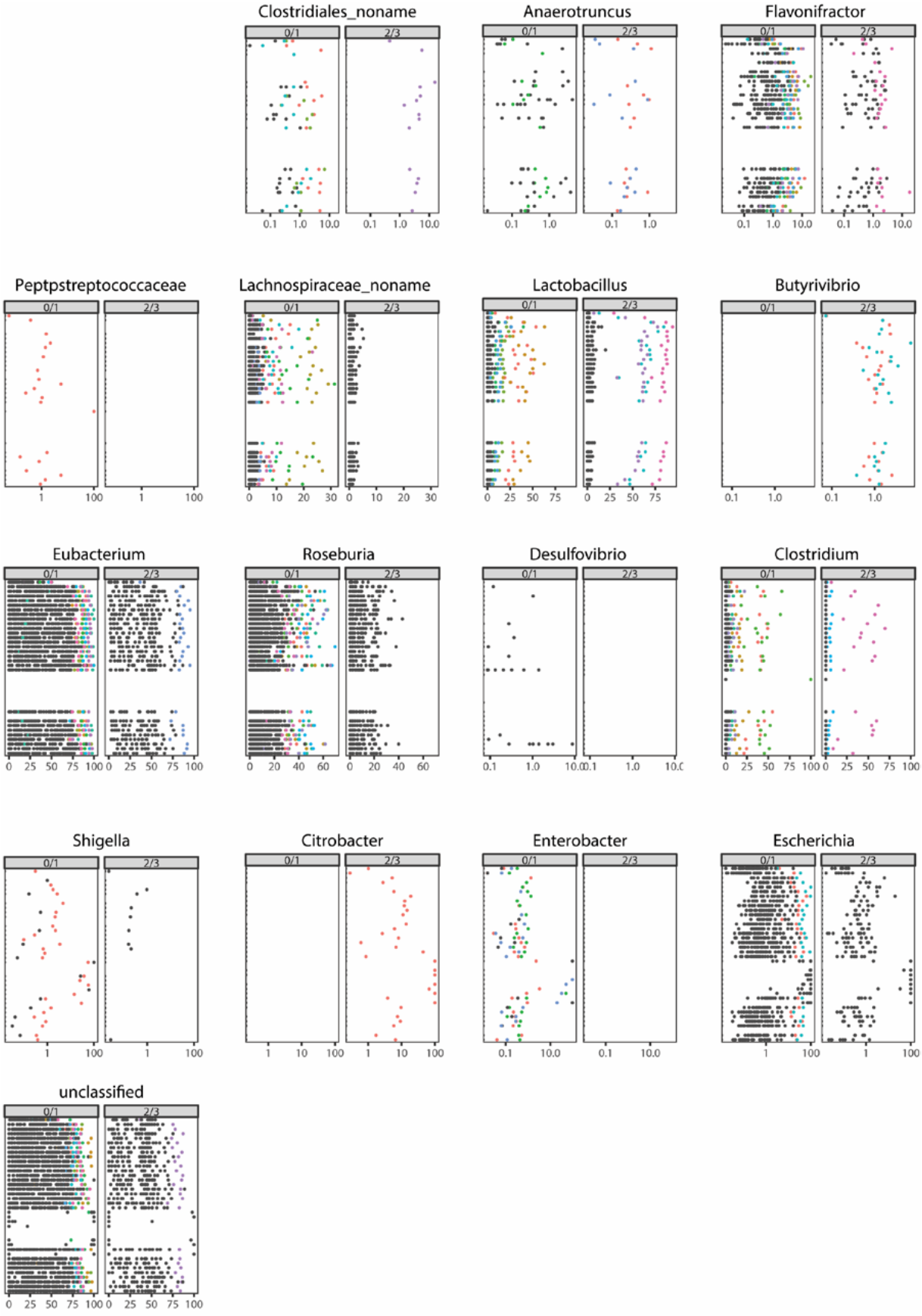
Supplementary Figure 13: Contribution of bacterial genera to lipopolysaccharide biosynthesis gene pathway in individual cases with NAFLD but not NASH and definite NASH

The proportion of reads mapping to KOs in the flagellar assembly pathway that originate from different bacterial genera, shown for individual NAFLD cases and obese controls. X-axis represents the percentage of mWGS reads mapping to a KO that originate from the genus in question. Y-axis represents KOs in the flagellar assembly pathway ranked as shown in Fig. 5A. Each data point corresponds to a single patient. Patients for which a bacterial taxon makes a substantial contribution to multiple KOs in the pathway are colored individually (colors are not consistent between panels).


Providencia
Clostridiales_noname
Anaerotruncus
Flavonifractor
Peptostreptococcaceae
Lachnospiraceae_noname
Lactobacillus
Butyrivibrio
Eubacterium
Roseburia
Desulfovibrio
Clostridium
Shigella
Citrobacter
Enterobacter
Escherichia
unclassified
control
case

Supplementary Figure 14: Contribution of bacterial genera to flagellar assembly gene pathway in individual cases and controls

The proportion of reads mapping to KOs in the flagellar assembly pathway that originate from different bacterial genera, shown for individual NAFLD cases and controls. X-axis represents the percentage of mWGS reads mapping to a KO that originate from the genus in question. Y-axis represents KOs in the flagellar assembly pathway ranked as shown in Fig. 5A. Each data point corresponds to a single patient. Patients for which a bacterial taxon makes a substantial contribution to multiple KOs in the pathway are colored individually (colors are not consistent between panels).


Clostridiales_noname
Anaerotruncus
Flavonifractor
Peptpstreptococcaceae
Lachnospiraceae_noname
Lactobacillus
Butyrivibrio
Eubacterium
Roseburia
Desulfovibrio
Clostridium
Shigella
Citrobacter
Enterobacter
Escherichia
unclassified
0/1
2/3

Supplementary Figure 15: Contribution of bacterial genera to flagellar assembly gene pathway in individual cases with absent-to-mild and moderate-to-severe fibrosis

The proportion of reads mapping to KOs in the flagellar assembly pathway that originate from different bacterial genera, shown for individual cases with either absent-to-mild, or moderate-to-severe fibrosis. X-axis represents the percentage of mWGS reads mapping to a KO that originate from the genus in question. Y-axis represents KOs in the flagellar assembly pathway ranked as shown in Fig. 5E. Each data point corresponds to a single patient. Patients for which a bacterial taxon makes a substantial contribution to multiple KOs in the pathway are colored individually (colors are not consistent between panels).