Računalna klasifikacija uzoraka mikrobnih zajednica probavnog trakta kod ljudi sa zdravom i bolesnom jetrom

Pavlinek, Eva

University of Zagreb
Faculty of Science
Department of Biology

Eva Pavlinek

Computational classification of gut microbial community from humans with healthy or diseased liver

Graduation thesis

Zagreb, 2020.

This thesis is created in the Bioinformatics group at the Division of Molecular Biology, under the supervision of Professor Kristian Vlahoviček and co-supervision of Assistant Maja Kuzman. The thesis is submitted for grading to the Department of Biology at the Faculty of Science, University of Zagreb, with the aim of obtaining the Master's degree in molecular biology.

I would like to thank my supervisor, Professor Kristian Vlahoviček, for the given opportunities, and my co-supervisor Maja for all the guidance and patience you offered.

Thanks to everyone else from the bioinformatics group for making this experience fun. I genuinely enjoyed every shared coffee, chocolate and day spent working with you.

I am extremely grateful to my friends for fulfilling last few years of my life, especially to those who knew how much this meant to me and never stopped encouraging me. To my family, who had to listen to my countless episodes of whining throughout my studies. Mom, thank you for having nerves of steel.

And lastly, Niko, thank you for all the 1000%.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Department of Biology

Graduation thesis

Computational classification of gut microbial community from humans with healthy or diseased liver

Eva Pavlinek
Division of Molecular Biology, Horvatovac 102A, 10000 Zagreb, Croatia

Metagenomics allows for a research of microorganisms which cannot be cultivated and explores them as entire microbial communities. Among many research directions that are opened through metagenomics, these studies also contribute to the investigation of human microbiome with the aim of discovering causes for different diseases, as well as inventing the methods for their diagnosis and treatment. Prokaryotes do not use all synonymous codons with the same frequency and those genes which are coded with preferred codons are optimised for translation. This effect is termed translational optimisation. By comparing individual genes to highly expressed gene set, such as ribosomal protein genes, we can predict their expressivity – their potential to be expressed. It has been demonstrated previously that this is an effect that can be observed not only at the level of a single species, but also at the level of entire microbial communities. Translational optimisation effect was used in this thesis as a basis for building a classification model based on a random forest algorithm. By training the model on the observed differences in codon usage bias between microbial communities from healthy and liver diseased individuals, we can predict with high accuracy the disease status in unknown samples. Also, an exploratory analysis of an entire dataset was performed to examine genes that are most important for the classification of samples. Pathway analysis was conducted to examine pathways in which the samples vary and to identify potential biomarkers for diagnosis of liver disease. Additional physiological information about the patients was observed in search for a correlation with the condition of samples.

(37 pages, 19 figures, 3 tables, 57 references, original in English)
Thesis deposited in the Central Biological Library
Key words: metagenomics, codon usage, translational optimisation, MELP, random forest, liver cirrhosis
Supervisor: Professor Kristian Vlahoviček, PhD
Assistant Supervisor: Maja Kuzman, MSc
Reviewers: Professor Kristian Vlahoviček, PhD
Assoc. Prof Damjan Franjević, PhD
Asst. Prof. Tomislav Ivanković, PhD
Substitution: Asst. Prof. Rosa Karlić, PhD
Thesis accepted: February 6th, 2020.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno matematički fakultet
Biološki odsjek Diplomski rad

Računalna klasifikacija uzoraka mikrobnih zajednica probavnog trakta kod ljudi sa zdravom i bolesnom jetrom

Eva Pavlinek
Zavod za molekularnu biologiju, Horvatovac 102A, 10000 Zagreb, Hrvatska

Metagenomska istraživanja omogućuju proučavanje cijelih mikrobioloških zajednica, naročito onih koje nije moguće uzgajati u laboratorijskim uvjetima. Značajno su doprinijela proučavanju ljudskog mikrobioma sa ciljem otkrivanja uzoraka, dijagnosticiranju te potencijalnom liječenju mnogih bolesti. Sinonimni kodoni nisu jednako zastupljeni kod prokariota. Geni koji su kodirani najpovoljnijim kodonima su optimizirani za translaciju te se ta pojava naziva translacijskom optimizacijom. Uspoređujući pojedine gene s visoko eksprimiranim setom gena poput gena za ribosomske proteine, moguće je odrediti njihovu ekspresivnost – potencijal gena da bude eksprimiran. Ovaj fenomen je prethodno uočen na razini pojedinih vrsta, ali i cijelih mikrobnih zajednica. U ovom diplomskom radu, translacijska optimizacija je poslužila kao osnova za izradu klasifikacijskog sustava temeljenog na algoritmu nasumičnih šuma (eng. Random Forest). Treniranjem sustava na uočenim razlikama u upotrebi kodona između mikrobnih zajednica zdravih pojedinaca i pojedinaca oboljelih od ciroze jetre, moguće je predvidjeti zdravstveno stanje novih uzoraka s visokom pouzdanošću. Također sam provela računalnu analizu svih metagenomskih uzoraka i usporedila ih s dodatnim fiziološkim podacima koji su bili dostupni o pacijentima. Proučavanjem gena sa značajno različitom translacijskom optimizacijom od očekivane, odredila sam razlike u metaboličkim procesima pojedinih uzoraka, a time i identificirala potencijalne biomarkere za neinvazivnu dijagnozu bolesti jetre temeljenu na mikrobnim zajednicama probavnog trakta.

(37 stranica, 19 slika, 3 tablice, 57 literaturnih navoda, jezik izvornika: engleski)
Rad je pohranjen u Središnjoj biološkoj knjižnici.
Ključne riječi: metagenomika, upotreba kodona, translacijska optimizacija, MELP, nasumične šume, ciroza jetre
Voditelj: Prof. dr. sc. Kristian Vlahoviček
Neposredni voditelj: Maja Kuzman, mag. biol. mol.
Ocjenitelji: Prof. dr. sc. Kristian Vlahoviček
izv. prof. dr. sc. Damjan Franjević
doc. dr. sc. Tomislav Ivanković
Zamjena: doc. dr. sc. Rosa Karlić
Rad prihvaćen: 6. veljače, 2020.

Abbreviations

AUC	Area Under the Curve
BMI	Body Mass Index
CU	Codon Usage
FPR	False Positive Rate
GAGE	Generally Applicable Gene-set Enrichment
KEGG-KO	Kyoto Encyclopaedia of Genes and Genomes - Orthology
MELP	MILC-based Expression Level Predictor
MILC	Measure Independent of Length and Composition
ORF	Open Reading Frame
PCA	Principal Component Analysis
RF	Random Forest
ROC	Receiver Operating Characteristic
TPR	True Positive Rate
UMAP	Uniform Manifold Approximation and Projection

CONTENTS

1. INTRODUCTION

1.1 Metagenomics

Using high-throughput sequencing, metagenomics enables direct analysis of genomes gathered directly from the environmental sample. These studies significantly contribute to research of whole microbial communities in their natural environment, especially making it useful to investigate organisms whose cultivation is impossible in laboratory conditions (Thomas *et al.*, 2012).. Various genomic methods are used with the aim of characterising microorganisms and their metabolic potential, with particular interest to mutual interactions between the microbial species and strains within the community (Tringe and Rubin, 2005).

1.1.1 Metagenomic applications

Analysis of metabolic potential of microbial communities through the investigation of their metabolic pathways could lead to discoveries which could be applied in biotechnology or biomedicine. It enables access to novel biocatalysts from metagenomes. For example, enzymes obtained from the microorganisms which live in extreme conditions could have industrial applications, such as in food or detergent industry (Steele *et al.*, 2009), synthesis of vitamin C (Eschenfeldt *et al.*, 2001) or biotin (Entcheva *et al.*, 2001) and many other. It also expanded the potential of discovering pharmaceutically important molecules. Metagenomic studies enabled discovery of microbial products which can be used as antibiotics (Brady and Clardy, 2004), as well as provided new information about antibiotic resistance mechanisms in microorganisms (Diaz-Torres *et al.*, 2003). It is also applied in ecological studies, where it could be used to investigate the genomic, temporal and spatial variability between the microbial communities and the environment (Delong *et al.*, 2006).

1.1.2 Metagenomics in human health

Metagenomics also has an enormous impact on study of human microbiome – a great number of microorganisms living in symbiosis with human organism. The most important role of human microbiome is in gastrointestinal and immune system. For this reason, more attention is dedicated to research with the aim of studying microbial role in human health (Hooper and Macpherson, 2010). The assumed number of microbes inhabiting human body varied throughout the years of research. The latest conclusions seem to estimate that number to the same order as the number of human cells, or approximately 3,8 x 10^{13} microbial cells for the reference man (a man between 20-30 years of age, weighing 70 kg, and being 170 cm tall) (Sender *et al.*, 2016). The number varies depending mostly on the gender, age and obesity.

Human health depends on human-microorganism mutualism. A primary function of gut microbiota is to enhance the human digestive system, mostly by degrading polysaccharides which humans cannot process on their own (Martens *et al.*, 2008). But it also has an important

role in protecting the host against pathogenic infections (Benson *et al.*, 2009), epithelial cell maturation, angiogenesis (Hooper *et al.*, 2001), lymphocyte development (He *et al.*, 2007; Ivanov *et al.*, 2008) and many other. It is also important to note that the disruption of balance between this mutualism can lead to pathogenicity. Such example is a Gram-positive bacteria *Enterococcus faecalis* which normally has a beneficial function in intestinal digestion, but can exceptionally become pathogenic and cause bacteraemia and endocarditis (Klare *et al.*, 2001). With the help of the metagenomics, great developments were made in diagnosing neurological disorders. Such example is the metagenomic sequencing of cerebrospinal fluid, a method which can identify a broad range of pathogens in a single test. This approach was used to successfully diagnose diseases such as meningitis and encephalitis (Wilson *et al.*, 2019). It was also observed that the gut microbiota is associated to the development and progression of neurological disorders. These interactions can be achieved not only through immune signalling, but also via bacterial metabolites and neural pathways, such as neurotransmitters (Slingerland and Stein-Thoeringer, 2018).

1.2 Translational optimisation in prokaryotes

Due to the degeneracy of the genetic code, not all amino acids are encoded with the equal number of synonymous codons. Prokaryotes do not use all the synonymous codons with the same frequency, a phenomenon termed codon usage bias (Grantham *et al.*, 1980). They preferentially select for certain codons based on the availability of cognate tRNAs (Ikemura, 1981; Yamao *et al.*, 1991) and genes enriched with such codons are said to be optimised for translation, allowing them to be translated more efficiently and accurately. This has an important role in gene expression and protein functional features since it affects processes from RNA processing to protein folding (Plotkin and Kudla, 2011).

It is proposed that all prokaryotes undergo translational selection (Supek *et al.*, 2010), but not to the same extent. Ribosomal protein genes are highly expressed and a representative example of translationally optimised genes. Organisms living in multiple habitats, because of their need for adaption to different environments, seem to exhibit higher degrees of codon usage bias (Botzman and Margalit, 2011).

Although there is a need for analysis of metagenomic samples as an entire community, there are not many computational methods for analysing the metabolism of entire metagenomes. One of the approaches for this challenge is the functional analysis of microbial communities based on the translational optimisation (Roller *et al.*, 2013). Since environmental adaptations in prokaryotes are reflected in codon optimisations, their study can be used for the identification of genes which are significant in these adaptations.

1.3 Codon usage analysis

Based on the analysis of codon usage, it is possible to measure gene's expressivity – its potential to be expressed (Supek and Vlahoviček, 2005). This approach can be used in a computational DNA-based analysis, demanding only metagenomic sequences of prokaryotic open reading frames and avoiding the expensive and laborious methods for gene expression

analysis based on the RNA expression and protein translation analysis. The drawback of such approach is that it lacks the information regarding the actual levels of gene expression.

1.3.1 Measuring codon usage bias

The measure for the codon usage quantification used in this thesis is called Measure Independent of Length and Composition (MILC) (Supek and Vlahoviček, 2005). It calculates the distance in codon usage between a gene and some expected distribution of codons while taking into consideration the bias introduced by different gene lengths, but with proposed minimum sequence length of 80 codons.

The individual contribution M_a od each amino acid a to the MILC statistic is:

$$M_a = 2\sum_c O_c \, ln\frac{O_c}{E_c} = 2\sum_c O_c \, ln\frac{f_c}{g_c} \quad (1)$$

where O_c stands for the observed count of the codon c in a gene and E_c represents the expected count of the same codon. Analogously, f_c is the frequency of the codon c in a gene and g_c is the expected frequency of the same codon. The sum of f or g over all codons for each amino acid should equal 1 and stop codons are excluded.

