FECAL MICROBIOME DISTINGUISHES ALCOHOL CONSUMPTION FROM ALCOHOLIC HEPATITIS BUT DOES NOT DISCRIMINATE DISEASE SEVERITY

Ekaterina Smirnova PhD[1], Puneet Puri M.D.[2], Mark D. Muthiah, MBBS, MRCP(UK)[3,4], Kalyani Daitya MSc[2], Robert Brown PhD[5], Naga Chalasani MD[6], Suthat Liangpunsakul MD[6], Vijay H. Shah MD[7], Kayla Gelow PhD[8], Mohammed S. Siddiqui M.D.[2], Sherry Boyett R.N.[2], Faridoddin Mirshahi MSc[2], Masoumeh Sikaroodi PhD[5], Patrick Gillevet PhD[5] and Arun J. Sanyal M.D.[2]

Running title: Alcoholic hepatitis associated changes in fecal microbiota

Primary and corresponding author:
Ekaterina Smirnova PhD
Title: Assistant Professor of Biostatistics
Address: Box 980032
Richmond, VA 23298-0032
Phone: (804) 827 0461
Fax: (804) 828 8900
Email: ekaterina.smirnova@vcuhealth.org

Co-primary author: Puneet Puri, M.D.
Title: Associate Professor of Medicine
Address: MCV Box 980341
Richmond, VA 23298-0341
Phone: (804) 828 6314
Fax: (804) 828 2992
Email: puneet.puri@va.gov

Senior author: Arun J. Sanyal MBBS., M.D.
Title: Professor of Medicine, Physiology and Molecular Pathology
Address: MCV Box 980341
Richmond, VA 23298-0341
Phone: (804) 828 6314
Fax: (804) 828 2992
Email: arun.sanyal@vcuhealth.org

Affiliations:
1. Department of Biostatistics, Virginia Commonwealth University, Richmond, VA
2. Div. of Gastroenterology, Hepatology and Nutrition, Dept. of Internal Medicine, Virginia Commonwealth University, Richmond, VA.
3. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

4. Division of Gastroenterology and Hepatology, National University Hospital, Singapore
5. Dept. of Microbiology, George Mason University, VA
6. Div. of Gastroenterology, Dept. of Internal Medicine, Indiana University, Indianapolis
7. Div. of Gastroenterology, Dept. of Internal Medicine, Mayo Clinic, Rochester, MN
8. Dept. of Biostatistics, Indiana University, Indianapolis, Indiana

MATERIALS AND METHODS

This study was performed by the sites involved in the TREAT consortium from 2014-2018. The TREAT consortium includes three clinical sites (Virginia Commonwealth University, Indiana University and Mayo Clinic), a data coordinating center (Indiana University) and a microbiome analysis core (George Mason University). This consortium is funded by the NIAAA and includes a registry of subjects with alcoholic hepatitis and controls consuming large amounts of alcohol who do not have clinically overt liver disease. Healthy non-drinking controls (HC) with no evidence of liver disease were enrolled solely at VCU. All subjects provided informed consent and the study was approved by the institutional review boards (IRB) at each center. Sample collection and bacterial DNA extraction was performed at VCU whereas the microbiome analysis was performed at George Mason University. The investigators have fully participated in the design, performance and analysis of the study and take full responsibility of the contents of the manuscript. The NIAAA did not participate in the conduct of the studies but provided feedback on the contents of the manuscript.

1. Patient Population:

Alcoholic Hepatitis was defined by the development of jaundice, hepatomegaly and elevated AST with AST:ALT ratio > 1 in an individual with a history of sustained heavy alcohol consumption (> 5 units daily) within 6 weeks of diagnosis in accordance with the

NIAAA consensus definition (16). Those with concomitant alternate etiologies of liver disease such as hepatitis C were excluded. Also, those with active gastrointestinal bleeding, sepsis and those on antibiotics at the time of diagnosis were excluded. Patients receiving lactulose or rifaximin for hepatic encephalopathy were also excluded. The severity of alcoholic hepatitis was defined by a MELD score less than or equal to 20 versus those with higher levels (17).

Individuals with suspected alcoholic hepatitis were initially evaluated and liver enzymes and functions measured along with computation of the MELD score. All patients were assessed clinically for infection and blood cultures obtained along with chest X-ray and urine examination. In those with ascites, a diagnostic paracentesis was performed and the presence of spontaneous bacterial peritonitis excluded. Stool studies for infection were performed as clinically indicated. Participants were considered to have met entry criteria if they met inclusion criteria and had none of the exclusion criteria. Based on the MELD score, patients were categorized to have moderate or severe alcoholic hepatitis (MAH or SAH).

A control population without an alcohol use disorder and obvious liver disease served as a healthy control group (referred to as HC in figures). These individuals were asymptomatic, had a normal physical examination, normal liver enzymes and functions and absence of sonographic evidence of liver disease or a CAP score < 250 db/sec on fibroscan (18). Another set of heavy drinking controls (referred to as HDC in figures) who were consuming more than 5 units of alcohol daily but had no overt evidence of liver disease (normal liver enzymes, normal liver function and absence of jaundice or hepatomegaly) were also included to evaluate the impact of heavy alcohol consumption without clinically evident alcoholic hepatitis.

2. Data collection and clinical evaluation:

Demographic, medical history, and clinical data were collected. Alcohol drinking questionnaires to determine the quantity and pattern of alcohol consumption included 10-question Alcohol Use Disorders Identification Test (AUDIT), the Time Line Follow-back (TLFB), as well as the National Institute on Alcohol Abuse and Alcoholism's (NIAA) six-question survey. Blood samples were collected for complete blood counts, metabolic panel, hepatic panel, and coagulation tests. The tests were performed at the local pathology and chemistry laboratory at each site. The MELD score, Child-Pugh (CP), and Maddrey's DF were also recorded and patients were managed as clinically indicated based on the standard of care at each clinical sector.

3. Stool collection:

A standardized approach to stool collection was established and a standard operating procedure put in place. This was based on prior studies of the stool microbiome (19). In this study, most heavy drinking controls were outpatient and stool samples were collected at the time of a visit to the clinical research center. For patients with alcoholic hepatitis and inpatient heavy drinking controls, stool samples were collected within 72 hours of the patient being admitted to the hospital and enrolled into the study. All clinical personnel involved in stool collection were formally trained and also provided written resources and a video on U-tube as additional resources. Fresh stool was collected in sterile plastic collection tubes that had a Puritan PurFlock Ultra swab (Cat# 25-3306-U) connected to the inside of the lid of the collection tube; one container was empty while the second container had 10 ml of RNA later. Approximately 500mg of feces was transferred to each container and then shaken thoroughly taking care to avoid spillage. The tubes were then placed in ziplock bags packed with ice for transportation to the laboratory.

Stool samples were shipped in dry ice to the clinical center at VCU where they were stored at -70° C until they were to be analyzed. This study is specifically related to samples collected in empty containers from which stool DNA was obtained from samples in batches. The fecal DNA was also stored at -70° C until it was ready to be shipped to the microbiome core at GMU. Such shipments were also made in dry ice.

4. Stool Microbiome Analysis:

We used the 16S rRNA to interrogate and characterize gut microbiome composition. Length Heterogeneity PCR (LH-PCR) fingerprinting was routinely used to rapidly survey our samples and standardize the community amplification. We then interrogated the microbial taxa associated with the gut mucosal microbiome using Multitag Sequencing (MTS) on the samples. This latter technique allows the rapid sequencing of multiple samples at one time yielding thousands of sequence reads per sample (19).

Bacterial Community Fingerprinting. LH-PCR was done as previously published (1). Briefly, total genomic DNA was extracted from tissue using Bio101 kit from MP Biomedicals Inc., Montreal, Quebec as per the manufacturer's instructions. About 10 ng of extracted DNA was amplified by PCR using a fluorescently labeled forward primer 27F (5'-(6FAM) AGAGTTTGATCCTGGCTCA G-3') and unlabeled reverse primer 355R' (5'-GCTGCCTCCCGTAGGAGT-3'). Both primers are universal primers for Bacteria (2). The LH-PCR products were diluted according to their intensity on agarose gel electrophoresis and mixed with ILS-600 size standards (Promega) and HiDi Formamide

(Applied Biosystems, Foster City, CA). The diluted samples were then separated on a ABI 3130xl fluorescent capillary sequencer (Applied Biosystems, Foster City, CA) and processed using the Genemapper™ software package (Applied Biosystems, Foster City, CA). Normalized peak areas were calculated using a custom PERL script by and operational taxonomic units (OTUs) constituting less than 1% of the total community from each sample were eliminated from the analysis to remove the variable low abundance components within the communities (19).

Multitag Sequencing (MTS): We then interrogated the microbial taxa associated with the gut fecal microbiome using multitag sequencing (MTS). This technique allows the rapid sequencing of multiple samples at one time, yielding thousands of sequence reads per sample (19). Specifically, we have generated a set of 96 emulsion PCR fusion primers that contain the Ion Torrrent PGM linkers on the 27F primer (AGAGTTTGATCCTGGCTCA G-3′) and 355R′ (5′-GCTGCCTCCCGTAGGAGT-3′) and different eight-base "barcode" between the A adapter and the 27F primer. Thus each fecal sample was amplified with unique barcoded forward 16S rRNA primers, and then up to 96 samples were pooled and subjected to emulsion PCR and sequenced using a Ion Torrent PGM sequencer (Thermo-Fisher). Data from each pooled sample were "deconvoluted" by sorting the sequences into bins based on the barcodes using custom PERL scripts. Reads were filtered based on quality scores and length. Thus, we were able to normalize each sample by the total number of reads from each barcode. We have noted that ligating tagged primers to PCR amplicons distorts the abundances of the communities, and thus it is critical to incorporate the tags during the original amplification step (19).

5. Fecal metabolite sample processing for short chain fatty acids

Sample storage and transport: Stool samples were shipped in dry ice to the clinical center at VCU where they were stored at -70° C until they were to be analyzed. The samples were then shipped in dry ice to Metabolon, Inc, where they were maintained at -80°C until processed. **Sample Preparation:** Samples were prepared using the automated MicroLab STAR® system from Hamilton Company. Several recovery standards were added prior to the first step in the extraction process for QC purposes. To remove protein, dissociate small molecules bound to protein or trapped in the precipitated protein matrix, and to recover chemically diverse metabolites, proteins were precipitated with methanol under vigorous shaking for 2 min (Glen Mills GenoGrinder 2000) followed by centrifugation. The resulting extract was divided into five fractions: two for analysis by two separate reverse phase (RP)/UPLC-MS/MS methods with positive ion mode electrospray ionization (ESI), one for analysis by RP/UPLC-MS/MS with negative ion mode ESI, one for analysis by HILIC/UPLC-MS/MS with negative ion mode ESI, and one sample was reserved for backup. Samples were placed briefly on a TurboVap® (Zymark) to remove the organic solvent. The sample extracts were stored overnight under nitrogen before preparation for analysis.

QA/QC: Several types of controls were analyzed in concert with the experimental samples: a pooled matrix sample generated by taking a small volume of each experimental sample (or alternatively, use of a pool of well-characterized human plasma) served as a technical replicate throughout the data set; extracted water samples served as process blanks; and a cocktail of QC standards that were carefully chosen not to interfere with the measurement of endogenous compounds were spiked into every analyzed sample, allowed instrument performance monitoring and aided chromatographic alignment. Instrument variability was determined by calculating the median relative standard deviation (RSD) for the standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the pooled matrix samples. **Ultrahigh Performance Liquid Chromatography-Tandem Mass Spectroscopy (UPLC-MS/MS):** All methods utilized a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution. The sample extract was dried then reconstituted in solvents compatible to each of the four methods. Each reconstitution solvent contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot was analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds. In this method, the extract was gradient eluted from a C18 column (Waters UPLC BEH C18-2.1x100 mm, 1.7 µm) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA). Another aliquot was also analyzed using acidic positive ion conditions, however it was chromatographically optimized for more hydrophobic compounds. In this method, the extract was gradient eluted from the same afore mentioned C18 column using methanol, acetonitrile, water, 0.05% PFPA and 0.01% FA and was operated at an overall higher organic content. Another aliquot was analyzed using basic negative ion optimized conditions using a separate dedicated C18 column.

The basic extracts were gradient eluted from the column using methanol and water, however with 6.5mM Ammonium Bicarbonate at pH 8. The fourth aliquot was analyzed via negative ionization following elution from a HILIC column (Waters UPLC BEH Amide 2.1x150 mm, 1.7 µm) using a gradient consisting of water and acetonitrile with 10mM Ammonium Formate, pH 10.8. The MS analysis alternated between MS and data-dependent MSn scans using dynamic exclusion. The scan range varied slighted between methods but covered 70-1000 m/z. Raw data files are archived and extracted as described below.

Bioinformatics for stool metabolites: The informatics system consisted of four major components, the Laboratory Information Management System (LIMS), the data extraction and peak-identification software, data processing tools for QC and compound identification, and a collection of information interpretation and visualization tools for use by data analysts. The hardware and software foundations for these informatics components were the LAN backbone, and a database server running Oracle 10.2.0.1 Enterprise Edition.

LIMS: The purpose of the Metabolon LIMS system was to enable fully auditable laboratory automation through a secure, easy to use, and highly specialized system. The scope of the Metabolon LIMS system encompasses sample accessioning, sample preparation and instrumental analysis and reporting and advanced data analysis. All of the subsequent software systems are grounded in the LIMS data structures. It has been modified to leverage and interface with the in-house information extraction and data visualization systems, as well as third party instrumentation and data analysis software.

Data Extraction and Compound Identification: Raw data was extracted, peak-identified and QC processed using Metabolon’s hardware and software. These systems are built on a web-service platform utilizing Microsoft’s .NET technologies, which run on high-performance application servers and fiber-channel storage arrays in clusters to provide active failover and load-balancing. Compounds were identified by comparison to library entries of purified standards or recurrent unknown entities. Metabolon maintains a library based on authenticated standards that contains the retention time/index (RI), mass to charge ratio (m/z), and chromatographic data (including MS/MS spectral data) on all molecules present in the library. Furthermore, biochemical identifications are based on three criteria: retention index within a narrow RI window of the proposed identification, accurate mass match to the library +/- 10 ppm, and the MS/MS forward and reverse scores between the experimental data and authentic standards. The MS/MS scores are based on a comparison of the ions present in the experimental spectrum to the ions present in the library spectrum. While there may be similarities between these molecules based on one of these factors, the use of all three data points can be utilized to distinguish and differentiate biochemicals. More than 3300 commercially available purified standard compounds have been acquired and registered into LIMS for analysis on all platforms for determination of their analytical characteristics. Additional mass spectral entries have been created for structurally unnamed biochemicals, which have been identified by virtue of their recurrent nature (both chromatographic and mass spectral). These compounds have the potential to be identified by future acquisition of a matching purified standard or by classical structural analysis.

Curation: A variety of curation procedures were carried out to ensure that a high quality data set was made available for statistical analysis and data interpretation. The QC and curation processes were designed to ensure accurate and consistent identification of true chemical entities, and to remove those representing system artifacts, mis-assignments, and background noise. Metabolon data analysts use proprietary visualization and interpretation software to confirm the consistency of peak identification among the various samples. Library matches for each compound were checked for each sample and corrected if necessary.

Metabolite Quantification and Data Normalization: Peaks were quantified using area-under-the-curve.

6. **Bioinformatical Data Analysis:**

RDP11 Bayesian Analysis: We identified the taxa present in each sample using the Bayesian analysis tool in Version 11 of the Ribosomal Database Project (RDP11). The abundances of the bacterial identifications were then normalized using a custom PERL script, and taxa present at >0.1% of the community were tabulated. We chose this cutoff because of our a priori assumption that taxa present in <0.1% of the community vary between individuals and have minimal contribution to the functionality of that community and that 20,000 reads per sample will only reliably identify community components that are >0.1% in abundance.

7. **Statistical Data Analysis:**

Software: The Statistical analysis was performed using R software. Full analysis script is available in the Supporting Information.

Filtering of rare taxa: Taxa that are observed in small number of samples and are likely to be sequencing and bioinformatics artifacts were removed using principled permutation filtering algorithm with default settings implemented in function **PERFect_perm()** in **R** software package **PERFect** (20).

Relative abundance across comparison groups: Significance of differences in phylum level taxa average relative abundances across two groups of samples (e.g. HC vs HDC) were tested using t-test for the difference in the means between two population proportions.

Alpha Diversity Analysis: Shannon alpha diversity was estimated using **estimate_richness()** function implemented in R package **phyloseq()**. Overall significance of differences in alpha diversity across four patient groups (HDC, HC, MAH, SAH) was tested using rank-based non-parametric Kruskal-Wallis test implemented in **R** function **kruskal.test()**. Dunn's test with Benjamini-Hochberg Controlling the false discovery rate: a practical and powerful approach to multiple testing, JRSS 1995 multiple comparisons adjustment implemented in R function **dunn.test()** in package **dunn.test** was used to access the significance between pairs of patient groups.

PCO Analysis of Community Structure: Principle Coordinates Analysis (PCoA) is an Eigen analysis performed on the sample pairwise distance derived from the taxa abundance table. Graphically, it is a rotation of a swarm of data points in multidimensional space so that the largest Eigen value denotes the Eigen vector or first principal component that accounts for the greatest variance. The second principal component is orthogonal to the first and accounts for the next highest amount of variance. The first few PCoA axes represent the greatest amount of variation in the data set. A Brays-Curtis distance metric on taxa relative abundance was used for species beta diversity. From the resulting scatter diagram, it was determined how species clustered with the study groups. The permutational analysis of variance (PERMANOVA) McArdie and Anderson. Fitting Multivariate models to community data: a comment on distance-based redundancy analysis, Ecology, 82(1) implemented in **R** function **adonis()** was used to access significance of beta diversity differences across four patient groups.

LEfSe: The linear discriminant analysis effect size (LEfSe) (3), an algorithm for biomarker discovery that identifies enrichment of abundant taxa or function between two or more groups, was used to compare all taxa at different taxonomic levels simultaneously (i.e., phylum, class, order, family, genus) between treatment groups. The non-parametric Kruskal-Wallis statistical test was used to compute differences among treatment groups and then paired Wilcoxon Rank sum tests among subgroups. This method uses a linear discriminant analysis (LDA) model which utilizes continuous independent variables to predict one dependent variable and provides an effect size for the significantly different taxa or metabolic function based on relative differences between two conditions; taking into account both variability and discriminatory power. Unless stated otherwise, alpha values of 0.05 were used for the Kruskal-Wallis rank sum test, and a threshold of >2.0 was chosen for logarithmic LDA score display. A series of bar graphs were constructed to show the relationship between significantly different metabolic functions or taxa at different phylogenetic levels differentiating clades with a common ancestor.

Predictive modeling: Random forest classification and regression models based on taxa genus level relative abundance data were used to build the prediction models of alcoholic hepatitis and MELD score, respectively. R function **randomForest()** with default settings implemented in package **randomForest** was used to fit the models. Models were accessed using K=3 fold cross validation procedure implemented in function **train()** in **R** package **caret**. Model significance was evaluated using **rf.significance()** function in **R** package **rfUtilities**. Heavy alcoholic consuming patients (i.e. HDC, MAH, and SAH groups) were considered in this analysis. For the prediction model for AH, the binary, AH not present (HDC group) versus AH present (SAH and MAH groups) was used as a response and taxa relative abundances as predictors. Predictive ability of AH classification model was accessed using area under the curve criteria (AUC) calculated via **roc()** function if **pROC** package in **R**. Variable significance was evaluated based on the mean Gini impurity decrease (MGID), where taxa with larger MGID values were more important for classification. For the prediction model of MELD score, MELD score was used as a continuous response and taxa relative abundances as predictors. Variable significance was evaluated based on the node purity increase (NPI), where taxa with larger values of NPI were more important for classification.

Inferred metagenomic analysis: The OTU table was used to generate inferred metagenome data using the online Galaxy interface for Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt, version 1.0.0) with default settings as described (4). Briefly, the abundance values of each OTU were normalized to their respective predicted 16S rRNA copy numbers and then multiplied by the respective predicted gene counts for metagenome prediction. The resulting core output was a list of Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologues and predicted gene count data for each sample. We used in-house scripts to parse the output into KEGG module categories for functional pathways and structural complex hierarchies using the KEGG database (**http://www.genome.jp/kegg/module.html**). The output matrix containing the relative abundance of KEGG orthologous groups (KO) per sample was processed with the online Galaxy interface for LEfSe with a threshold logarithmic LDA score set at 2.0 and ranked.

DETAILED COMPARISON BETWEEN SPECIFIC GENERA WITHIN LACNOSPIRACEAE AND RUMINOCOCCACEAE FAMILIES

Within the *Lachnospiraceae* family, the genera reduced in both SAH and MAH as compared to HDC were *Clostridium cluster XIVb*, *Incertaesedis*, *Roseburia*, *Ruminococcus*, *Anaerostipes*, *Eisenbergiella*, and *Syntrophococcus*, while the genera reduced in SAH but not in MAH included *Blautia*, *Coprococcus*, *Dorea*, *Fusicatenibacter*, *Clostridium cluster XIVa*, *Lachnobacterium*, *Lactonifactor*, and *Robinsoniella*. Within the *Ruminococcaceae* family, the genera reduced in both SAH and MAH as compared to HDC were *Clostridium cluster IV*, *Subdoligranulum*, *Acetanaerobacterium*, and *Anaerotruncus*, while the genera reduced in SAH but not in MAH included *Ruminococcus*, *Anaerofilum*, *Ethanoligenens*, *Flavonifractor*, *Faecalibacterium*, *Gemmiger*, *Hydrogenoanaerobacterium*, *Intestimonas*, *Oscillibacter*, and *Sporobacter*.

REFERENCES

1. Mills DK, Fitzgerald K, Litchfield CD, Gillevet PM. A comparison of DNA profiling techniques for monitoring nutrient impact on microbial community composition during bioremediation of petroleum-contaminated soils. Journal of microbiological methods. 2003; 54(1):57-74.

2. Lane DJ. 16S/23S rRNA sequencing. In ‘Nucleic acid techniques in bacterial systematics’.(Eds E Stackebrandt, M Goodfellow), 1991;115–175.

3. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome biology. 2011; 12(6):R60.

4. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Thurber RL, Knight R, Beiko RG. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature biotechnology. 2013; 31(9):814.

SUPPLEMENTARY FIGURES

S1: Time Between Last Drink and Study Enrollment

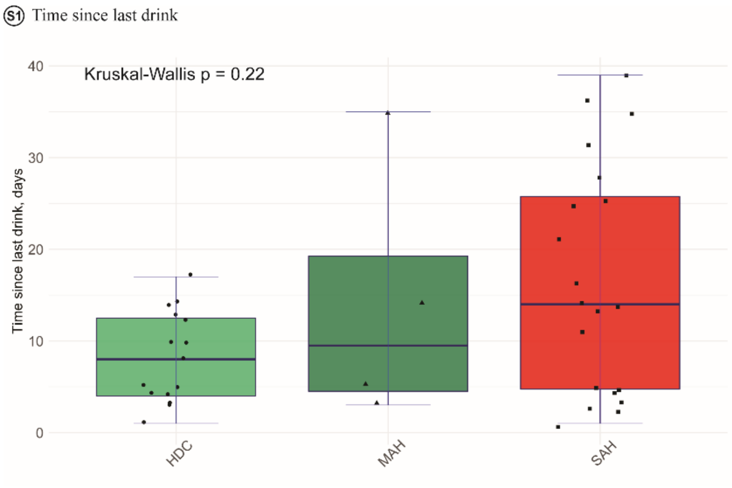


Figure S1: Time since last drink, in days, by study participant groups.