The complete difference in codon usage is defined as:

$$MILC = \frac{\sum_a M_a}{L} - C \quad (2)$$

The sum of contributions of all amino acids is divided by L, the gene length in codons and the correction factor C is subtracted. The correction factor is used to make up for the codon usage bias which can be overestimated in shorter sequences and is calculated as:

$$C = \frac{\sum_a (r_a - 1)}{L} - 0.5 \quad (3)$$

where r_a denotes the number of possible codons for the amino acid a. If expected and observed codon distributions are similar, MILC can assume negative values. Therefore, a constant of 0.5 is subtracted.

1.3.2 Prediction of gene expressivity

This approach can be used for predicting gene expressivity by using statistic MELP (MILC-based Expression Level Predictor) (Supek and Vlahoviček, 2005), where:

$$MELP = \frac{MILC_{set}}{MILC_{reference}} \quad (4)$$

Statistic MELP represents the ratio of MILC distance of a gene's CU from an average CU of a set of genes and a MILC distance from a reference set of genes, in this case a set of ribosomal protein genes.

These distances for each gene can be plotted in B plots, where each gene is represented by a dot and a characteristic crescent moon shape is seen, as shown in Figure 1. Those genes which have MELP value greater than 1 are considered to have high expressivity.

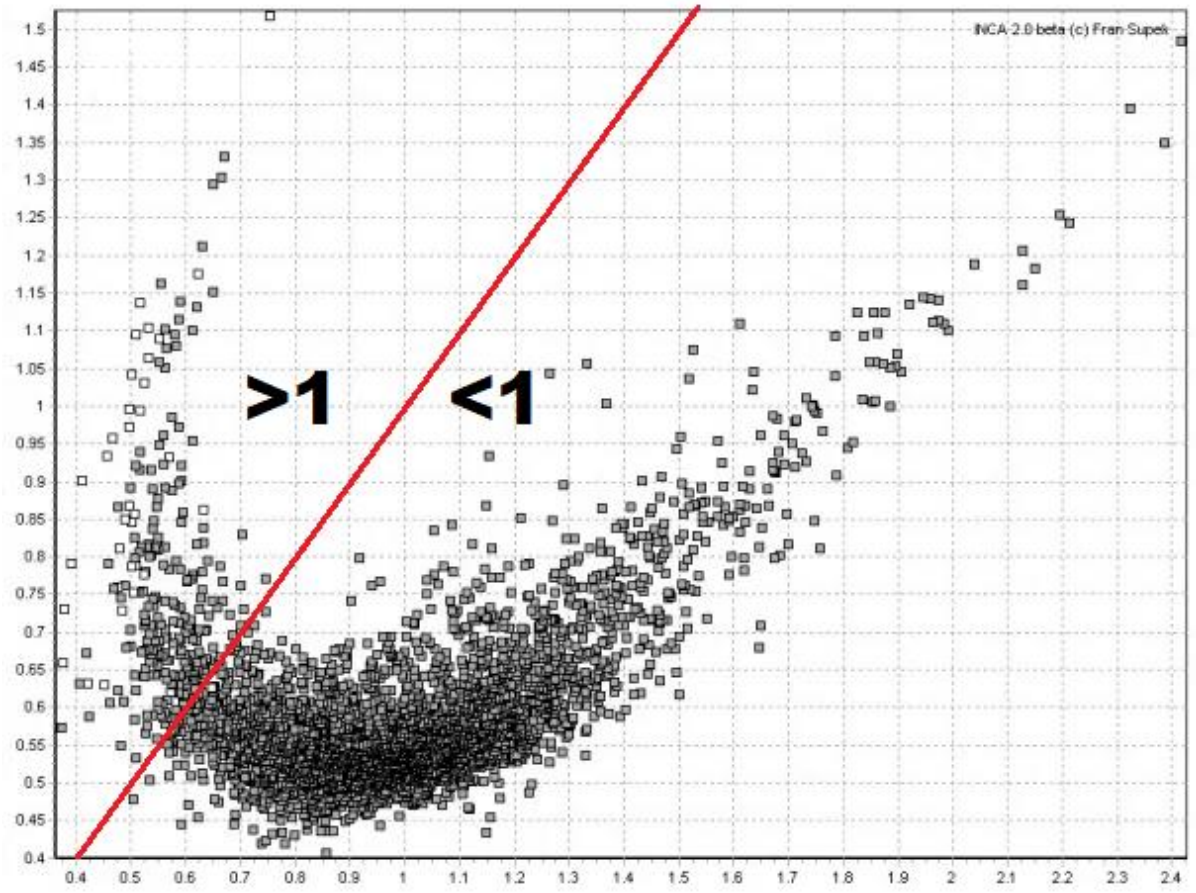

Figure 1: Plots of the *E. coli* genome made using MILC statistics. The distance of codon usage of a gene from *E. coli* ribosomal genes was plotted on the x axis, and the distance of codon usage of a gene from the average codon usage of *E. coli* was plotted on the y axis. The red line represents where the MELP statistics equals to 1, >1 where it is greater than 1 and <1 where it is lower than 1. White squares represent ribosomal protein genes, while all other genes are represented by grey squares. Taken and adjusted from Supek and Vlahoviček, 2005.

1.4 Random forest classifier

There are many classification algorithms in machine learning, but the one used in this thesis is the random forest (RF) classifier (Breiman, 2001). It is based on the decision trees, where the general idea is to build multiple decision trees which are ensembled into a forest, and a definitive classification is made based upon combined trees. Those decision trees vote on how to classify each observation in the given dataset. It examines every feature and in each step searches for one which splits the observations so that the resulting classes are as different from each other as possible. The modification and the advantage of random forests is the use of

bagging. Bagging (Bootstrap Aggregation) is the process in which a random sample with replacement from the dataset is taken, of the same size as the entire dataset, and used for the build of each tree (Breiman, 1996). Another advantage of the random forest classifier is the random sampling of features. In the normal decision tree, every possible feature is considered in each node splitting step, while the random forest can select only a feature from a random subspace of all features (Figure 2). The number of features taken for each tree is usually a square root or a third of the total number of predictors, but the classifier can also be trained in search for the optimal number of predictors.

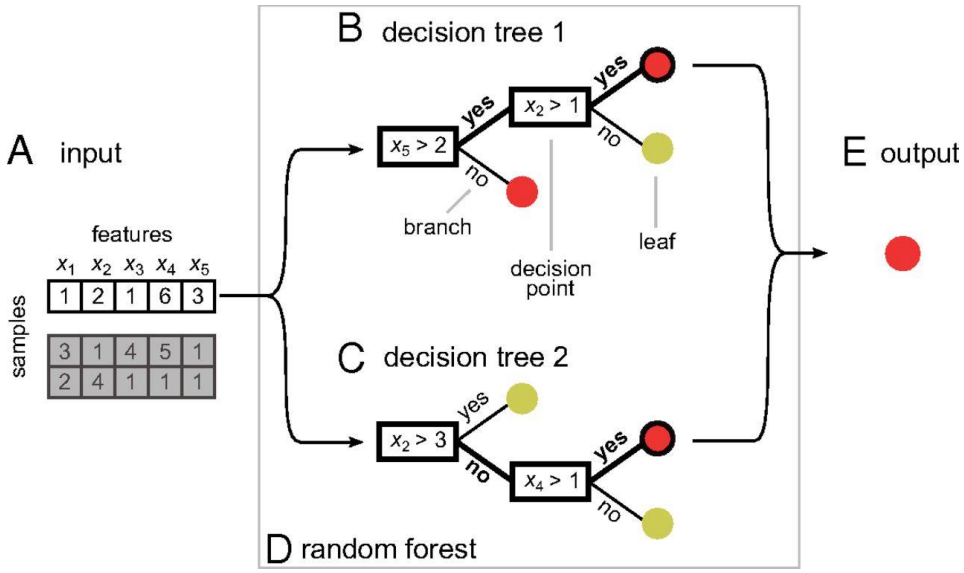

Figure 2: A representation of RF algorithm. Depicted is a dataset with 3 samples and 5 features (A) upon which a random forest is built. Two decision trees (B and C) are shown for classifying the samples, where random sets of only 2 features are used and the classification of the first sample is illustrated. Features are used as decision points where the branches fork and samples are assigned to branches depending on a feature value. The branches terminate in leaves which represent the classes (red or yellow). Multiple decision trees are built for each sample and a random forest (D) combines their votes and concludes the final class prediction (E). This process is repeated for every sample in the dataset. (from Denisko and Hoffman, 2018)

When using classification methods, a given dataset with beforehand known classes is often split into a training and a test set. The model is trained on the training data and then tested for accuracy on the test set. When used for classification, the accuracy is represented as a misclassification error – the percentage of incorrectly classified samples. Given the new dataset, the same model can be used to classify new observations. Therefore, a random forest classifier can be used as a valuable prediction tool. Due to the bagging and random feature sampling, the correlation between the built trees in a random forest is lower which improves predictive accuracy and prevents overfitting to the training data. Another significant benefit of a random forest is that it estimates the variable importance which can tell which features contribute the most to the classification.

1.5. Liver disease

The liver has many vital functions in maintaining the metabolic homeostasis. It processes dietary amino acids, lipids, carbohydrates and vitamins and stores glycogen. It also produces bile, clotting factors and is important for metabolising toxins and cholesterol (Si-Tayeb *et al.*, 2010). Cirrhosis is a progressive scarring condition of the liver which disrupts its structure and functions. It results from acute or chronic liver injuries, mostly caused by alcohol abuse, hepatitis virus infection of obesity (Nishikawa and Osaki, 2015). The scar tissue is formed during its recovery and extensive scarring can lead to life-threatening condition and a necessary liver transplantation.

1.5.1 Diagnosis

Many patients with cirrhotic livers are asymptomatic. The diagnosis includes physical examination, laboratory evaluation and radiologic studies. No serologic test can diagnose cirrhosis accurately, but many tests are performed in search for abnormality in liver functions. These tests usually include a complete blood count with platelets and a prothrombin time test, as well as the analysis of serum enzymes, bilirubin, albumin and creatinine concentrations (Dufour *et al.*, 2000; Nishikawa and Osaki, 2015). Patients with liver cirrhosis have prolongated prothrombin time, decreased serum albumin and bilirubin and creatinine elevations.

There is no radiographic test considered a diagnostic standard, although various studies can indicate the presence of cirrhosis. Ultrasonography is often used as it is the least expensive and does not pose a radiation exposure risk (Šimonovský, 1999).

Lastly, if serologic and radiographic evaluation have failed to confirm a diagnosis of cirrhosis, a liver biopsy is performed to learn its cause and to determine its extensiveness. In conclusion, the diagnosis often includes extensive, long-term, laborious and invasive methods.

1.5.2 Gut microbiota in liver diseases

The liver and gut are strongly connected. The hepatic portal system receives blood from the gut and the liver secrets bile into the intestinal lumen. Cirrhosis is often followed by the bacterial translocation, a migration of bacteria or their products from the gut to the blood circulation or other organs, which can further the progression of liver damage (Fouts *et al.*, 2012). Gut flora alternations also include higher concentrations of toxic acetaldehyde in the lumen, produced through bacterial metabolism of alcohol, and the enhanced production of pro-inflammatory cytokines which propagates a systemic inflammatory state. Some studies demonstrated that probiotics may modify the intestinal microbiota and benefit the treatment of the liver damage (Cesaro *et al.*, 2011).

2. GOALS OF THE RESEARCH

The goal of this research is to study whether metagenomic samples from the healthy individuals and individuals with liver disease can be distinguished based on the differences in translationally optimised sets of genes. Only genes under translational optimisation (i.e. with high expressivity measures based on the MELP statistics) will be used. For this purpose, a classification model based on the random forest algorithm will be built in a language and environment for statistical computing R (R Core Team, 2018). Samples will be split into a training set, for building the classification model, and a test set, for testing the model's accuracy. Different combinations of variables will be used to train the model in search for the optimal set of predictors for the classification.

Next, the exploratory analysis will be applied to the whole dataset to investigate which genes might potentially be relevant for distinguishing healthy and cirrhotic samples. This will be achieved by examining the variable importances in assembled random forest. Also, additional biological data is provided about the patients, which will be analysed in a search for a correlation with obtained classes.

Furthermore, by investigating which genes have significantly higher translational optimisation than the expected and comparing them between the samples, further analysis will be conducted to associate genes with their metabolic pathways. Such approach might identify potential biomarkers for diagnosis of liver disease.

3. MATERIALS AND METHODS

3.1 Initial data

The original data is obtained from the previous research (Qin *et al.*, 2014), where 161 metagenomic samples from the intestinal tract are provided, 80 from the healthy individuals and 81 from the individuals with liver disease. The samples were obtained from the individuals of Han Chinese origin. Additional phenotype information about the patients is presented in Table 1.

Table 1: Phenotype information of individuals

Feature	Description	Range of values
Gender		male or female
Age (years)	-	18 – 78
BMI (kg/m^2)	body mass index	17,58 – 29,03
Cirrhotic	-	yes or no
HBV related	-	yes or no
Alcohol related	-	yes or no
Crea (μmol/L)	the concentration of serum creatine	30,00 – 117,00
Alb (g/L)	the concentration of serum albumin	15,20 – 57,60
TB (μmol/L)	the concentration of total bilirubin	5,00 – 580,00

These sequences were previously assembled and open reading frames (ORFs) were predicted with their corresponding identifiers from the Kyoto Encyclopaedia of Genes and Genomes - Orthology (KEGG - KO) Database (Fabijanić and Vlahoviček, 2016).

3.2 Data pre-processing

3.2.1 Prediction of gene expressivity

Codon usage (CU) frequencies for each ortholog in each sample were calculated by implementing the functions from the coRdon package (Elek *et al.*, 2019) in R. Genes shorter than 80 codons were filtered out in further processing and stop codons were excluded. Next, by computing MELP values for each gene, their expressivity is predicted, with ribosomal genes used as a reference set.

3.2.2 Functional annotation and enrichment analysis

To identify the most significantly enriched or depleted functions in the set of annotated genes, first, the counts of genes annotated to each KO category among all the genes in a sample are calculated, as well as the counts of those which are predicted to have high expressivity (MELP

value greater than 1). The enrichment analysis is performed by scaling and transforming the gene counts by MA transformation, performing the binomial test with correction for multiple testing. The features of pre-processed data and their descriptions are shown in Table 2.

Table 2: The features of processed initial data

feature	description
samples	sample name
category	KO identifier
all	the count of corresponding gene (KO) in the sample
gt_1	the count of corresponding gene with MELP value > 1 in the sample
enrich	a measure of gene enrichment in the sample
M	a scaled counts ratio
A	a mean average of scaled counts
pvals	p-value obtained by binomial test
padj	adjusted p-value for multiple testing

Genes with high expressivity values (MELP values greater than 1) are regarded as optimised for translation and only those were used in further analysis. Samples were annotated by their condition, belonging to either the healthy or the diseased group.

3.3 Random forest classifier

3.3.1 Model training

The data was split into a training test, which contained 80% of all the data, with the equal parts of samples from healthy and diseased individuals, and a test set of the remaining 20% of the data. The features explained in the Table 2 were used as predictors. Using the 5-fold cross-validation from the caret package (Kuhn, 2019), the Random forest classifier from the ranger package (Marvin *et al.*, 2017) was trained on the training set, with the condition as a response variable. The aim of the model training is to obtain the optimal number of predictors for the RF classifier. The model is trained for different number of predictors and the one with the highest accuracy is elected as the best one. The number of trees used for the training was 10000, the method was set to “ranger” and trees were trained for 50 different numbers of predictors. This cross-validation method divides the given set into 5 parts of equal sizes, and successively uses one of them as a test set while other 4 make the training set (Figure 3). The test error is obtained by averaging errors of all 5 validations. This process is repeated for different number of predictors in RF and results with the optimal number of predictors which generated the lowest test error.

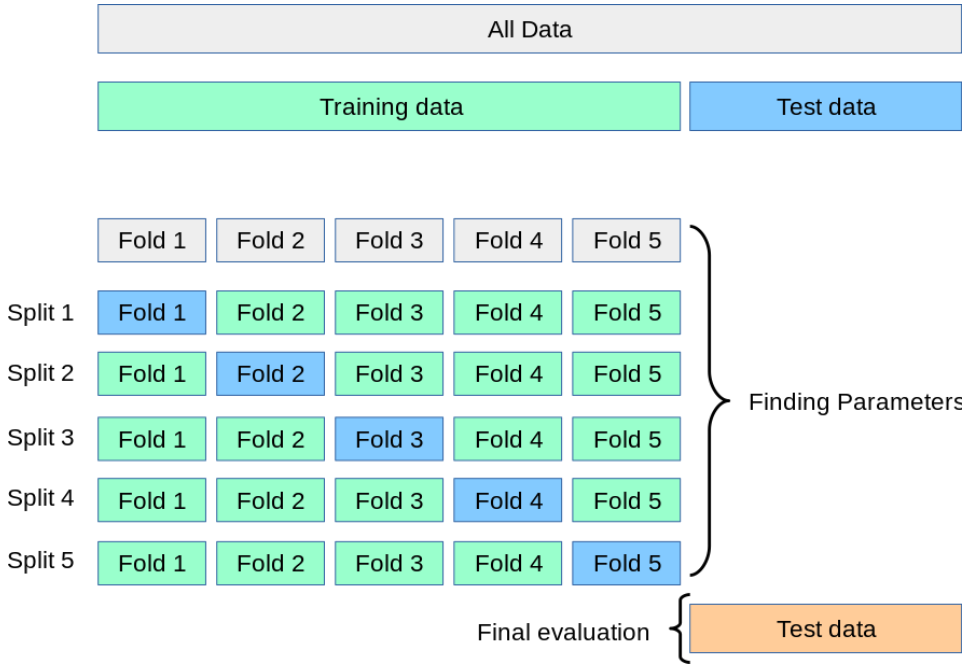

Figure 3: The illustration of 5-fold cross-validation. Training data is split into 5 parts, where in each step, 4 of them are used as a training set and one is used as a test set. This is repeated for each split and the final training error is averaged over all splits. After finding the optimal parameters, trained model can be used to predict the error on a new test data. (Taken from https://scikit-learn.org/stable/modules/cross_validation.html)

3.3.2 Random forest assembly

Based on the optimal number of predictors given by cross-validation, a RF classifier was trained on the entire training set, with the *ntree* parameter set to 10000. The computed RF classifier was applied to predict the condition of the samples from the test set and the test error was computed. The performance of the RF classifier was also evaluated with the Receiver operating characteristic (ROC) analysis – a performance measurement for classification problem. ROC curve is a graphical plot showing the true positive rate (TPR) of the model against the false positive rate (FPR) at various threshold settings. These measures are calculated as:

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

where TP stands for true positive; FN for false negative, FP for false positive and TN for true negative. These measures are also called sensitivity and specificity, respectively, where sensitivity indicates how well can the model identify the true positives, that is the samples from the liver diseased individuals and specificity indicates how well can the model identify the true negatives, that is the samples from the healthy individuals. The area under the curve (AUC) is a measure of model's ability to distinguishing between classes. Its value is between 0 and 1

and higher it is, better the model is at predicting the classes. This was performed using the pROC package (Xavier *et al.*, 2011).

3.3.3 Graphical representation of classification

The RF classifier also calculates the proximity measures among the samples, which is based on the frequency that pairs of the samples are in the same terminal nodes. Those proximity measures were used to create a heatmap with the package pheatmap (Kolde, 2019), a graphical representation of data where proximity measures between all samples are represented as colours. Red colour corresponds to higher proximity values and illustrates samples which are more similar, while the blue colour illustrates more distant samples. The intensity of these colours corresponds to the similarity of samples.

3.4 Feature selection based on the graphical data analysis

3.4.1 Principal Component Analysis (PCA)

Since not every feature shown in Table 2. contributes equally to the separation of samples based on their condition, feature selection was performed through Principal Component Analysis (PCA). PCA is a technique used for dimensionality reduction, exploration and visualisation of the data (James *et al.*, 2013). It transforms the original dataset and produces the principal components - the linear combinations of the original variables which explain most of the variability in the original set and are mutually uncorrelated. The first component has the largest variance, and each subsequent component has lower variance. Considerable differences between the observations in the data can be visualised by plotting the principal components.

3.4.2 Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) is another technique used for dimensionality reduction (McInnes *et al.*, 2018). It is based on the distance between observations rather than the source features and as a result it does not have an equivalent of the linear combinations of the variables. While it lacks the strong interpretability of PCA, it can emphasize the differences between the samples. By visualising the observations on UMAP plots using different predictors, it is possible to select which are the most beneficial for the division of the samples.

By analysing how each feature, and their combinations, contributed to the parting of the data, the most favourable features were chosen. A new RF classifier was trained, built and tested as previously described using only those features.

3.5 Exploratory analysis of entire dataset

Exploratory analysis was conducted to investigate which genes have the most important role in distinguishing the samples from the healthy individuals and individuals with the diseased liver. A new RF model was built using all samples from the dataset. This approach is suitable for examining the given data and its characteristics but cannot be tested as a prediction tool on the same dataset since it uses all samples for the training.

3.5.1 Selection of important predictors based on the empirical p-values

To select only the most important predictors for classification, considering that RF computes slightly different variable importance values each time, p-values for each predictor have been calculated. For this purpose, 1000 RFs have been built on the original dataset, as well as 1000 RFs on permuted data, which is the original dataset with predictor values randomly sampled among each predictor. Some variables are excluded during the computation of each decision tree in RF assembly. The mean decrease in classification accuracy is calculated for each excluded variable. The more the accuracy decreases due to the exclusion of certain variable, the more important it is for the classification of the data. Such variable importance is computed for each RF, for the original and permuted data. P-values were calculated for each predictor by calculating how many of the importance values from the original data are smaller than the greatest importance value from the permuted data, and were adjusted for multiple testing by Bonferroni correction. Only the predictors with the adjusted p-value lower than 0.1 were chosen for the final classification model.

3.5.2 Random forest assembly with important predictors

The complete process of RF assembly was repeated on the entire dataset (training and test set combined). The model was trained and built as previously described, the data separation was analysed through PCA and a heatmap based on the computed proximity measures from the RF model was made.

3.5.3 Wilcoxon-Mann-Whitney rank sum test

Another method used to test which predictors have significantly different values between the samples from the healthy individuals and individuals with the diseased liver is the Wilcoxon-Mann-Whitney rank sum test. This test does not assume a certain distribution of the data, but presumes that observations within each samples as well as the samples amongst themselves are independent of one another. It ranks the measures in groups and tests whether the groups have significantly different means of ranks. In the end, it was analysed how many predictors which resulted with p-value lower than 0.05 by this test intersected with the predictors deemed important by RF.

3.5.4 Codon usage analysis

Differences in codon frequencies between the samples were analysed. Only genes longer than 80 codons which had MELP values greater than 1 were used. Codon usage was examined for all codons, with the stop codons excluded. Median of the number of codons for each gene in each sample was calculated and PCA was applied to this data.

3.6 Metabolic pathway analysis

Samples from the healthy and diseased individuals were compared through GAGE (Generally Applicable Gene-set Enrichment) analysis using the gage package (Luo *et al.*, 2009). When provided with sample names, KO of each gene and the enrichment measure from the initial data, this package analyses significantly different metabolic pathways between the reference and sample set. As a result, it outputs upregulated and downregulated genes, as well as associated metabolic pathways. The analysis was run for all samples and only those genes deemed important by the RF classifier.

3.7 Analysis of phenotype information

Disease status was compared to provided phenotype information to examine whether the samples are grouped together based on any given physiological feature. First, PCA was applied only to the additional data to investigate how gender, age, body mass index (BMI), albumin, bilirubin and creatinine concentrations affect the separation of the samples based on their condition. Then, the samples in PCA plot previously made using the important predictors were labelled by these features, including whether cause for the diseased liver was alcohol or HBV related.

Lastly, Kruskal-Wallis rank sum test was applied to test in which features the samples from the healthy individuals and individuals with diseased liver differ. Like the Wilcoxon-Mann-Whitney rank sum test, this test uses the relative position of the data in a rank ordering and does not assume a certain distribution of the data. Those features where the Kruskal-Wallis test resulted with p-value lower than 0.05 were considered significantly different.

4. RESULTS

4.1 Initial processing of the data

The initial data had 3744 different orthologs. Distribution of their lengths in codons is shown in Figure 4.