S2: Effects of Acid Reducing and Treatment Medication

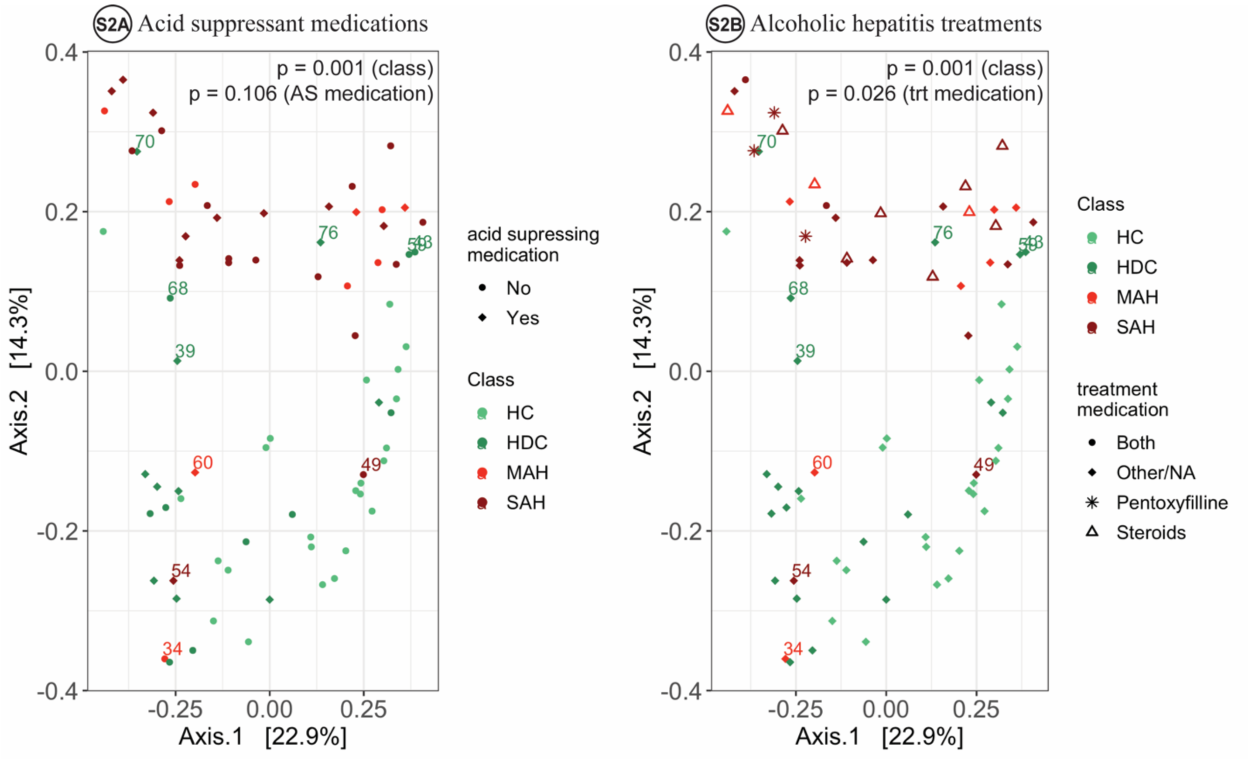


Figure 2. Principal coordinate analysis (PCoA) Bray-Curtis distance plots depicting the relationships between the microbiomes with respect to alcohol use and alcoholic hepatitis study groups. Each point represents a study participant, colored by disease group and shaped by the prescription of acid suppressant (AS) medications (S2A) and treatment (trt) medication (S2B). PERMANOVA test values for microbiota differences between patient and medication groups are displayed as text in top right corner of each subfigure.

FECAL MICROBIOME DISTINGUISHES ALCOHOL CONSUMPTION FROM ALCOHOLIC HEPATITIS BUT DOES NOT DISCRIMINATE DISEASE SEVERITY

Smirnova, E. et al

2020-02-10

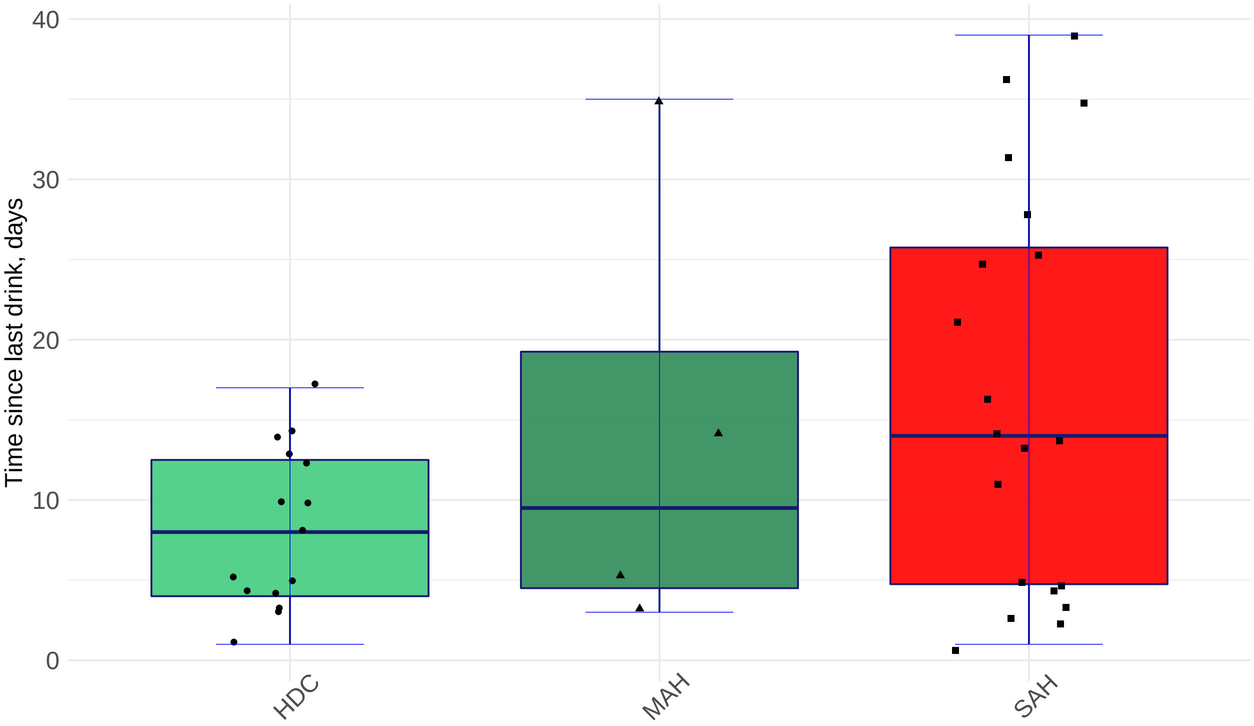
This document is intended to provide the data analysis details, additional model quality accessment reports and R code statements used to produce results and figures reported in the main manuscript. We further provide the estimates and significance tests for group comparisons that were not found significant and were not reported in the manuscript.

R packages used in this analysis

```r
# library(devtools) install_github('katiasmirn/PERFect')
library(pheatmap)
library(readr)
library(genefilter)
library(RColorBrewer)
library(Matrix)
library(ggplot2)
library(reshape2)
library(phyloseq)
require(plyr)
require(dplyr)
library(Hmisc)  # needed for labelling
library(readr)
library(tidyr)
library(here)
library(tableone)
library(pander)
library(ggplot2)
library(reshape2)
library(gridExtra)
library(readxl)
library(Matrix)
library(dunn.test)
library(kableExtra)
library(magrittr)
library(pROC)
library("randomForest")
library("rfUtilities")
library(caret)
library(corrplot)
```

Time since last drink

Supplementary figure 1 looking at the time since last drink and corresponding significance tests



```
$HDC
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    1.0     4.0     8.0     8.2    12.5    17.0       5 

$MAH
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   3.00    4.50    9.50   14.25   19.25   35.00       6 

$SAH
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   1.00    4.75   14.00   16.55   25.75   39.00       4 
```

```
# pairwise test between groups
dunn.res <- dunn.test(df$drink_diff, df$Class, kw = TRUE, method = "bh")
```

```
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 3.0462, df = 2, p-value = 0.22

                           Comparison of x by group                            
                             (Benjamini-Hochberg)                              
Col Mean-|
Row Mean |        HDC        MAH
---------+----------------------
     MAH |  -0.712943
         |     0.3569
         |
     SAH |  -1.741471  -0.353518
         |     0.1224     0.3618

alpha = 0.05
```

```
Reject Ho if p <= alpha/2
```

```
dunn.res.df <- data.frame(comparison = dunn.res$comparisons, difference = round(dunn.res$Z,
    2), p_value = round(dunn.res$P.adjusted, 2))
kable(dunn.res.df) %>% kable_styling("striped", full_width = F)
```

comparison	difference	p_value
HDC - MAH	-0.71	0.36
HDC - SAH	-1.74	0.12
MAH - SAH	-0.35	0.36

Phylum level proportions test

	Phyla average percentage			
	HC	HDC	MAH	SAH
Actinobacteria	2	4	3	6
Bacteroidetes	46	26	39	31
Candidatus.Saccharibacteria	0	0	0	0
Chloroflexi	0	0	0	0
Cyanobacteria.Chloroplast	0	0	0	0
Firmicutes	49	62	53	48
Fusobacteria	2	0	0	1
Proteobacteria	2	7	5	14
Spirochaetes	0	0	0	0
Synergistetes	0	0	0	0
Tenericutes	0	0	0	0
Verrucomicrobia	0	0	0	0

Significance of differences in phylum level taxa average relative abundances across two groups of samples (e.g. HC vs HDC) were tested using t-test for the difference in the means between two population proportions.

Note: code is shown only one group comparison, significance for other groups is evaluated similarly

```
# average across groups
Firmicutes_HC_HDC <- t.test(x = meta.phyla[meta.phyla$Class == "HC", "Firmicutes"],
    y = meta.phyla[meta.phyla$Class == "HDC", "Firmicutes"], paired = FALSE,
    var.equal = FALSE, alternative = "two.sided")$p.value
```

comparison	pval
Firmicutes_HC_HDC	0.08
Firmicutes_HDC_SAH	0.09
Bacteroides_HC_HDC	0.01
Bacteroidetes_SAH_HDC	0.59
Proteobacteria_SAH_HDC	0.20
Proteobacteria_SAH_HC	0.01

Significance of the ratio between proportions in two sample groups phyla were tested using t-test approximation for the difference in the means between two population log(ratios).

```
ratio.test <- function(x.class, y.class, taxa.num, taxa.denom, alternative = c("two.sided",
    "less", "greater"), test = TRUE) {
    pval = NA
```

```r
    x.est <- mean(meta.phyla[meta.phyla$Class == x.class, taxa.num])/mean(meta.phyla[meta.phyla$Class ==
        x.class, taxa.denom])
    y.est <- mean(meta.phyla[meta.phyla$Class == y.class, taxa.num])/mean(meta.phyla[meta.phyla$Class ==
        y.class, taxa.denom])
    if (test == TRUE) {
        x.r <- meta.phyla[meta.phyla$Class == x.class, taxa.num]/meta.phyla[meta.phyla$Class ==
            x.class, taxa.denom]
        y.r <- meta.phyla[meta.phyla$Class == y.class, taxa.num]/meta.phyla[meta.phyla$Class ==
            y.class, taxa.denom]

        pval <- t.test(x = log10(x.r), y = log10(y.r), paired = FALSE, var.equal = FALSE,
            alternative = alternative)$p.value
    }
    return(list(x.est = x.est, y.est = y.est, pval = pval))
}

Firm_to_Bact_HDC_vs_HC <- ratio.test(x.class = "HDC", y.class = "HC", taxa.num = "Firmicutes",
    taxa.denom = "Bacteroidetes", alternative = "greater")

df <- data.frame(test = c("HDC", "HC"), est = c(Firm_to_Bact_HDC_vs_HC$x.est,
    Firm_to_Bact_HDC_vs_HC$y.est), pval = c(Firm_to_Bact_HDC_vs_HC$pval, NA))
```

test	est	pval
HDC	2.39	0.02
HC	1.05	NA

Filtering

Taxa that are observed in small number of samples and are likely to be sequencing and bioinformatics artifacts were removed using principled permutation filtering algorithm with default settings implemented in function PERFect_perm() in R software package PERFect (https://bioconductor.org/packages/devel/bioc/html/PERFect.html).