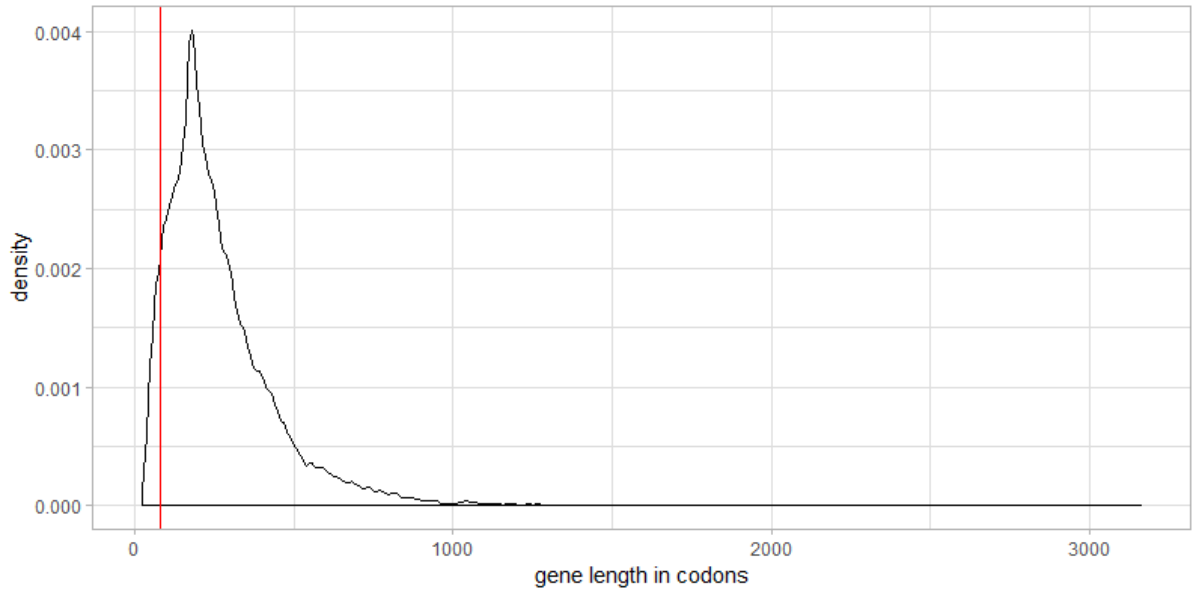

Figure 4: A distribution of gene lengths in codons. Red line represents the length of 80 codons.

After filtering the data and excluding the genes shorter than 80 codons, the new dataset contained 3294 orthologs.

4.2 Building the random forest classifier

4.2.1 Random forest assembly with all predictors

After splitting the data, there were 129 samples in the training and 32 samples in the test set. The optimal RF model obtained by 5-fold cross-validation was the one which used 18 variables. The RF classifier trained on the training set resulted with a training error of 31,78%. A heatmap based on the obtained proximity measures from the training set is shown in Figure 5. When the built RF model was used to classify the samples from the test set, the misclassification error was 25,00%. From the 8 misclassified samples, 3 were from the healthy individuals and 5 from the individuals with diseased liver. The ROC curve for this model is shown in Figure 6 and the corresponding AUC value was 0,773.

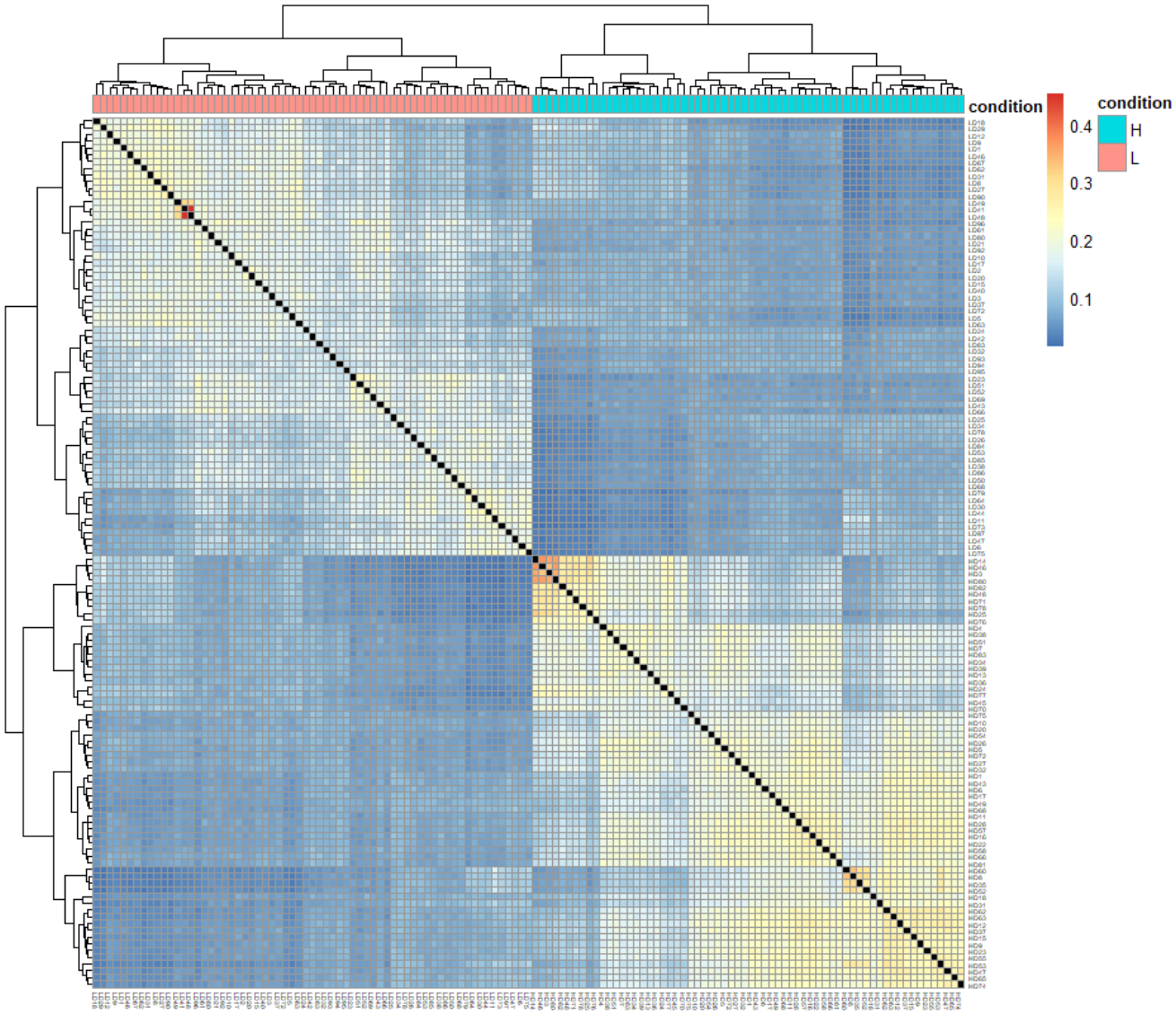

Figure 5: A heatmap based on the proximity measures from the RF classifier for the training set, with all predictors. Red colour represents closer samples, blue colour represents more distant samples. The black squares in diagonal direction represent the proximity measures between the same sample. The label above the heatmap corresponds to the condition of samples, where blue colour represents the samples from the healthy individuals and the red colour represents the samples from the liver diseased individuals. Two groups of samples can be distinguished, one in the upper left area, representing the samples from the liver diseased individuals, and the other one in the lower right area, representing the samples from the healthy individuals.

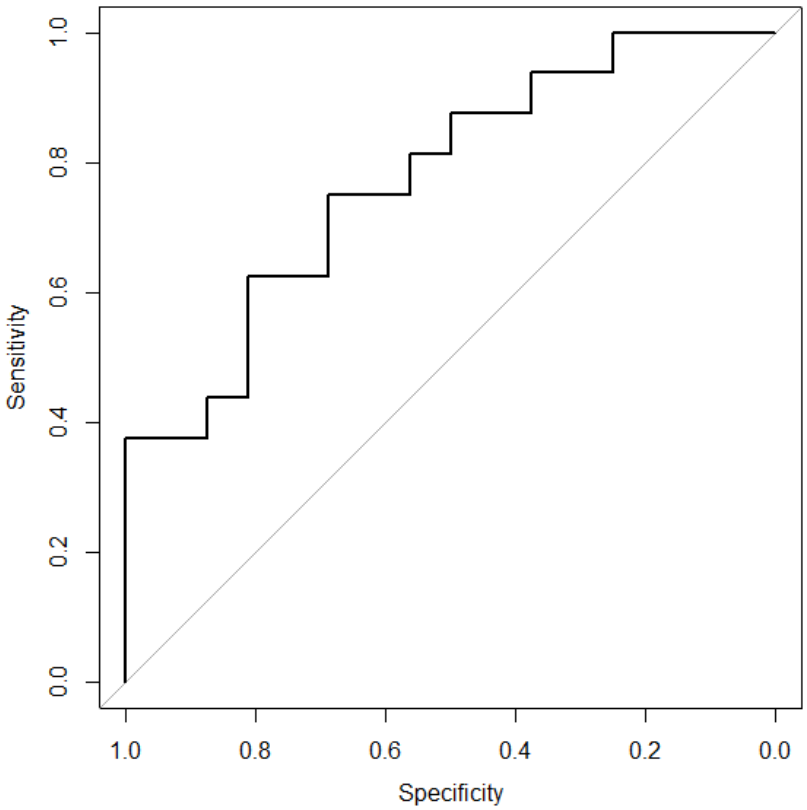

Figure 6: ROC curve for the RF classifier when all predictors were used. Sensitivity demonstrates the model's ability to detect samples from the liver diseased individuals, while the specificity demonstrates the model's ability to detect samples from the healthy individuals. The graph indicates that the model is not able to do both well simultaneously. If the threshold values are set to detect the samples from the diseased individuals with high accuracy, the samples from the healthy individuals will be misclassified in higher rate. The AUC value is 0,773.

4.2.2 Feature selection based on the graphical data analysis

After examining the data with UMAP plots, the best separation of samples based on their condition is achieved using the predictors *enrich* and *M*. The comparison of the data parting when only those predictors are used as opposed to all predictors is shown in Figure 7.

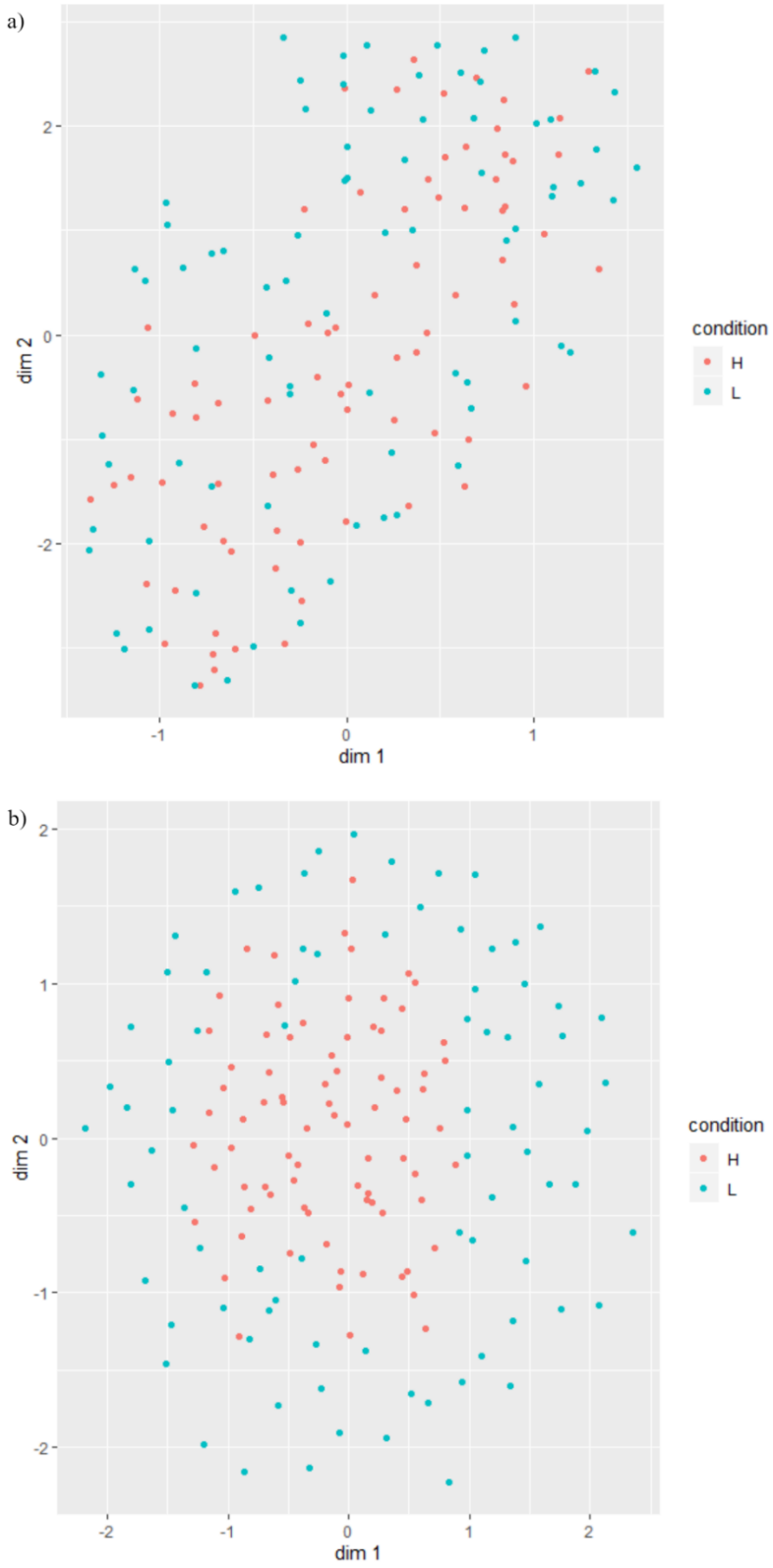

Figure 7: UMAP plots showing the samples coloured based on their condition when all predictors are used (a) and when only predictors *enrich* and *M* are used (b). The red samples (H) are from the healthy individuals and the blue samples (L) are from the liver diseased individuals. The distinction between samples is better when only *enrich* and *M variables* are used (b).

4.2.3 Random Forest assembly with predictors *enrich* and *M*

The optimal RF model gained by 5-fold cross-validation was the one which used 4733 variables. The RF classifier trained on the training set resulted with a training error of 31,78%. When the built RF model was used to classify the samples from the test set, the misclassification error was 18,75%. The ROC curve evaluating the model's performance is shown in Figure 8 and the heatmap based on the obtained proximity measures from the training set is shown in Figure 9. This model performed better than the model which uses all predictors and has lower misclassification rate and higher AUC value. From 6 samples which were misclassified, only 1 was from the healthy individuals and 5 were from the individuals with diseased liver.

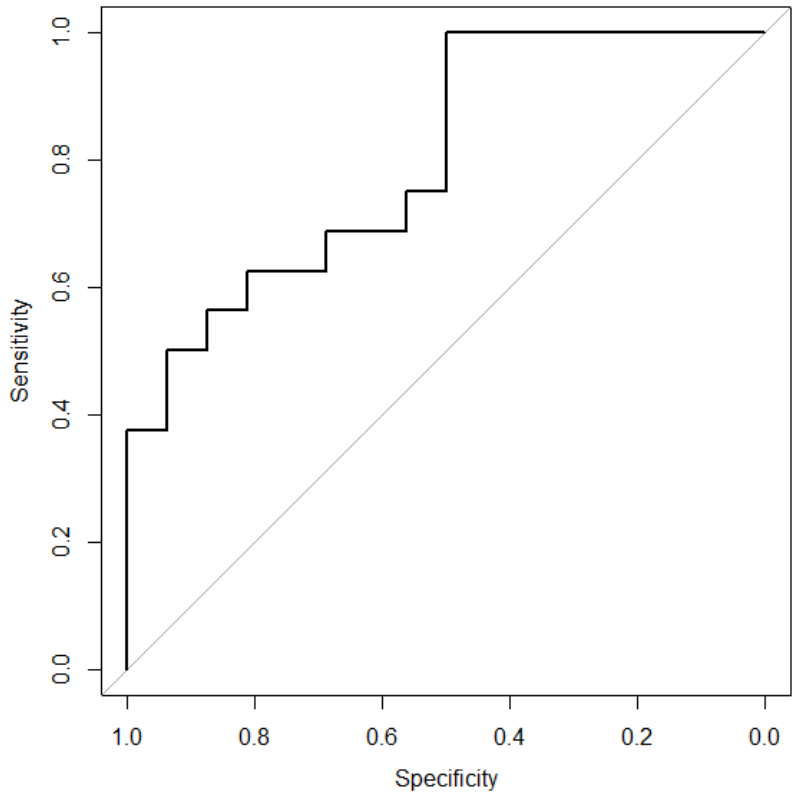

Figure 8: ROC curve for the RF classifier when predictors *enrich* and *M* were used. Sensitivity demonstrates the model's ability to detect samples from the liver diseased individuals, while the specificity demonstrates the model's ability to distinct samples from the healthy individuals. The graph indicates that the model is not able to do both well simultaneously. If the threshold values are set to detect the samples from the diseased individuals with high accuracy, the samples from the healthy individuals will be misclassified in higher rate. The AUC value is 0,801.

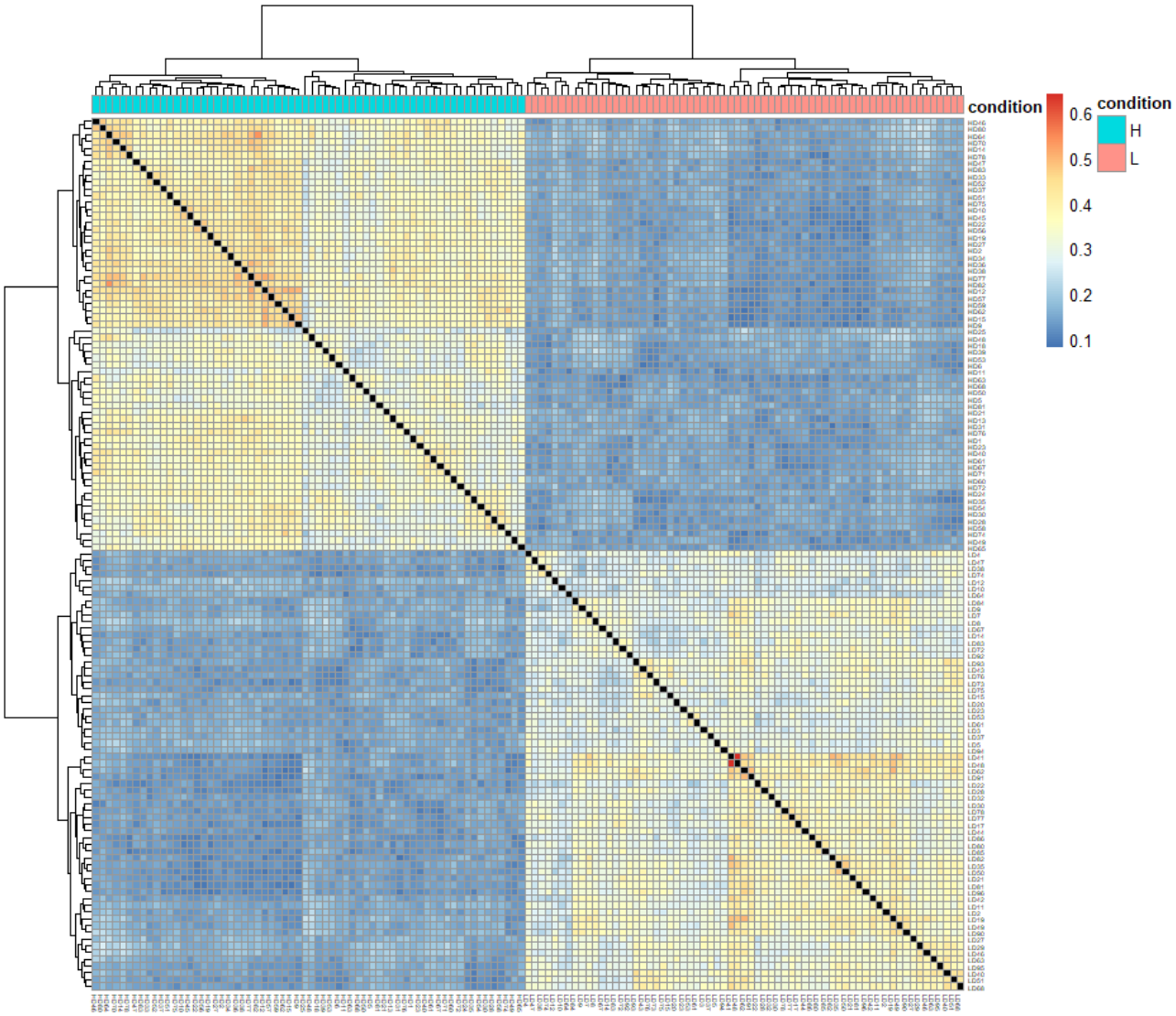

Figure 9: A heatmap based on the proximity measures from the RF classifier for the training set, with predictors *enrich* and *M*. Red colour represents closer samples, blue colour represents more distant samples. The black squares in diagonal direction represent the proximity measures between the same samples. The label above the heatmap corresponds to the condition of samples, where blue colour represents the samples from the healthy individuals and the red colour represents the samples from the liver diseased individuals. Two groups of samples can be distinguished, one in the upper left area, representing the samples from the healthy individuals, and the other one in the lower right area, representing the samples from the liver diseased individuals.

4.3 Exploratory analysis of all samples

4.3.1 Feature selection based on the calculated p-values

After calculating adjusted p-values for each predictor based on the variable importances computed by RF classifier, there were 524 predictors with adjusted p-value lower than 0.1. Top 20 predictors which had the highest importance measured in mean decrease in impurity are shown in Figure 10.

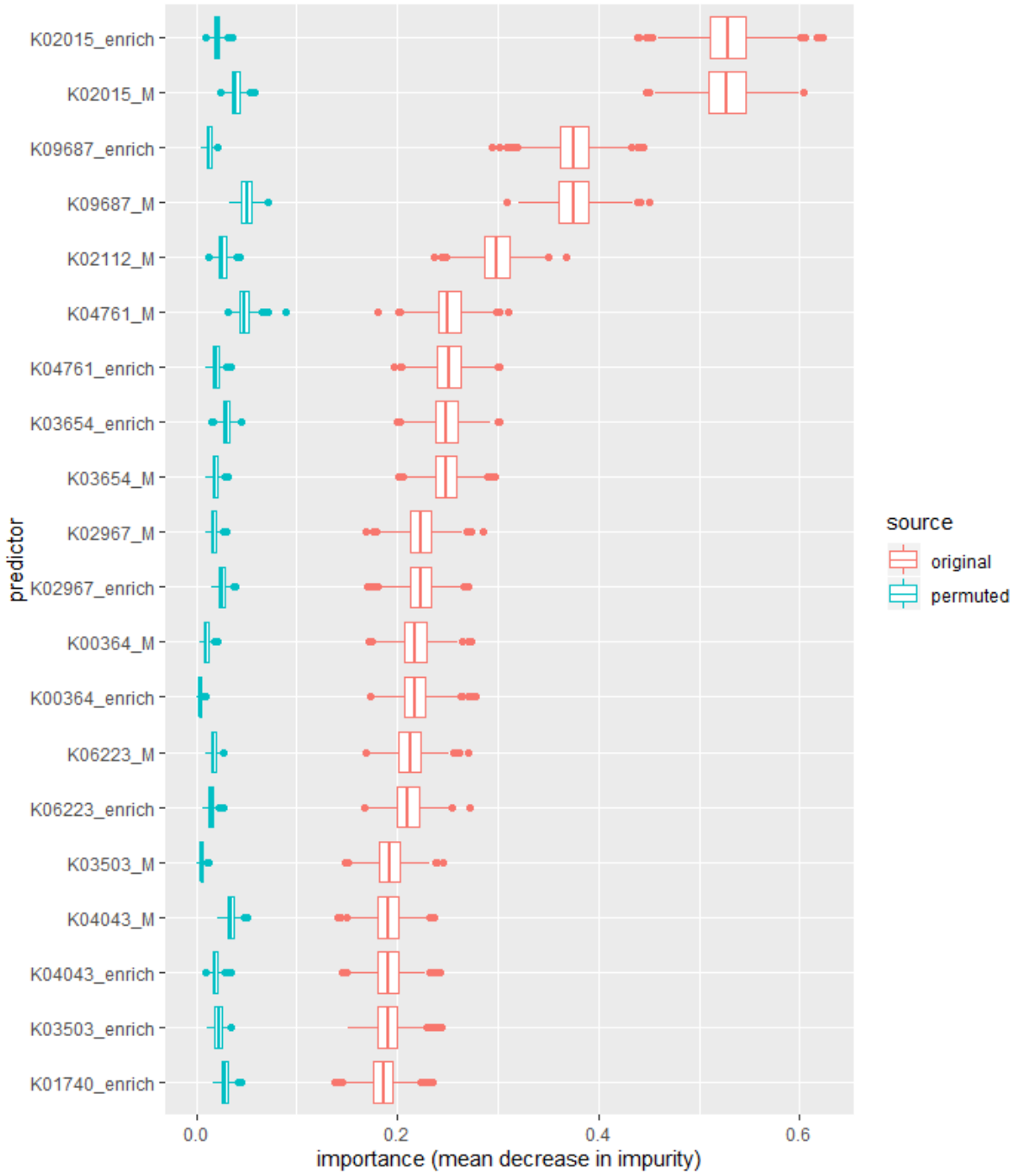

Figure 10: Variable importances for top 20 predictors obtained from the RF models built on the original (red) and the permuted data (blue).

A comparison of separation of the samples based on their condition when all predictors are used and only predictors *enrich* and *M* with adjusted p-value lower than 0.1 is shown in Figure 11. This PCA plot shows that the samples from the individuals with the diseased liver are more diverse than the samples from the healthy individuals.