We apply permutation PERFect method with abundance ordering to the counts data to identify significant taxa. Here we reduce the data form 345 to 150 most abundant taxa.

```r
# basic 5% abundance filtering
abund <- sort(apply(mtx.count, 1, nnzero), decreasing = TRUE)
taxa <- names(abund[(abund/length(abund)) > 0.05])  #62 taxa that appear in 5% of the samples
mtx.count.5p <- mtx.count[match(taxa, rownames(mtx.count)), ]

# apply PERFect
library(PERFect)
NP <- NP_Order(t(mtx.count))
Order_Ind <- match(NP, colnames(t(mtx.count)))
FL <- FiltLoss(t(mtx.count))
dFL <- DiffFiltLoss(t(mtx.count), Order_Ind = Order_Ind, Plot = TRUE, Taxa_Names = NP)

grid.arrange(FL$p_FL, dFL$p_FL, ncol = 2)


# apply permutations PERFect with NP ordering
res_perm <- PERFect_perm(X = t(mtx.count))
```

```
mtx.count <- t(res_perm$filtX)
# res_sim<- PERFect_sim(X=t(mtx.count)) dim(res_sim$filtX)

saveRDS(mtx.count, "mtx.count.RDS")
```

Alpha and Beta diversity

Differences in taxa composition across Healthy Controls (HC; n=24), Alcoholic Controls (HDC; n=20), Mild Alcoholics (MAH; n=10) and Severe Alcoholics (SAH; n=24) groups were compared using Shannon alpha diversity measure. Kruskal-Wallis test was used to test significance of alpha diversity differences between groups.

```
alpha_div <- estimate_richness(physeq, measures = richness_measures)
# all(rownames(meta) == rownames(alpha_div))
alpha_div <- cbind(alpha_div, meta)

count <- aggregate(Shannon ~ Class, alpha_div, sum)
mean <- aggregate(Shannon ~ Class, alpha_div, mean)
median <- aggregate(Shannon ~ Class, alpha_div, median)
sd <- aggregate(Shannon ~ Class, alpha_div, sd)
IQR <- aggregate(Shannon ~ Class, alpha_div, IQR)

div.summary <- data.frame(Class = count$Class, count = count$Shannon, mean = mean$Shannon,
    median = median$Shannon, sd = sd$Shannon, IRQ = IQR$Shannon)
```

Dunn test results to compare alpha diversity across patient groups. Here, we used Benjamini-Hochberg correction for multiple testing implemented in dunn.test() function in R.

```
# overall test
ks.pval <- round(kruskal.test(Shannon ~ Class, data = alpha_div)$p.value, 2)
# pairwise test between groups
dunn.res <- dunn.test(alpha_div$Shannon, alpha_div$Class, kw = TRUE, method = "bh")
```

comparison	difference	p_value
HC - HDC	1.04	0.30
HC - MAH	1.25	0.63
HDC - MAH	0.40	0.41
HC - SAH	0.13	0.45
HDC - SAH	-0.91	0.27
MAH - SAH	-1.15	0.38

Beta diversity was accessed using Bray-Curtis dissimilarity measure based on relative abundance data; PERMANOVA test was performed to access the differences in beta diversity.

```
library(vegan)
metadata <- as(sample_data(physeq.prop), "data.frame")

res <- adonis(distance(physeq.prop, method = "bray") ~ Class, data = metadata)
permanova.pval <- round(res$aov.tab[1, 6], 3)
if (permanova.pval < 0.01) {
    permanova.pval <- "< 0.001"
}
```

```
physeq_ordination <- ordinate(physeq.prop, method = "PCoA", distance = "bray")

beta.div <- physeq.prop %>% plot_ordination(physeq_ordination, color = "Class",
    shape = "Class")
```

Random forest predictive modeling

Random forest classification and regression models based on taxa genus level relative abundance data were used to build the prediction models of alcoholic hepatitis and MELD score, respectively. R function randomForest() with default settings implemented in package randomForest was used to fit the models. Models were accessed using K=3 fold cross validation procedure implemented in function train() in R package caret.

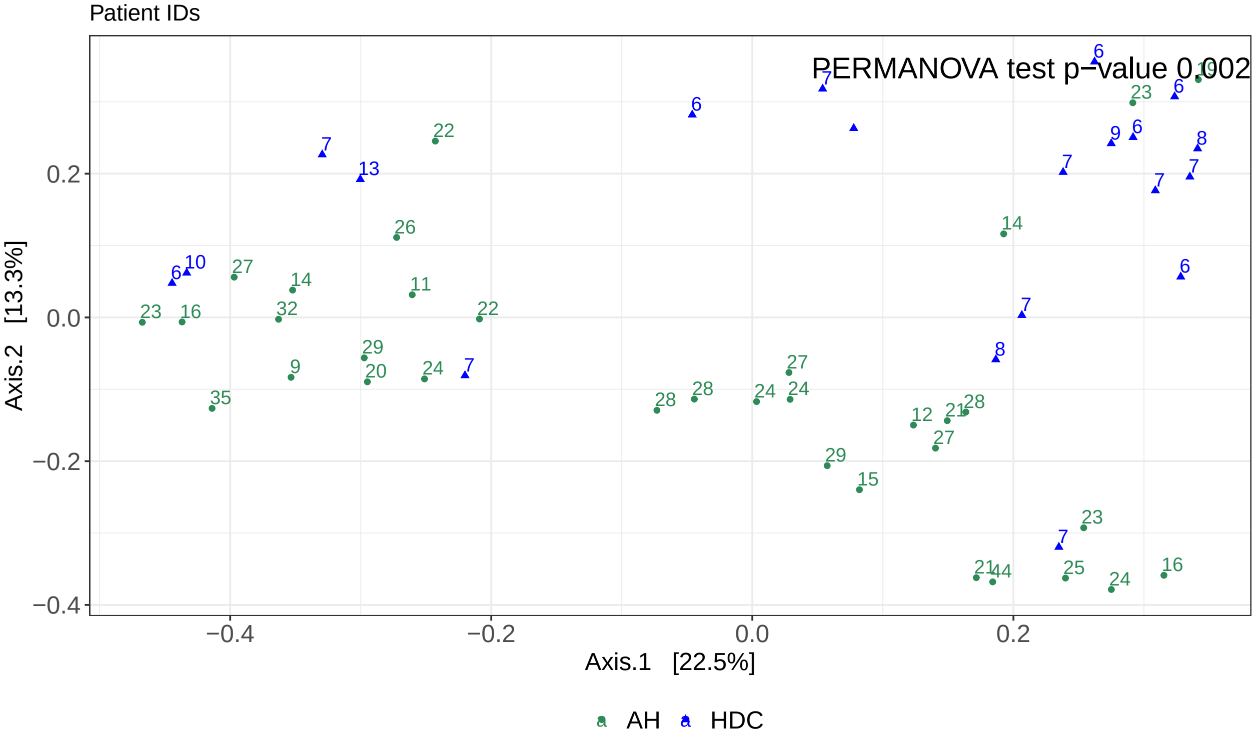
Prediction of alcoholic hepatitis (among all heavy drinking patients) based on the genus level taxa

We start by subsetting the data to all heavy drinking patients, that is groups HDC, MAH and SAH combined. Differences in these groups are visualized using Bray-Curtis PCoA plots and tested using PERMANOVA implemented in R package adonis().

```
physeq.AH <- subset_samples(physeq.prop, Class %in% c("HDC", "MAH", "SAH"))
# trim 0 OTUs
physeq.AH %<>% taxa_sums() %>% is_greater_than(0) %>% prune_taxa(physeq.AH)

metadata.AH <- sample_data(physeq.AH)
metadata.AH$Class_AH <- as.character(metadata.AH$Class)
metadata.AH$Class_AH <- factor(ifelse(metadata.AH$Class_AH == "HDC", "HDC",
    "AH"))

physeq.AH <- phyloseq(otu_table(physeq.AH), metadata.AH, tax_table(physeq.AH))

metadata.AH <- as(sample_data(physeq.AH), "data.frame")

# subset of AH patients
res <- adonis(distance(physeq.AH, method = "bray") ~ Class_AH, data = metadata.AH)
res
```

```

Call:
adonis(formula = distance(physeq.AH, method = "bray") ~ Class_AH,      data = metadata.AH)

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

          Df SumsOfSqs MeanSqs F.Model     R2 Pr(>F)
Class_AH   1    1.0346 1.03459  3.2955 0.0596  0.002 **
Residuals 52   16.3251 0.31394         0.9404
Total     53   17.3597                 1.0000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
permanova.pval.AH <- round(res$aov.tab[1, 6], 3)
```



Random forest categorical prediction model

In categorical prediction model, we create a binary outcome variable with two levels:

1. HDC - heavy drinking controls (i.e. patients without alcoholic hepatitis)
2. AH - patients with alcoholic hepatitis

```
# categorical prediction model data
OTU.pred <- t(otu_table(physeq.AH))
y <- ifelse(metadata.AH$Class.AH.num == 1, "AH", "HDC")
data.AH <- data.frame(y, OTU.pred)
```

Accuracy is the percentage of correctly classified instances out of all instances. It is more useful on a binary classification than multi-class classification problems because it can be less clear exactly how the accuracy breaks down across those classes (e.g. you need to go deeper with a confusion matrix).

Kappa or Cohen's Kappa is like classification accuracy, except that it is normalized at the baseline of random chance on your dataset. It is a more useful measure to use on problems that have an imbalance in the classes (e.g. 70-30 split for classes 0 and 1 and you can achieve 70% accuracy by predicting all instances are for class 0).

```
nfolds = 3
fit_control <- trainControl( method = "cv",
                             number =    nfolds,
                             classProbs = TRUE,
                             savePredictions = TRUE)

RF_classify_cv <- train(y~. , data = data.AH, method="rf", ntree=501 ,
                             tuneGrid=data.frame( mtry=sqrt(dim(OTU.pred)[2]) ), #rule of thumb: sqrt(# variable
                             preProcess=c("center", "scale"),
                             trControl=fit_control)
```

```
RF_classify_cv$results
```

```
      mtry  Accuracy    Kappa AccuracySD   KappaSD
1 12.20656 0.7962963 0.547504  0.1398117 0.3261615
```

Summary of the final cross-validated random forest classification model

```
# final classification model
RF_classify <- RF_classify_cv$finalModel
RF_classify
```

```

Call:
 randomForest(x = x, y = y, ntree = 501, mtry = param$mtry)
               Type of random forest: classification
                     Number of trees: 501
No. of variables tried at each split: 12

        OOB estimate of  error rate: 22.22%
Confusion matrix:
    AH HDC class.error
AH  29   5   0.1470588
HDC  7  13   0.3500000
```

Model significance was evaluated using rf.significance() function in R package rfUtilities. Heavy alcoholic consuming patients (i.e. HDC, MAH, and SAH groups) were considered in this analysis. For the prediction model for AH, the binary, AH not present (HDC group) versus AH present (SAH and MAH groups) was used as a response and taxa relative abundances as predictors.

```
RF_classify_sig <- rf.significance(x = RF_classify, xdata = OTU.pred, nperm = 1000,
    ntree = 501)
# saveRDS(RF_classify_sig, file = paste0(path, '/Code/MELD_classify.rds'))
RF_classify_sig
```

```
Number of permutations:  1000
p-value:  0
Model signifiant at p = 0
     Model OOB error:  0.2222222
     Random OOB error:  0.4259259
     min random global error: 0.2222222
     max random global error:  0.5740741
     min random within class error: 0.55
     max random within class error:  0.55
```

Predictive accuracy

First, we calculate and plot AUC for the final predictive model using roc() function in R package pROC.

```
pred <- predict(RF_classify, type = "prob")[, "AH"]  #predict cases
auc <- roc(ifelse(data.AH$y == "AH", 1, 0), pred)  #set to 1 if AH -- i.e. case
```

Cross-validated AUC values for each fold of prediction are listed below.

```
auc_list <- list()

for (i in 1:nfolds) {
    df <- subset(RF_classify_cv$pred, Resample == paste0("Fold", i))
    # make predictions for each fold, calculate auc, average across folds,
    auc_list[[i]] <- roc(ifelse(df[, "obs"] == "AH", 1, 0), df[, "AH"])
}
cv_auc_vec <- sapply(auc_list, function(x) x$auc[1])

round(cv_auc_vec, 3)
```