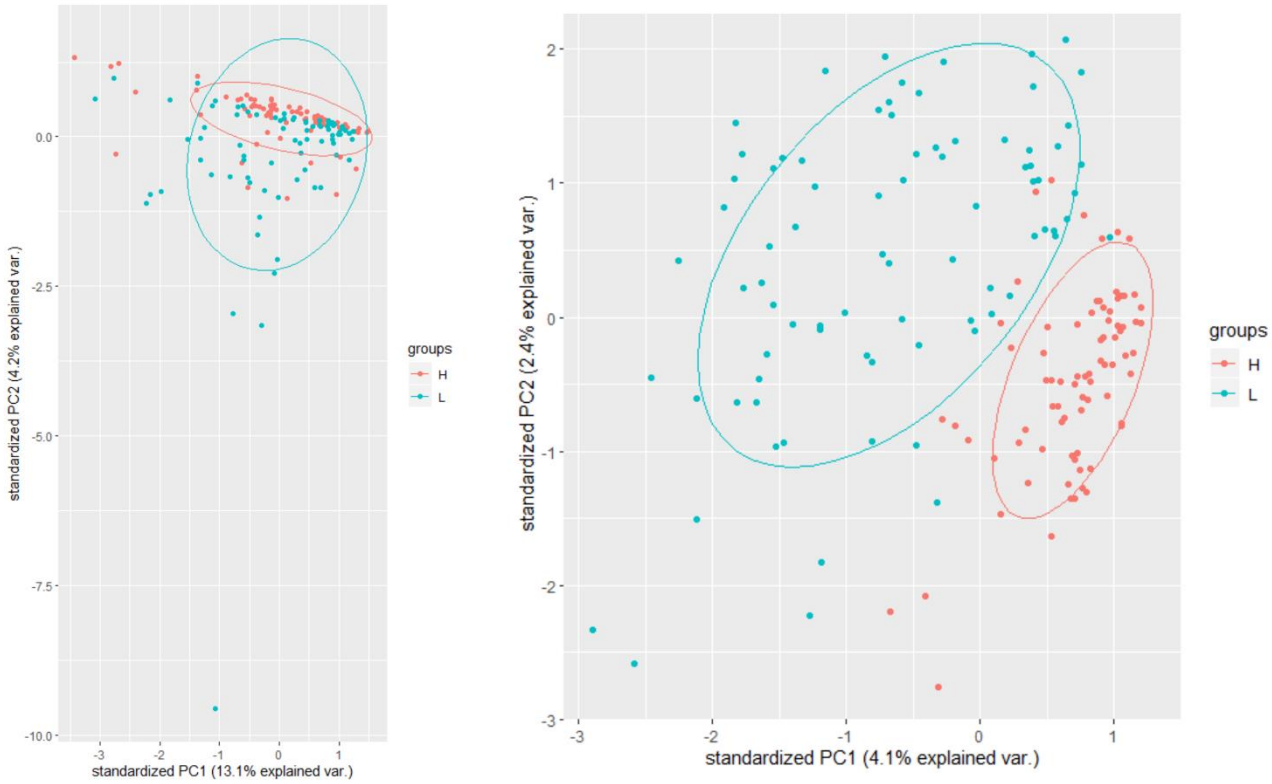

Figure 11: PCA plots showing the separation of the samples based on their condition, when a) all predictors are used and b) when only predictors enrich and M with adjusted p-value lower than 0.1 are used. The red samples (H) are from the healthy individuals and the blue samples (L) are from the liver diseased individuals. Using only a subset of predictors (b) improves the separability of samples based on their condition.

The exploratory analysis of all samples showed that it is possible to distinguish the samples from the healthy individuals and individuals with the diseased liver based only on a subset of predictors. In contrast, using all genes to classify the samples lowers the accuracy of the classification model and doesn't separate the samples well, as shown in Figure 11.

4.3.2 Random Forest assembly on all samples

The predicted optimal number of predictors for RF model trained on all samples with predictors *enrich* and *M* which had adjusted p-value lower than 0.1 was 3. A heatmap based on the obtained proximity measures from the trained model is shown in Figure 12. It reveals two clearly distinguishable groups of samples, where the samples from the liver diseased individuals are in the upper left corner, and the samples from the healthy individuals are in the lower right corner.

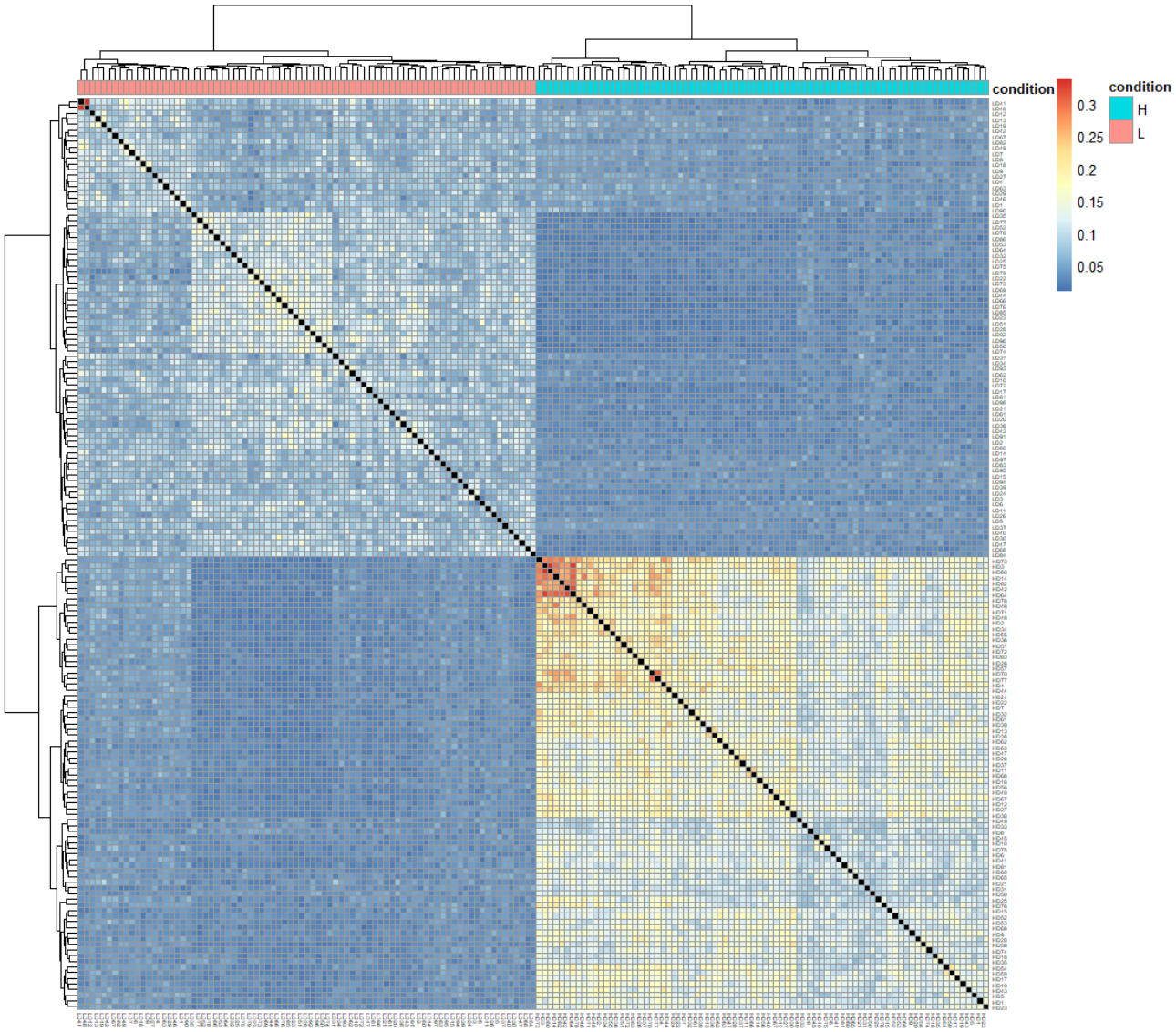

Figure 12: A heatmap based on the proximity measures from the RF built on all samples with predictors *enrich* and *M* which had adjusted p-value lower than 0.1. Red colour represents closer samples, blue colour represents more distant samples. The black squares in diagonal direction represent the proximity measures between the same samples. The label above the heatmap corresponds to the condition of samples, where blue colour represents the samples from the healthy individuals and the red colour represents the samples from the liver diseased individuals. Two groups of samples can be distinguished, one in the upper left area, representing the samples from the liver diseased individuals, and the other one in the lower right area, representing the samples from the healthy individuals.

4.3.3 Wilcoxon-Mann-Whitney rank sum test

Wilcoxon-Mann-Whitney rank sum test applied to all samples with predictors *enrich* and *M* resulted with 702 predictors with p-value lower than 0.5. Of those, 225 were the same as those deemed important based on variable importances, as shown in Figure 13. The separation of the samples based on the combination of these predictors is illustrated in Figure 14. When the data is separated using only the predictors obtained by the Wilcoxon-Mann-Whitney rank sum test, the samples did not separate well (Figure 14a). Using the intersection of these samples and the ones deemed important by the RF classifier improved the division (Figure 14b), but it was inferior to the separation achieved by the predictors from the RF classifier alone.

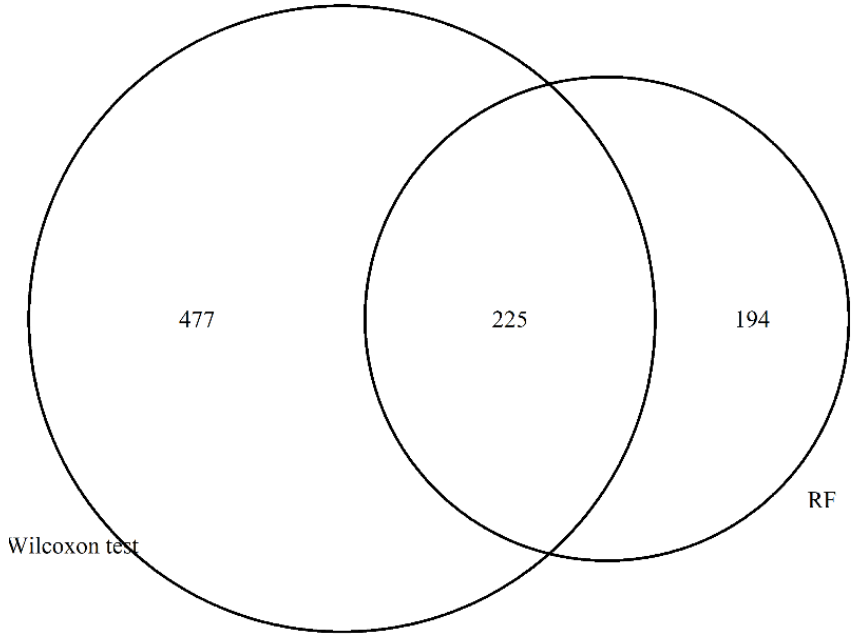

Figure 13: Venn diagram showing the intersection of predictors with p-value lower than 0.05 obtained by Wilcoxon test and by variable importances computed by RF.