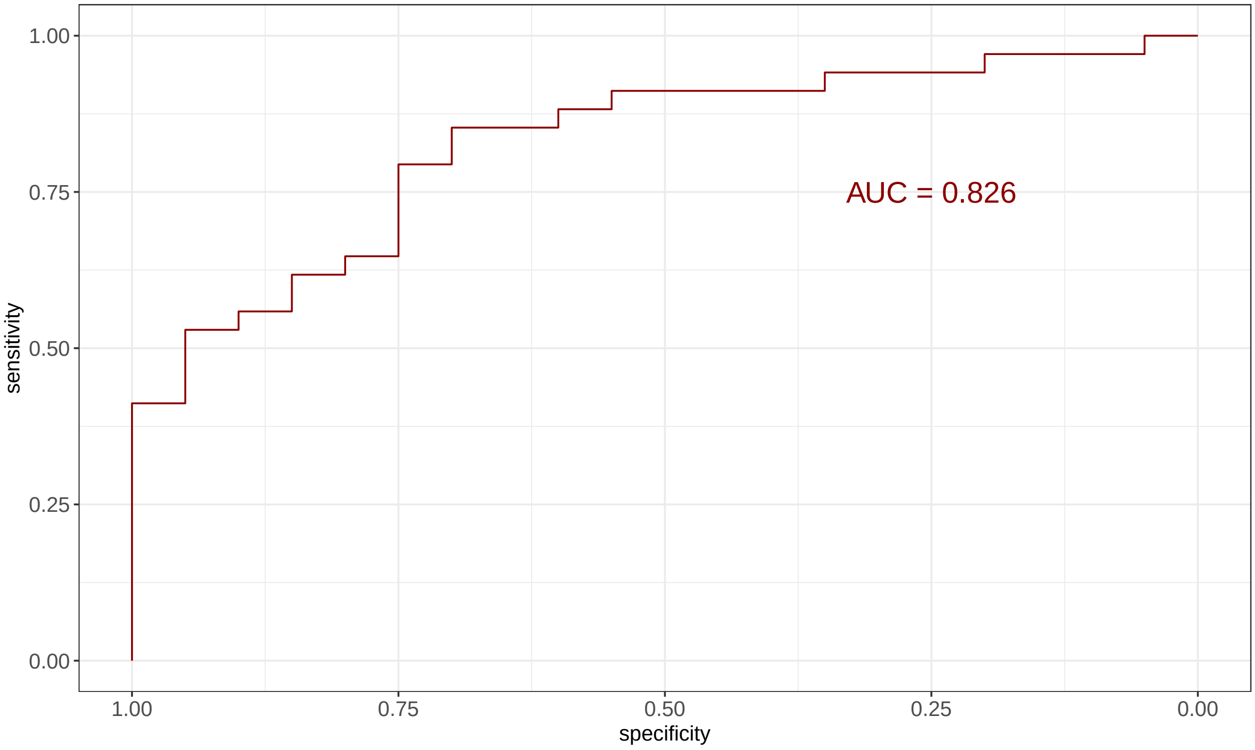
```
[1] 0.792 0.792 0.986
```

Averaged cross validated AUC

```
[1] 0.857
```

Plot of the final alcoholic hepatitis classification model AUC

```
p.roc <- ggroc(auc, color = "darkred") + annotate("text", x = 0.25, y = 0.75,
    label = paste("AUC =", round(auc$auc[1], 3)), color = "darkred", size = 6) +
    theme_bw() + theme(axis.title.x = element_text(size = 12), axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12), axis.title.y = element_text(size = 12))
p.roc
```



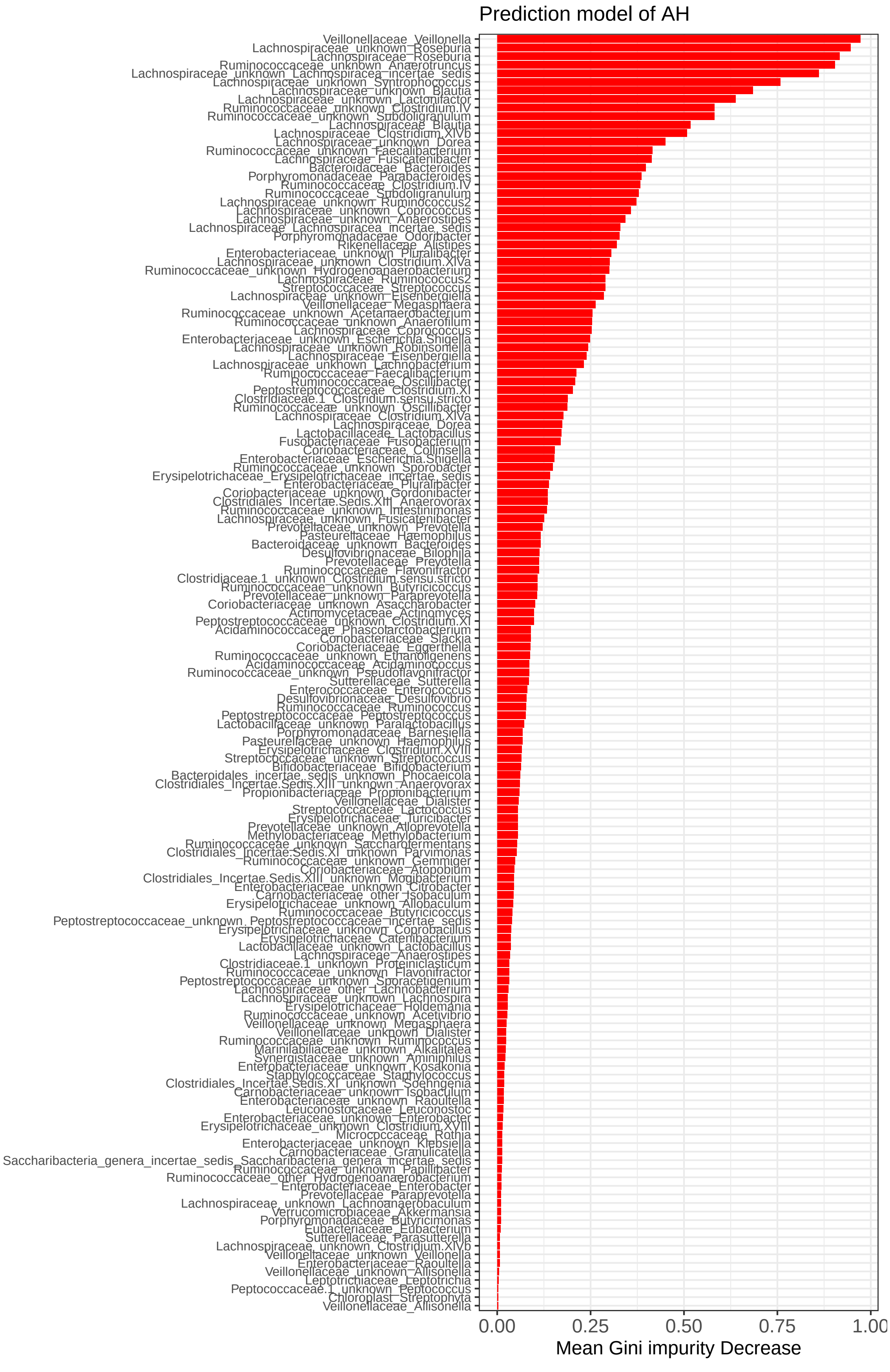
```
ggsave(paste0(path, "/Code/Plots/Figure4_ROC.tiff"), p.roc, width = 6, height = 6)
```

Notice, since sample size is small, we cannot use many folds in cross validation model. Thus, in each fold cross validated AUC is evaluated based on 3 points only, which leads to high variability in AUC. For this reason, we observe the effect that the cross-validated AUC averaged across 3 folds is slightly larger than the final model AUC.

Variable importance

The Mean Decrease Gini measures the average gain of purity by splits of a given variable. If the variable is useful, it tends to split mixed labeled nodes into pure single class nodes. Splitting by a permuted variables tend neither to increase nor decrease node purities. Permuting a useful variable, tend to give relatively large decrease in mean gini-gain. GINI importance is closely related to the local decision function, that random forest uses to select the best available split.

```r
RF_classify_imp <- as.data.frame(importance(RF_classify_cv$finalModel))
RF_classify_imp$features <- rownames(RF_classify_imp)
RF_classify_imp_sorted <- arrange(RF_classify_imp, desc(MeanDecreaseGini))
```


Prediction model of AH
Veillonellaceae_Veillonella
Lachnospiraceae_unknown_Roseburia
Lachnospiraceae_Roseburia
Ruminococcaceae_unknown_Anaerotruncus
Lachnospiraceae_unknown_Lachnospiracea_incertae_sedis
Lachnospiraceae_unknown_Syntrophococcus
Lachnospiraceae_unknown_Blautia
Lachnospiraceae_unknown_Lactonifactor
Ruminococcaceae_unknown_Clostridium.IV
Ruminococcaceae_unknown_Subdoligranulum
Lachnospiraceae_Blautia
Lachnospiraceae_Clostridium.XIVb
Lachnospiraceae_unknown_Dorea
Ruminococcaceae_unknown_Faecalibacterium
Lachnospiraceae_Fusicatenibacter
Bacteroidaceae_Bacteroides
Porphyromonadaceae_Parabacteroides
Ruminococcaceae_Clostridium.IV
Ruminococcaceae_Subdoligranulum
Lachnospiraceae_unknown_Ruminococcus2
Lachnospiraceae_unknown_Coprococcus
Lachnospiraceae_unknown_Anaerostipes
Lachnospiraceae_Lachnospiracea_incertae_sedis
Porphyromonadaceae_Odoribacter
Rikenellaceae_Alistipes
Enterobacteriaceae_unknown_Pluralibacter
Lachnospiraceae_unknown_Clostridium.XlVa
Ruminococcaceae_unknown_Hydrogenoanaerobacterium
Lachnospiraceae_Ruminococcus2
Streptococcaceae_Streptococcus
Lachnospiraceae_unknown_Eisenbergiella
Veillonellaceae_Megasphaera
Ruminococcaceae_unknown_Acetanaerobacterium
Ruminococcaceae_unknown_Anaerofilum
Lachnospiraceae_Coprococcus
Enterobacteriaceae_unknown_Escherichia.Shigella
Lachnospiraceae_unknown_Robinsoniella
Lachnospiraceae_Eisenbergiella
Lachnospiraceae_unknown_Lachnobacterium
Ruminococcaceae_Faecalibacterium
Ruminococcaceae_Oscillibacter
Peptostreptococcaceae_Clostridium.XI
Clostridiaceae.1_Clostridium.sensu.stricto
Ruminococcaceae_unknown_Oscillibacter
Lachnospiraceae_Clostridium.XlVa
Lachnospiraceae_Dorea
Lactobacillaceae_Lactobacillus
Fusobacteriaceae_Fusobacterium
Coriobacteriaceae_Collinsella
Enterobacteriaceae_Escherichia.Shigella
Ruminococcaceae_unknown_Sporobacter
Erysipelotrichaceae_Erysipelotrichaceae_incertae_sedis
Enterobacteriaceae_Pluralibacter
Coriobacteriaceae_unknown_Gordonibacter
Clostridiales_Incertae.Sedis.XIII_Anaerovorax
Ruminococcaceae_unknown_Intestinimonas
Lachnospiraceae_unknown_Fusicatenibacter
Prevotellaceae_unknown_Prevotella
Pasteurellaceae_Haemophilus
Bacteroidaceae_unknown_Bacteroides
Desulfovibrionaceae_Bilophila
Prevotellaceae_Prevotella
Ruminococcaceae_Flavonifractor
Clostridiaceae.1_unknown_Clostridium.sensu.stricto
Ruminococcaceae_unknown_Butyricicoccus
Prevotellaceae_unknown_Paraprevotella
Coriobacteriaceae_unknown_Asaccharobacter
Actinomycetaceae_Actinomyces
Peptostreptococcaceae_unknown_Clostridium.XI
Acidaminococcaceae_Phascolarctobacterium
Coriobacteriaceae_Slackia
Coriobacteriaceae_Eggerthella
Ruminococcaceae_unknown_Ethanoligenens
Acidaminococcaceae_Acidaminococcus
Ruminococcaceae_unknown_Pseudoflavonifractor
Sutterellaceae_Sutterella
Enterococcaceae_Enterococcus
Desulfovibrionaceae_Desulfovibrio
Ruminococcaceae_Ruminococcus
Peptostreptococcaceae_Peptostreptococcus
Lactobacillaceae_unknown_Paralactobacillus
Porphyromonadaceae_Barnesiella
Pasteurellaceae_unknown_Haemophilus
Erysipelotrichaceae_Clostridium.XVIII
Streptococcaceae_unknown_Streptococcus
Bifidobacteriaceae_Bifidobacterium
Bacteroidales_incertae_sedis_unknown_Phocaeicola
Clostridiales_Incertae.Sedis.XIII_unknown_Anaerovorax
Propionibacteriaceae_Propionibacterium
Veillonellaceae_Dialister
Streptococcaceae_Lactococcus
Erysipelotrichaceae_Turicibacter
Prevotellaceae_unknown_Alloprevotella
Methylobacteriaceae_Methylobacterium
Ruminococcaceae_unknown_Saccharofermentans
Clostridiales_Incertae.Sedis.XI_unknown_Parvimonas
Ruminococcaceae_unknown_Gemmiger
Coriobacteriaceae_Atopobium
Clostridiales_Incertae.Sedis.XIII_unknown_Mogibacterium
Enterobacteriaceae_unknown_Citrobacter
Carnobacteriaceae_other_Isobaculum
Erysipelotrichaceae_unknown_Allobaculum
Ruminococcaceae_Butyricicoccus
Peptostreptococcaceae_unknown_Peptostreptococcaceae_incertae_sedis
Erysipelotrichaceae_unknown_Coprobacillus
Erysipelotrichaceae_Catenibacterium
Lactobacillaceae_unknown_Lactobacillus
Lachnospiraceae_Anaerostipes
Clostridiaceae.1_unknown_Proteiniclasticum
Ruminococcaceae_unknown_Flavonifractor
Peptostreptococcaceae_unknown_Sporacetigenium
Lachnospiraceae_other_Lachnobacterium
Lachnospiraceae_unknown_Lachnospira
Erysipelotrichaceae_Holdemania
Ruminococcaceae_unknown_Acetivibrio
Veillonellaceae_unknown_Megasphaera
Veillonellaceae_unknown_Dialister
Ruminococcaceae_unknown_Ruminococcus
Marinilabiliaceae_unknown_Alkalitalea
Synergistaceae_unknown_Aminiphilus
Enterobacteriaceae_unknown_Kosakonia
Staphylococcaceae_Staphylococcus
Clostridiales_Incertae.Sedis.XI_unknown_Soehngenia
Carnobacteriaceae_unknown_Isobaculum
Enterobacteriaceae_unknown_Raoultella
Leuconostocaceae_Leuconostoc
Enterobacteriaceae_unknown_Enterobacter
Erysipelotrichaceae_unknown_Clostridium.XVIII
Micrococcaceae_Rothia
Enterobacteriaceae_unknown_Klebsiella
Carnobacteriaceae_Granulicatella
Saccharibacteria_genera_incertae_sedis_Saccharibacteria_genera_incertae_sedis
Ruminococcaceae_unknown_Papillibacter
Ruminococcaceae_other_Hydrogenoanaerobacterium
Enterobacteriaceae_Enterobacter
Prevotellaceae_Paraprevotella
Lachnospiraceae_unknown_Lachnoanaerobaculum
Verrucomicrobiaceae_Akkermansia
Porphyromonadaceae_Butyricimonas
Eubacteriaceae_Eubacterium
Sutterellaceae_Parasutterella
Lachnospiraceae_unknown_Clostridium.XIVb
Veillonellaceae_unknown_Veillonella
Enterobacteriaceae_Raoultella
Veillonellaceae_unknown_Allisonella
Leptotrichiaceae_Leptotrichia
Peptococcaceae.1_unknown_Peptococcus
Chloroplast_Streptophyta
Veillonellaceae_Allisonella
0.00
0.25
0.50
0.75
1.00
Mean Gini impurity Decrease