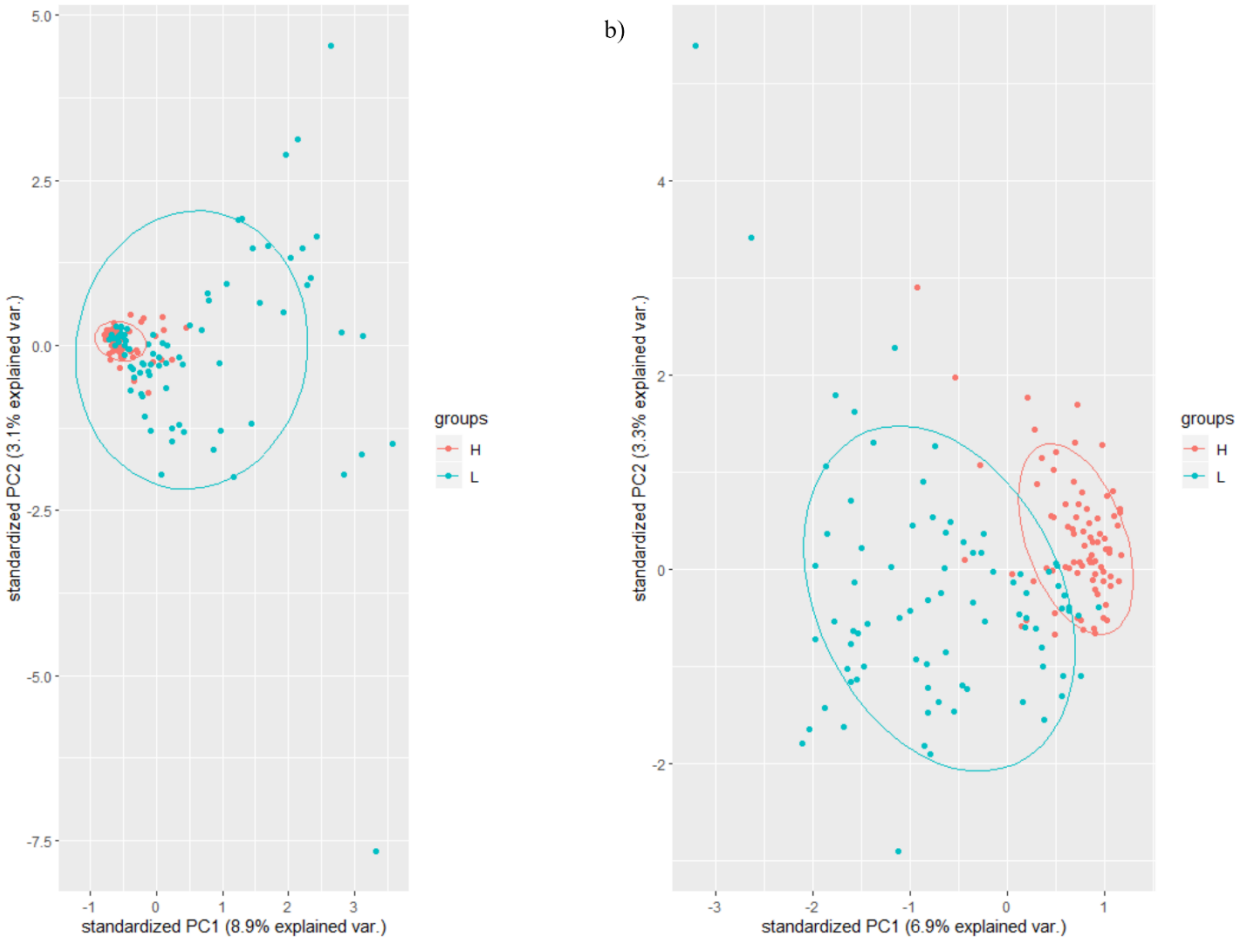

Figure 14: PCA plots showing the separation of the samples based on their condition when a) predictors with p-value lower than 0.05 from the Wilcoxon-Mann-Whitney rank sum test are used and b) when the intersection of predictors from the a) and the predictors deemed important by the RF classifier are used. The red samples (H) are from the healthy individuals and the blue samples (L) are from the liver diseased individuals. Samples separate notably only when the intersection of the predictors is used (b).

4.3.4 Codon usage analysis

The PCA plot showing how the samples separate based only on the codon frequencies is shown in Figure 15. It demonstrates that the differences in codon frequencies between the samples from the healthy individuals and individuals with the diseased liver alone are not enough to distinguish them, proving that the MELP statistics is essential for the data separation.

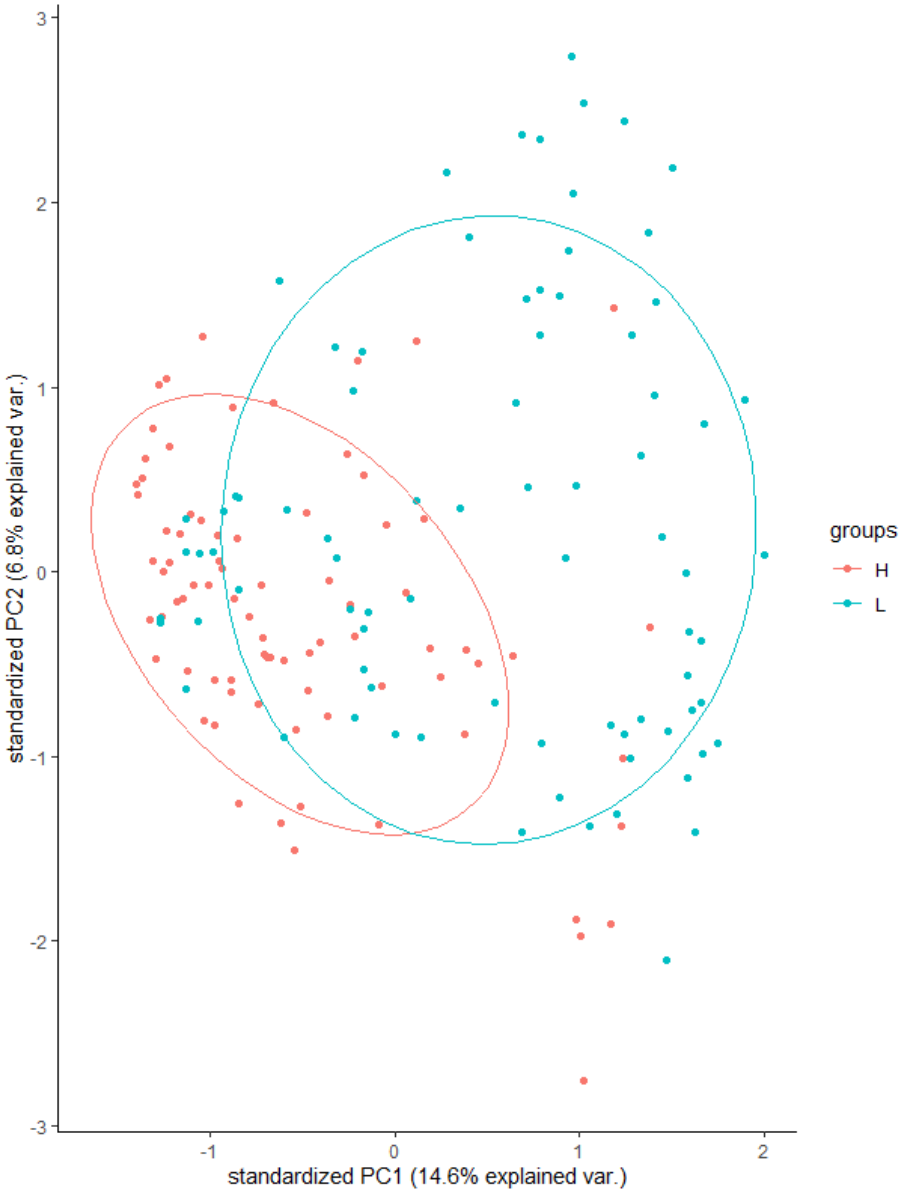

Figure 15: PCA plot showing the separation of the samples based on their codon frequencies. The red samples (H) are from the healthy individuals and the blue samples (L) are from the liver diseased individuals.

4.4 Metabolic pathway analysis

Gene enrichment was compared between the samples from the healthy individuals and individuals with diseased liver. GAGE analysis resulted with 5 significantly up-regulated and no significantly down-regulated pathways, as shown in Figure 16.

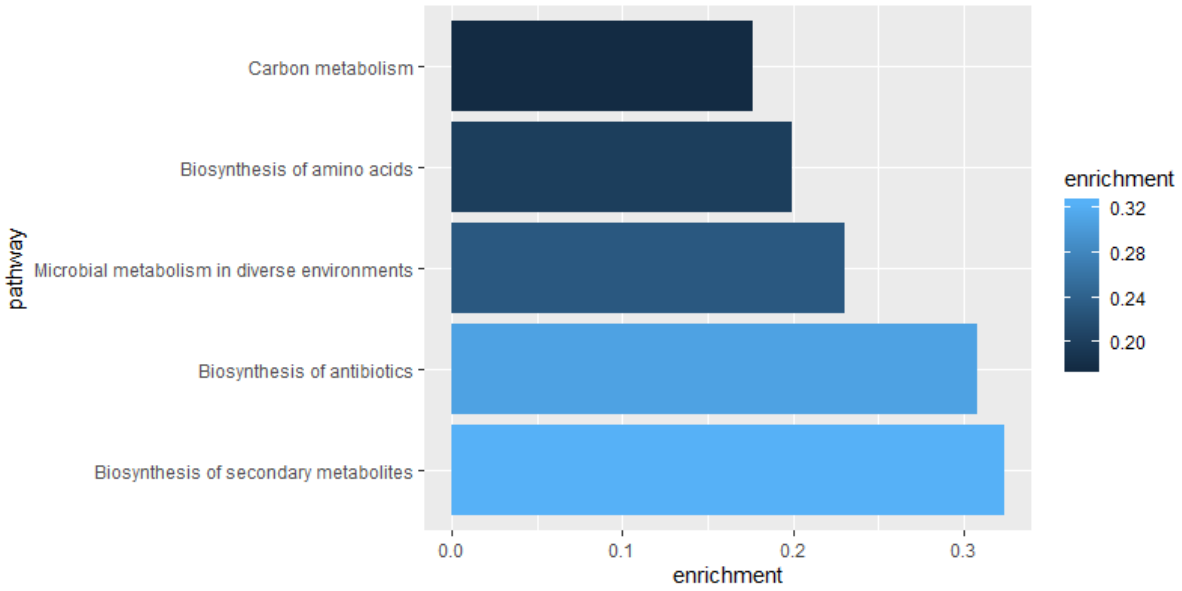

Figure 16: The comparison of enriched metabolic pathways between the samples from the healthy individuals and individuals with diseased liver. Shown pathways are enriched in the samples from the liver diseased individuals compared to the healthy individuals .

A more detailed inspection of orthologs connected to the pathway biosynthesis of antibiotics was carried out. There were 10 genes which KEGGREST database connected to this pathway that were significantly enriched in samples from the diseased individuals. Their KO identifiers and the corresponding enzymes and pathways are shown in Table 3.

Table 3: KO identifiers with the corresponding enzymes and pathways for orthologs which are connected to the pathway biosynthesis of antibiotics

KO	Enzyme	pathway
K00036	glucose-6-phosphate 1-dehydrogenase	pentose phosphate pathway
K00163	pyruvate dehydrogenase E1 component	pyruvate oxidation
K00164	2-oxoglutarate dehydrogenase E1 component	citrate cycle
K00825	kynurenine/2-aminoadipate aminotransferase	lysine degradation
K00832	aromatic-amino-acid transaminase	phenylalanine and tyrosine biosynthesis
K00891	shikimate kinase	shikimate pathway
K00927	phosphoglycerate kinase	glycolysis / gluconeogenesis
K01662	1-deoxy-D-xylulose-5-phosphate synthase	C5 isoprenoid biosynthesis
K01736	chorismate synthase	shikimate pathway
K11176	IMP cyclohydrolase	inosine monophosphate biosynthesis

4.5 The analysis of the phenotype information

4.5.1 Principal Component Analysis

The PCA applied to the additional biological information about samples is shown in Figure 17.

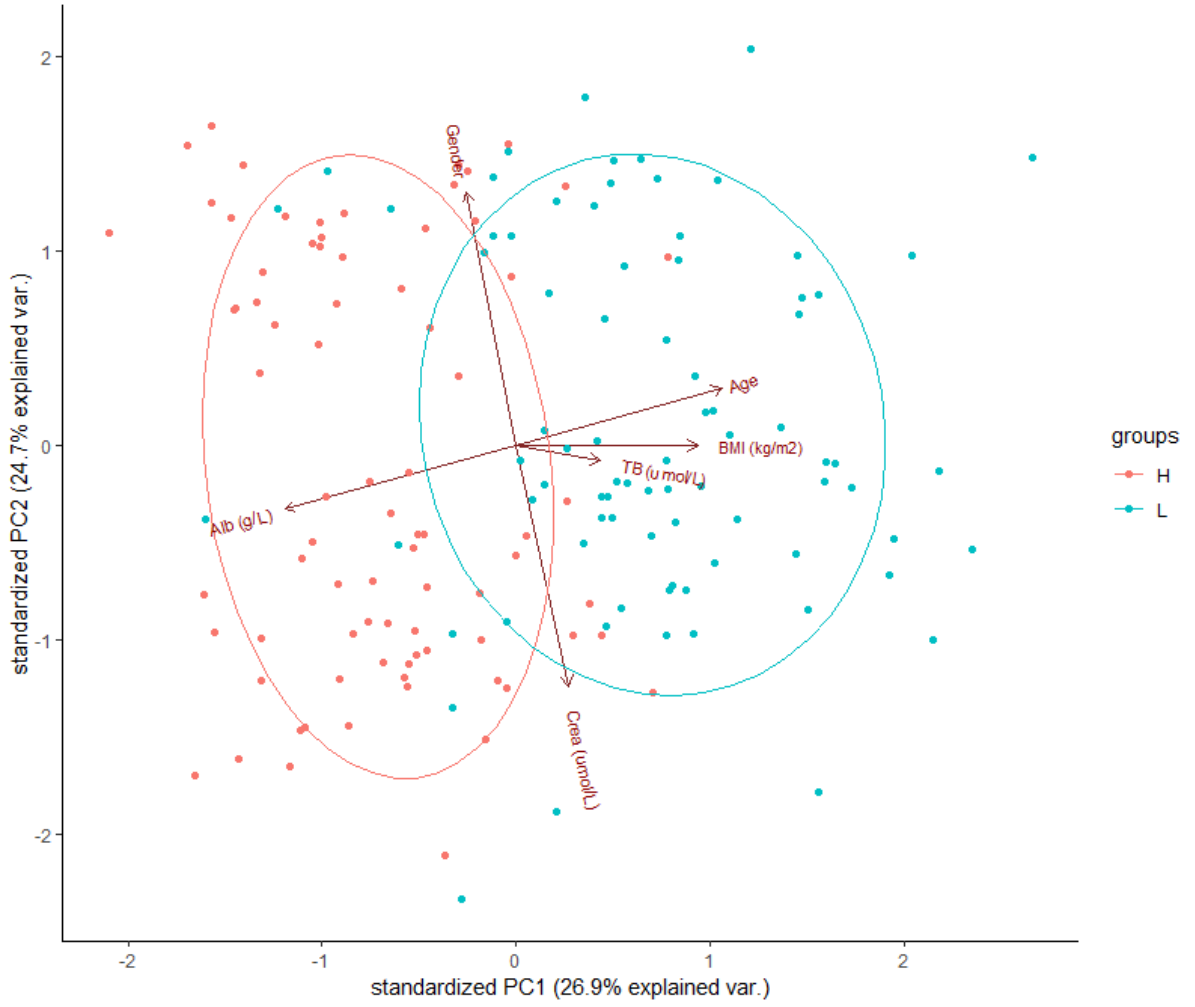

Figure 17: A PCA plot showing the parting of samples based on their phenotype information. The samples are coloured by their condition, where red colour represents the samples from the healthy individuals (H) and blue colour the samples from the individuals with diseased liver (L). The red arrows represent how the features affect the data parting. The samples which are in the pointing direction of an arrow have higher values of those features.