Random forest classification results ranked genera according to their predictive importance for the alcoholic hepatitis classification model. However, often taxa genera are observed in low abundance and some taxa might appear in only a few cases and not in controls. Classification models may be sensitive to such rare taxa, therefore it is important to confirm that higly ranked predictive taxa appear in multiple samples and examine their abundance levels.

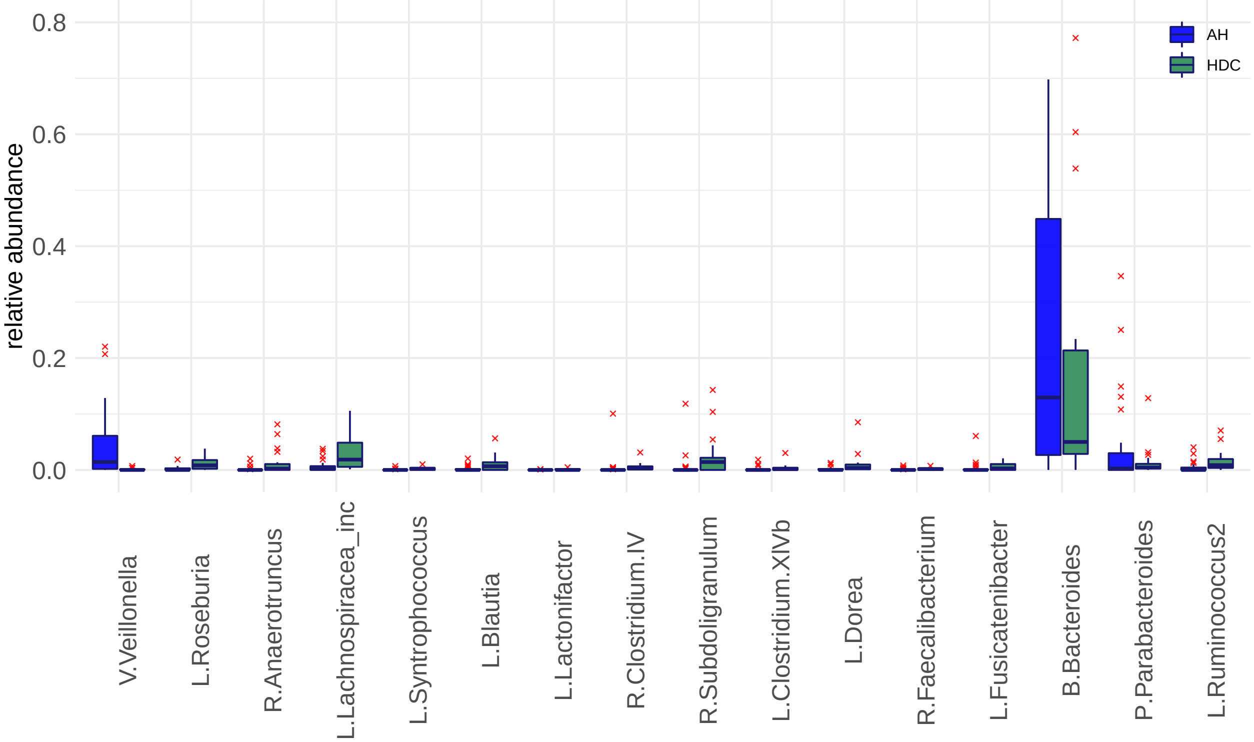
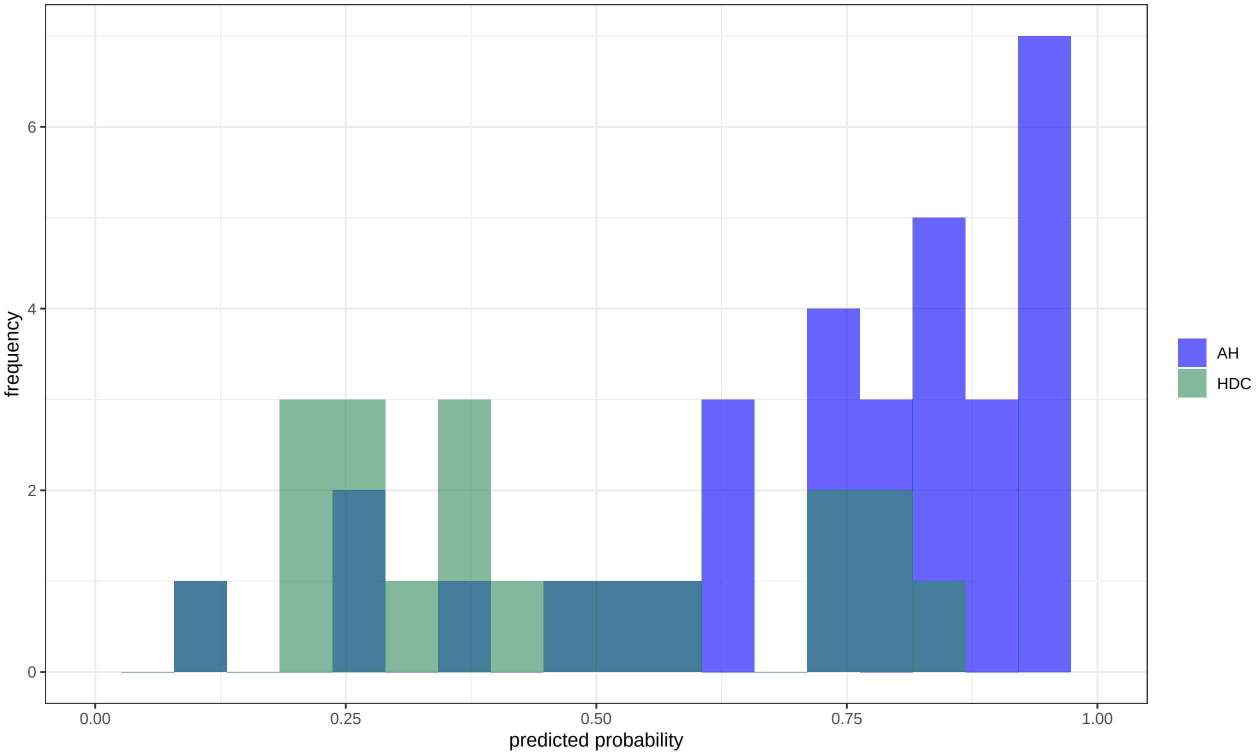


Table below ranks taxa by their relative contribution to the predictive model and lists the number of samples their are present in by HDC and AH groups. Overall, a feature is more reliable for the prediction model if it is observed in multiple samples and the difference in mean abundance is larger across two groups.

	Mean			Number present	
features	Decrease Gini	HDC	AH	HDC	AH
Veillonellaceae Veillonella	0.972	0.001	0.040	6	26
Lachnospiraceae unknown Roseburia	0.946	0.011	0.002	17	16
Lachnospiraceae Roseburia	0.917	0.008	0.003	17	11
Ruminococcaceae unknown Anaerotruncus	0.904	0.013	0.001	14	5
Lachnospiraceae unknown Lachnospiracea incertae sedis	0.861	0.029	0.006	20	23
Lachnospiraceae unknown Syntrophococcus	0.758	0.002	0.000	12	3
Lachnospiraceae unknown Blautia	0.684	0.011	0.002	14	10
Lachnospiraceae unknown Lactonifactor	0.638	0.001	0.000	9	2
Ruminococcaceae unknown Clostridium.IV	0.582	0.005	0.003	15	6
Ruminococcaceae unknown Subdoligranulum	0.582	0.024	0.005	13	6
Lachnospiraceae Blautia	0.517	0.049	0.015	19	21
Lachnospiraceae Clostridium.XlVb	0.508	0.004	0.001	11	3
Lachnospiraceae unknown Dorea	0.450	0.010	0.001	15	11
Ruminococcaceae unknown Faecalibacterium	0.415	0.002	0.001	13	5
Lachnospiraceae Fusicatenibacter	0.413	0.005	0.003	14	9
Bacteroidaceae Bacteroides	0.398	0.205	0.240	19	29
Porphyromonadaceae Parabacteroides	0.386	0.014	0.038	19	20
Ruminococcaceae Clostridium.IV	0.383	0.005	0.001	13	7
Ruminococcaceae Subdoligranulum	0.379	0.019	0.007	13	6
Lachnospiraceae unknown Ruminococcus2	0.373	0.015	0.004	16	15

Finally, we examine the histogram of predicted probabilities for HDC (conrtol) and AH (case) patient groups. The data is colored by the actual patient status. The x-axis represents the range of predicted probabilities (from 0 to 1) for each patient under the final cross validated random forest model and the height of each bar corresponds to the number of patients within the probability range. Overlapping colors indicate the patients with and without alcoholic hepatitis who were predicted within the same probability range.

A good classification model would have a bimodal histogram with lower predicted probability values for HDC and higher predicted probability values for the patients with AH. Visually, the two patient categories separate well, which confirms good classification ability of the model.



Continous random forest MELD score model

We work with the subset of heavy drinking patients (AH, MAH and SAH) groups and model MELD score as a continuous response and microbial genera as predictors.

Summary of the cross validated random forest model accuracy measures.

```
#continuous prediction model data
OTU.pred.regress <- t(otu_table(physeq.AH))[!is.na(metadata.AH$MELD_SCORE),]
MELD <- metadata.AH$MELD_SCORE[!is.na(metadata.AH$MELD_SCORE)]

fit_control <- trainControl( method = "cv",
                    number=5,
                    savePredictions = TRUE)

#use default p/3 splits
RF_regress_cv <- train( OTU.pred.regress , y=MELD, method="rf", ntree=501 ,
              tuneGrid=data.frame( mtry=round(dim(OTU.pred.regress)[2]/3)) ,#rule of thumb: (# variable
              trControl=fit_control )

RF_regress_cv$results
```

```
  mtry     RMSE  Rsquared      MAE   RMSESD RsquaredSD     MAESD
1   50 8.479771 0.2108729 7.054021 1.102481 0.06473199 0.7938433
```

```r
RF_regress <- RF_regress_cv$finalModel
```

Summary of the model significance test.

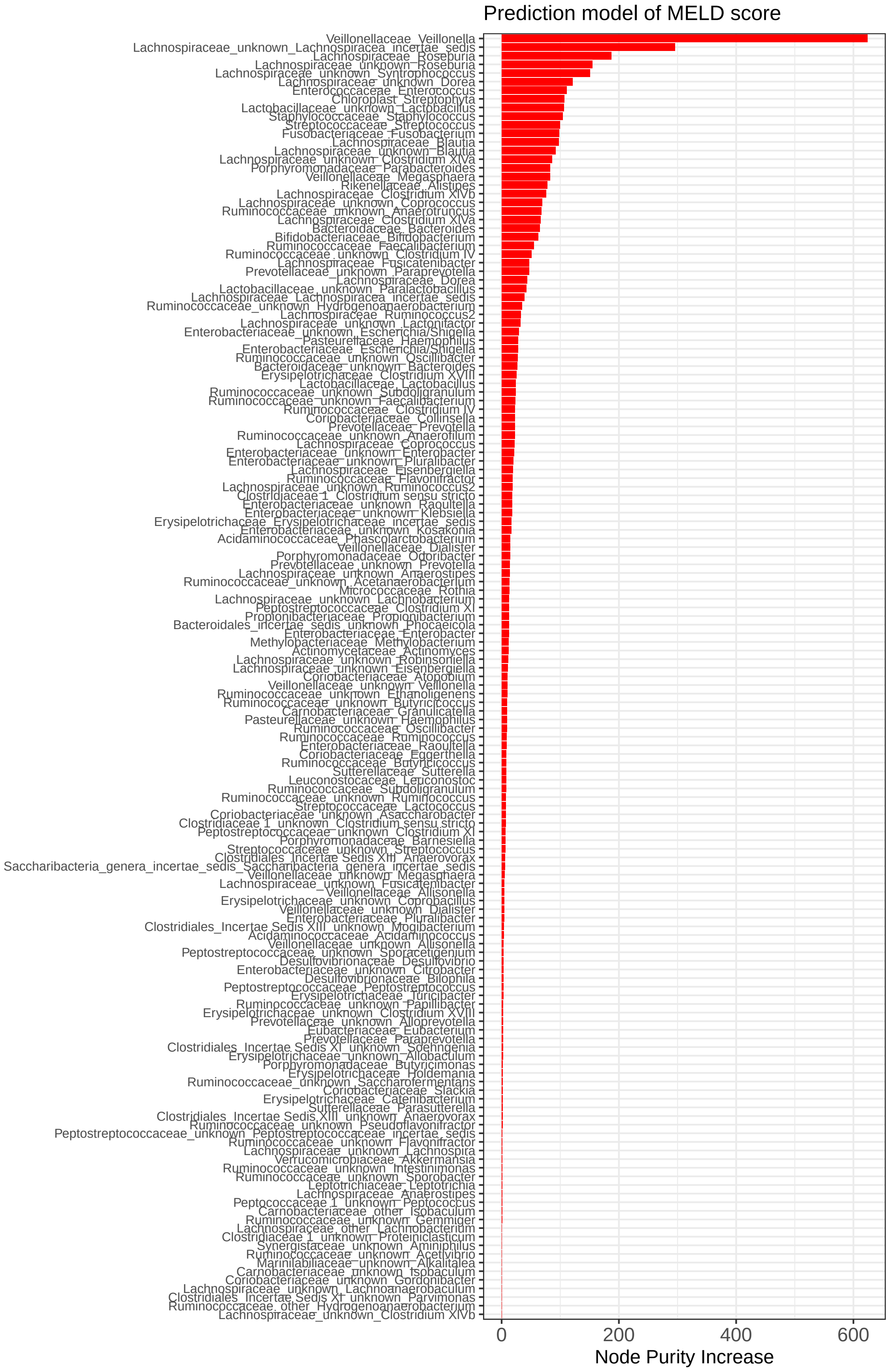
```r
RF_regress_sig <- rf.significance(x = RF_regress, xdata = OTU.pred.regress,
    nperm = 1000, ntree = 501)
RF_regress_sig
```

```
Number of permutations:  1000
p-value:  0.002
Model signifiant at p = 0.002
     Model R-square:  0.1484297
     Random R-square:  -0.1426828
     Random R-square variance:  0.009711888
```

Variable importance measures

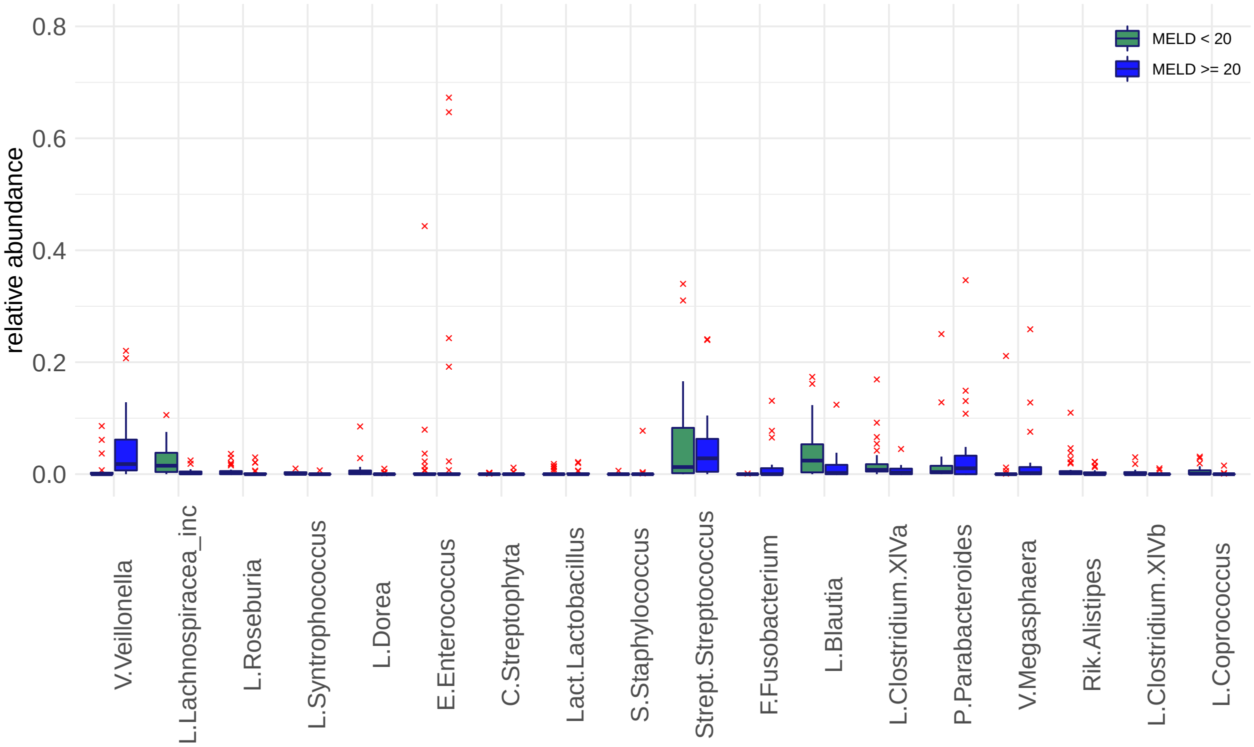
We rank microbial genera according to their contribution to the node purity increase (NPI), where taxa with larger values of NPI are more important for classification.

```r
RF_regress_imp <- as.data.frame(RF_regress$importance)
RF_regress_imp$features <- rownames(RF_regress_imp)
RF_regress_imp_sorted <- arrange(RF_regress_imp, desc(IncNodePurity))
```


Prediction model of MELD score
Veillonellaceae_Veillonella
Lachnospiraceae_unknown_Lachnospiracea_incertae_sedis
Lachnospiraceae_Roseburia
Lachnospiraceae_unknown_Roseburia
Lachnospiraceae_unknown_Syntrophococcus
Lachnospiraceae_unknown_Dorea
Enterococcaceae_Enterococcus
Chloroplast_Streptophyta
Lactobacillaceae_unknown_Lactobacillus
Staphylococcaceae_Staphylococcus
Streptococcaceae_Streptococcus
Fusobacteriaceae_Fusobacterium
Lachnospiraceae_Blautia
Lachnospiraceae_unknown_Blautia
Lachnospiraceae_unknown_Clostridium XlVa
Porphyromonadaceae_Parabacteroides
Veillonellaceae_Megasphaera
Rikenellaceae_Alistipes
Lachnospiraceae_Clostridium XlVb
Lachnospiraceae_unknown_Coprococcus
Ruminococcaceae_unknown_Anaerotruncus
Lachnospiraceae_Clostridium XlVa
Bacteroidaceae_Bacteroides
Bifidobacteriaceae_Bifidobacterium
Ruminococcaceae_Faecalibacterium
Ruminococcaceae_unknown_Clostridium IV
Lachnospiraceae_Fusicatenibacter
Prevotellaceae_unknown_Paraprevotella
Lachnospiraceae_Dorea
Lactobacillaceae_unknown_Paralactobacillus
Lachnospiraceae_Lachnospiracea_incertae_sedis
Ruminococcaceae_unknown_Hydrogenoanaerobacterium
Lachnospiraceae_Ruminococcus2
Lachnospiraceae_unknown_Lactonifactor
Enterobacteriaceae_unknown_Escherichia/Shigella
Pasteurellaceae_Haemophilus
Enterobacteriaceae_Escherichia/Shigella
Ruminococcaceae_unknown_Oscillibacter
Bacteroidaceae_unknown_Bacteroides
Erysipelotrichaceae_Clostridium XVIII
Lactobacillaceae_Lactobacillus
Ruminococcaceae_unknown_Subdoligranulum
Ruminococcaceae_unknown_Faecalibacterium
Ruminococcaceae_Clostridium IV
Coriobacteriaceae_Collinsella
Prevotellaceae_Prevotella
Ruminococcaceae_unknown_Anaerofilum
Lachnospiraceae_Coprococcus
Enterobacteriaceae_unknown_Enterobacter
Enterobacteriaceae_unknown_Pluralibacter
Lachnospiraceae_Eisenbergiella
Ruminococcaceae_Flavonifractor
Lachnospiraceae_unknown_Ruminococcus2
Clostridiaceae 1_Clostridium sensu stricto
Enterobacteriaceae_unknown_Raoultella
Enterobacteriaceae_unknown_Klebsiella
Erysipelotrichaceae_Erysipelotrichaceae_incertae_sedis
Enterobacteriaceae_unknown_Kosakonia
Acidaminococcaceae_Phascolarctobacterium
Veillonellaceae_Dialister
Porphyromonadaceae_Odoribacter
Prevotellaceae_unknown_Prevotella
Lachnospiraceae_unknown_Anaerostipes
Ruminococcaceae_unknown_Acetanaerobacterium
Micrococcaceae_Rothia
Lachnospiraceae_unknown_Lachnobacterium
Peptostreptococcaceae_Clostridium XI
Propionibacteriaceae_Propionibacterium
Bacteroidales_incertae_sedis_unknown_Phocaeicola
Enterobacteriaceae_Enterobacter
Methylobacteriaceae_Methylobacterium
Actinomycetaceae_Actinomyces
Lachnospiraceae_unknown_Robinsoniella
Lachnospiraceae_unknown_Eisenbergiella
Coriobacteriaceae_Atopobium
Veillonellaceae_unknown_Veillonella
Ruminococcaceae_unknown_Ethanoligenens
Ruminococcaceae_unknown_Butyricicoccus
Carnobacteriaceae_Granulicatella
Pasteurellaceae_unknown_Haemophilus
Ruminococcaceae_Oscillibacter
Ruminococcaceae_Ruminococcus
Enterobacteriaceae_Raoultella
Coriobacteriaceae_Eggerthella
Ruminococcaceae_Butyricicoccus
Sutterellaceae_Sutterella
Leuconostocaceae_Leuconostoc
Ruminococcaceae_Subdoligranulum
Ruminococcaceae_unknown_Ruminococcus
Streptococcaceae_Lactococcus
Coriobacteriaceae_unknown_Asaccharobacter
Clostridiaceae 1_unknown_Clostridium sensu stricto
Peptostreptococcaceae_unknown_Clostridium XI
Porphyromonadaceae_Barnesiella
Streptococcaceae_unknown_Streptococcus
Clostridiales_Incertae Sedis XIII_Anaerovorax
Saccharibacteria_genera_incertae_sedis_Saccharibacteria_genera_incertae_sedis
Veillonellaceae_unknown_Megasphaera
Lachnospiraceae_unknown_Fusicatenibacter
Veillonellaceae_Allisonella
Erysipelotrichaceae_unknown_Coprobacillus
Veillonellaceae_unknown_Dialister
Enterobacteriaceae_Pluralibacter
Clostridiales_Incertae Sedis XIII_unknown_Mogibacterium
Acidaminococcaceae_Acidaminococcus
Veillonellaceae_unknown_Allisonella
Peptostreptococcaceae_unknown_Sporacetigenium
Desulfovibrionaceae_Desulfovibrio
Enterobacteriaceae_unknown_Citrobacter
Desulfovibrionaceae_Bilophila
Peptostreptococcaceae_Peptostreptococcus
Erysipelotrichaceae_Turicibacter
Ruminococcaceae_unknown_Papillibacter
Erysipelotrichaceae_unknown_Clostridium XVIII
Prevotellaceae_unknown_Alloprevotella
Eubacteriaceae_Eubacterium
Prevotellaceae_Paraprevotella
Clostridiales_Incertae Sedis XI_unknown_Soehngenia
Erysipelotrichaceae_unknown_Allobaculum
Porphyromonadaceae_Butyricimonas
Erysipelotrichaceae_Holdemania
Ruminococcaceae_unknown_Saccharofermentans
Coriobacteriaceae_Slackia
Erysipelotrichaceae_Catenibacterium
Sutterellaceae_Parasutterella
Clostridiales_Incertae Sedis XIII_unknown_Anaerovorax
Ruminococcaceae_unknown_Pseudoflavonifractor
Peptostreptococcaceae_unknown_Peptostreptococcaceae_incertae_sedis
Ruminococcaceae_unknown_Flavonifractor
Lachnospiraceae_unknown_Lachnospira
Verrucomicrobiaceae_Akkermansia
Ruminococcaceae_unknown_Intestinimonas
Ruminococcaceae_unknown_Sporobacter
Leptotrichiaceae_Leptotrichia
Lachnospiraceae_Anaerostipes
Peptococcaceae 1_unknown_Peptococcus
Carnobacteriaceae_other_Isobaculum
Ruminococcaceae_unknown_Gemmiger
Lachnospiraceae_other_Lachnobacterium
Clostridiaceae 1_unknown_Proteiniclasticum
Synergistaceae_unknown_Aminiphilus
Ruminococcaceae_unknown_Acetivibrio
Marinilabiliaceae_unknown_Alkalitalea
Carnobacteriaceae_unknown_Isobaculum
Coriobacteriaceae_unknown_Gordonibacter
Lachnospiraceae_unknown_Lachnoanaerobaculum
Clostridiales_Incertae Sedis XI_unknown_Parvimonas
Ruminococcaceae_other_Hydrogenoanaerobacterium
Lachnospiraceae_unknown_Clostridium XlVb
0
200
400
600
Node Purity Increase

While prediction model of MELD score models MELD as a continuum, for the purposes of visualizing relative abundance of taxa, we separate patients into two groups that are used in clinical parctice to classify the disease state:

1. MELD < 20
2. MELD >= 20



		Mean		Number present	
features	Inc Node Purity	MELD<20	MELD>20	MELD<20	MELD>20
Veillonellaceae Veillonella	623.666	0.008	0.046	11	21
Lachnospiraceae unknown Lachnospiracea incertae sedis	295.665	0.024	0.004	26	16
Lachnospiraceae Roseburia	187.341	0.006	0.003	19	8
Lachnospiraceae unknown Roseburia	154.475	0.009	0.002	20	12
Lachnospiraceae unknown Syntrophococcus	150.635	0.002	0.000	14	1
Lachnospiraceae unknown Dorea	121.263	0.008	0.001	20	5
Enterococcaceae Enterococcus	110.739	0.022	0.072	7	9
Chloroplast Streptophyta	106.553	0.000	0.001	5	3
Lactobacillaceae unknown Lactobacillus	105.876	0.002	0.002	5	8
Staphylococcaceae Staphylococcus	104.217	0.000	0.003	1	3
Streptococcaceae Streptococcus	99.351	0.058	0.082	24	23
Fusobacteriaceae Fusobacterium	98.168	0.000	0.014	1	12
Lachnospiraceae Blautia	97.549	0.039	0.012	23	16
Lachnospiraceae unknown Blautia	92.142	0.008	0.001	18	5
Lachnospiraceae unknown Clostridium XlVa	85.785	0.022	0.006	27	17
Porphyromonadaceae Parabacteroides	82.724	0.021	0.039	23	15
Veillonellaceae Megasphaera	82.320	0.008	0.022	4	13
Rikenellaceae Alistipes	78.080	0.011	0.004	16	10
Lachnospiraceae Clostridium XlVb	75.868	0.003	0.001	12	2
Lachnospiraceae unknown Coprococcus	69.438	0.005	0.001	14	3

Prediction of disease severity

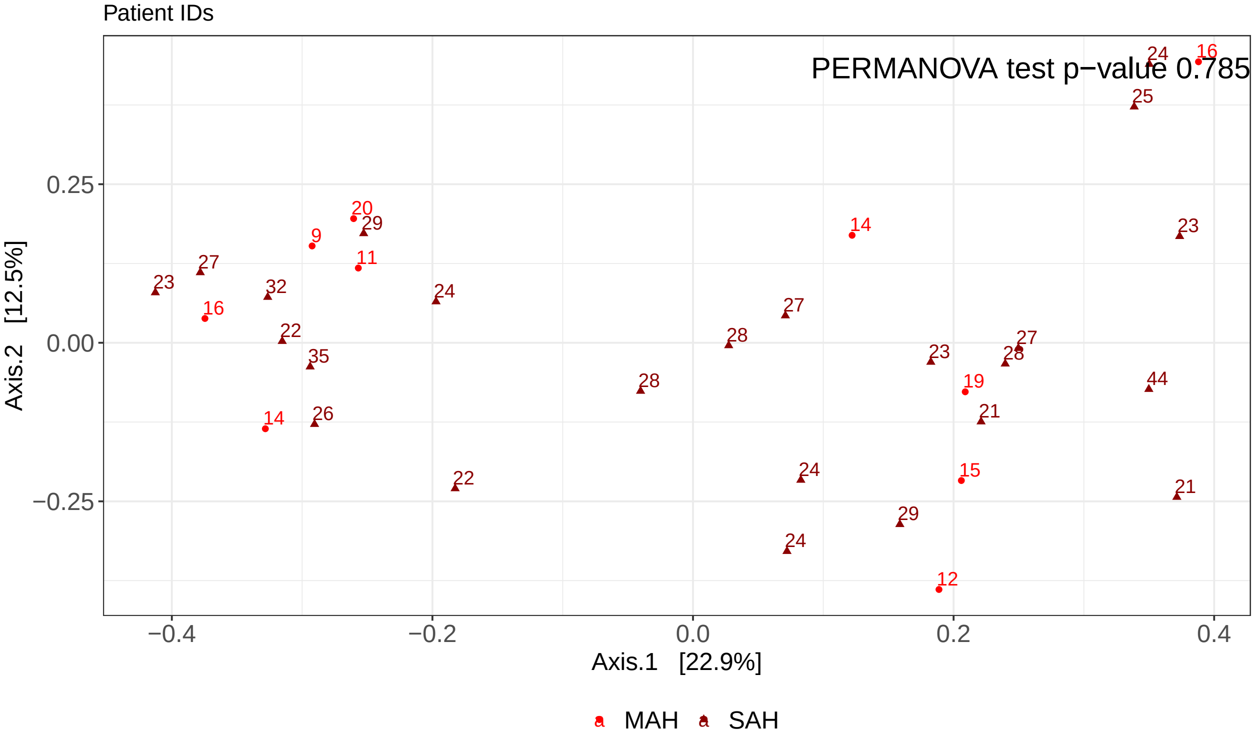
For the purposes of disease severity prediction, we select patients with alcoholic hepatitis, that is MAH and SAH groups. First, we visualize the differences between two groups using PCoA plots and test the significance of taxa composition differences using PERMANOVA.

```r
physeq.AH <- subset_samples(physeq.prop, Class %in% c("MAH", "SAH"))
metadata.AH <- sample_data(physeq.AH)
metadata.AH$Class_AH <- factor(metadata.AH$Class)

physeq.AH <- phyloseq(otu_table(physeq.AH), metadata.AH, tax_table(physeq.AH))

metadata.AH <- as(sample_data(physeq.AH), "data.frame")
```

```r
# subset of AH patients
res <- adonis(distance(physeq.AH, method = "bray") ~ Class, data = metadata.AH)
res
```

```

Call:
adonis(formula = distance(physeq.AH, method = "bray") ~ Class,      data = metadata.AH)

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

          Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
Class      1    0.2261 0.22612 0.68746 0.02103  0.785
Residuals 32   10.5256 0.32893         0.97897
Total     33   10.7518                 1.00000
```

```r
permanova.pval.AH <- round(res$aov.tab[1, 6], 3)
```



Random forest categorical prediction model

We apply cross-validated random forest model to classify disease severity. Model results indicate lack of classification accuracy.

```
      mtry  Accuracy       Kappa AccuracySD    KappaSD
1 12.24745 0.6792929 -0.05128205 0.08310345 0.08882312
```

Summary of the final cross-validated random forest classification model

```

Call:
 randomForest(x = x, y = y, ntree = 501, mtry = param$mtry)
               Type of random forest: classification
                     Number of trees: 501
No. of variables tried at each split: 12

        OOB estimate of  error rate: 32.35%
Confusion matrix:
    MAH SAH class.error
MAH   1   9  0.90000000
SAH   2  22  0.08333333
```

Significance tests of the final model indicate that the model is not significant

```
Number of permutations:  1000
p-value:  0.255
Model not signifiant at p = 0.255
     Model OOB error:  0.3235294
     Random OOB error:  0.3235294
     min random global error: 0.1470588
     max random global error:  0.4411765
     min random within class error: NA
     max random within class error:  NA
```

Random forest continous MELD score modeling

We apply cross-validated random forest model with MELD score as a continuum to the subset of patients with alcoholic hepatitis. Model results indicate lack of fit.

```
  mtry     RMSE   Rsquared     MAE   RMSESD  RsquaredSD    MAESD
1   50 7.063764 0.00326356 5.44227 1.956252 0.001279559 1.471655
```

Summary of the final cross validated model

```

Call:
 randomForest(x = x, y = y, ntree = 501, mtry = param$mtry) 
               Type of random forest: regression
                     Number of trees: 501
No. of variables tried at each split: 50

          Mean of squared residuals: 60.8545
                    % Var explained: -21.85
```

Summary of the significance test results

```
RF_regress_sig <- rf.significance(x = RF_regress, xdata = OTU.pred.regress,
    nperm = 1000, ntree = 501)
RF_regress_sig
```

```
Number of permutations:  1000 
p-value:  0.776 
Model not signifiant at p = 0.776 
     Model R-square:  -0.2564966 
     Random R-square:  -0.1571444 
     Random R-square variance:  0.01566613 
```