PCA plots displaying how the separation of the samples based on the important predictors is associated to the condition and the cause of the diseased liver are shown in Figure 18.

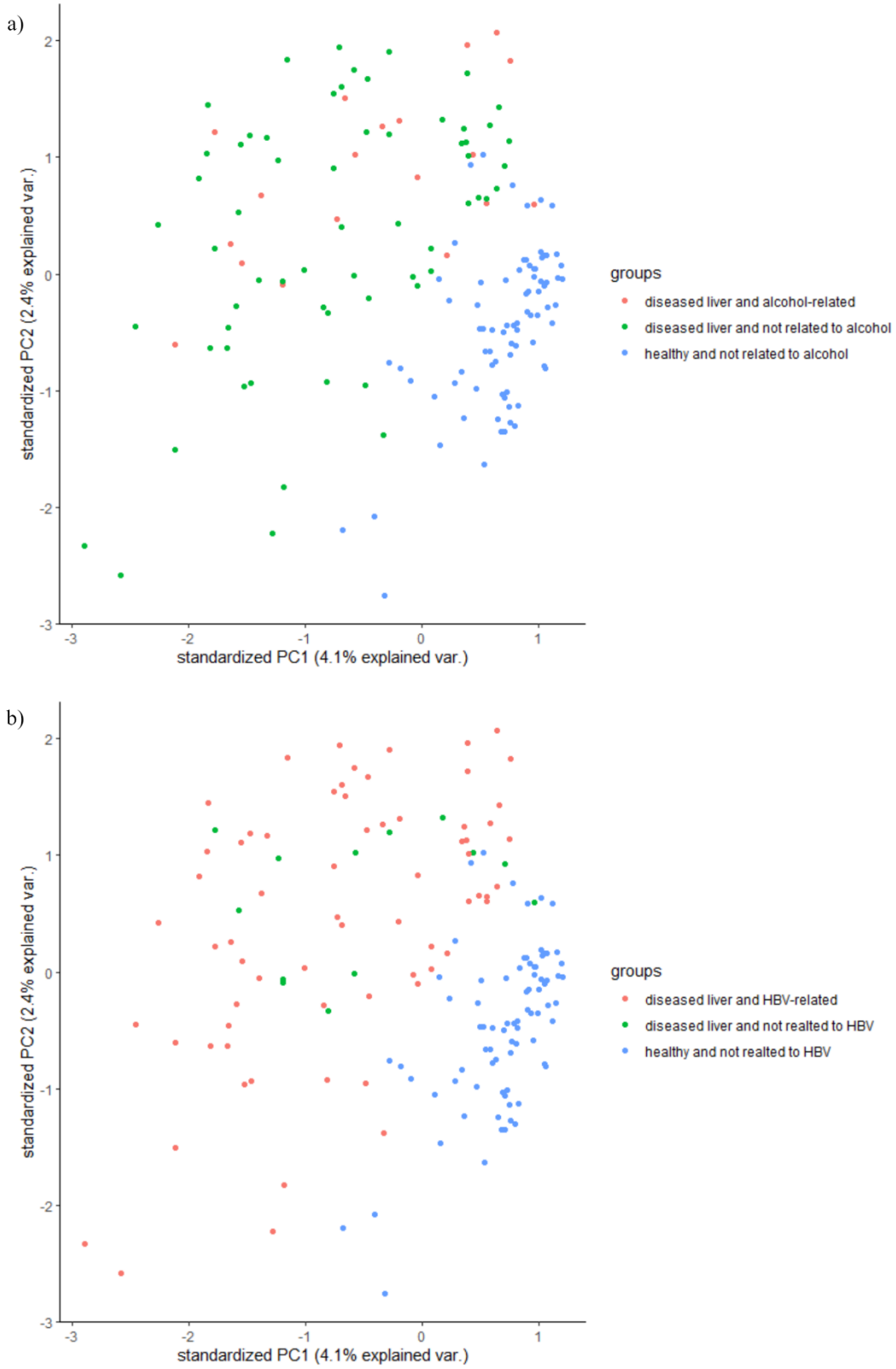

Figure 18: PCA plots showing which samples are associated to the alcohol-related cause (a) and HBV-related cause (b). The plot shows no grouping of the samples from the diseased individuals based on the cause.

4.5.2 Kruskal-Wallis test

A Kruskal-Wallis test resulted with no significant differences for the age, BMI and creatinine concentrations between the samples from the healthy individuals and individuals with diseased liver. It also resulted with a significant difference between the albumin concentrations with the p-value of 0.01567 and the total bilirubin concentrations with the p-value of 0.01957. The differences between these measures are shown in Figure 19.

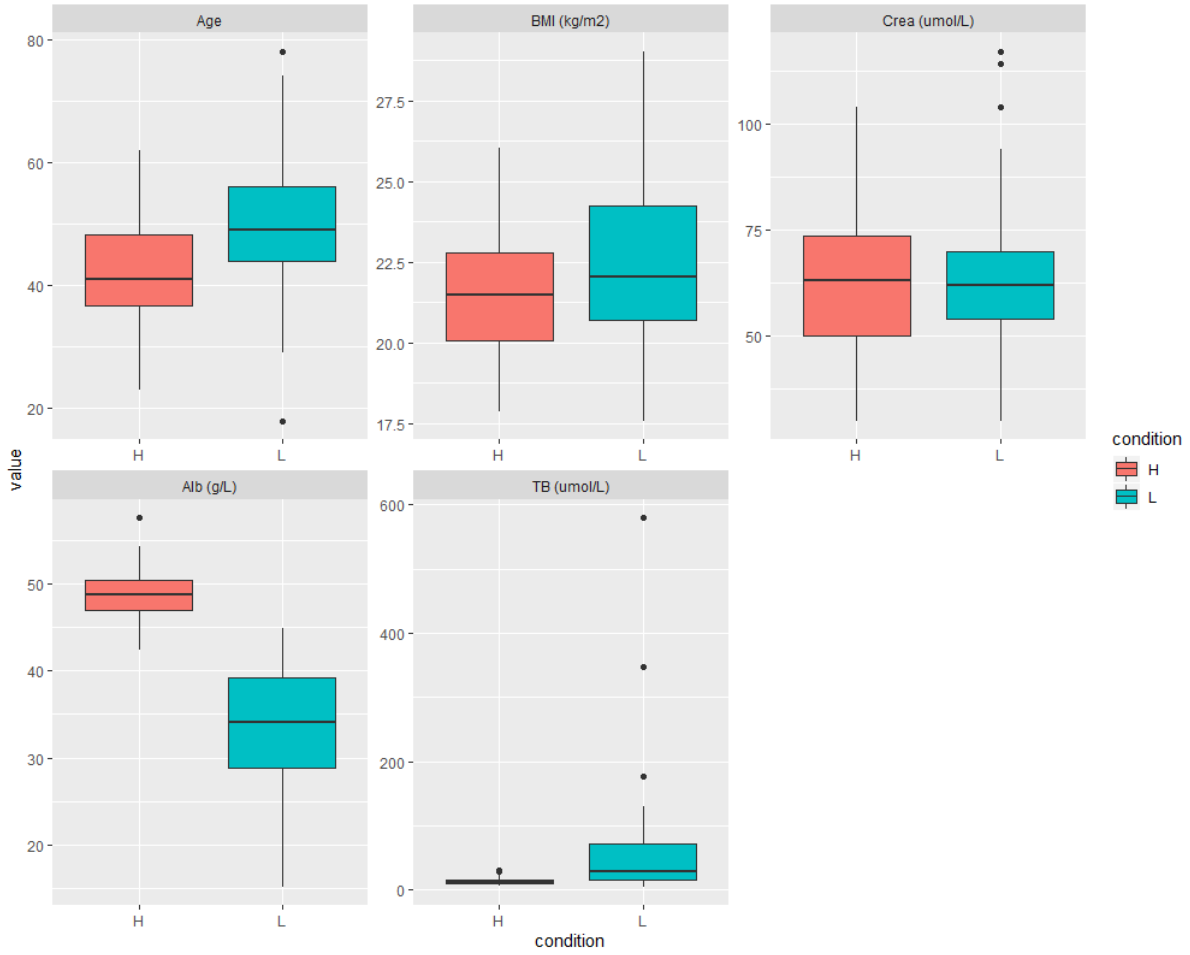

Figure 19: The differences between the variable measures for the samples from the healthy individuals (H) coloured red and the samples from the individuals with diseased liver (L) coloured blue. BMI = body mass index; Crea = creatinine concentration; Alb = albumin concentration; TB = total bilirubin concentration.

5. DISCUSSION

The authors of the original paper built a classification model based on the Support Vector Machine (SVM) algorithm (Qin *et al.*, 2014). First, they employed the Wilcoxon test to identify genes which were differentially abundant between the healthy and liver diseased individuals, choosing only those with the p-value lower than 0,01. The top 15 gut microbial gene markers were selected as the optimal subset of genes for patient discrimination. Its performance was assessed by ROC analysis and the obtained AUC value for the validation set was 0,836. The AUC value obtained with the RF classifier in this thesis based on the 418 predictors was 0,801, showing that it is inferior to the SVM classifier. Mentioned classifiers used different predictors, where the authors used the counts of genes as predictors, while the predictors in this thesis were based on the MELP statistics for each gene, resulting with larger number of predictors in RF model. It is also important to note that their training set had 181 samples (83 healthy and 98 liver diseased individuals) and the test set had 56 samples (31 healthy and 25 liver diseased individuals). Since not all samples were available, only 161 samples from their training set were used in this thesis. It was possible to separate the samples from the entire dataset based on their condition, which implies that codon usage bias in metagenomes holds enough information to train an effective classification model based on the RF algorithm. There are studies demonstrating that even the gut microbial communities composition from healthy individuals differ considerably (Huttenhower *et al.*, 2012). Most of the misclassified samples in this thesis were of individuals with the diseased liver, suggesting that they are more challenging to classify. This indication is supported by Figure 11b which shows that the samples from liver diseased individuals are more diverse than those from the healthy individuals. Using more samples to train the model would possibly considerably increase its accuracy. If the boundary between the samples based on their condition is not linear, the SVM algorithm can transform the dataset and project it into a higher dimension in which it could be linearly separable (James *et al.*, 2013). This approach might also improve the classification accuracy and it would be possible to implement it using the same MELP statistics as for the RF classifier.

Previous studies of the gut microbiome in human cirrhosis mostly focused on the comparison of species abundancies between healthy and diseased individuals (Acharya and Bajaj, 2019) and more detailed functional analyses are needed. GAGE analysis resulted with 5 up-regulated pathways, where the most up-regulated pathway is associated with the biosynthesis of secondary metabolites. This is probably due to the bacterial overgrowth in the intestinal tract and their production of endotoxins which are common in cirrhosis (Fouts *et al.*, 2012; Augustyn *et al.*, 2019). Previous research also observed a functional shift in microbiome toward endotoxin protein synthesis in cirrhosis (Acharya and Bajaj, 2019) and the top 15 markers which the authors identified also included biosynthesis of other secondary metabolites. Microbial metabolism in diverse environments is also enriched. The cause of this might be the need to adapt to the modified environment resulted by the gut flora alternations. While digested amino acids are metabolised by microorganisms in the intestinal tract and nutritionally impact their composition, diversity and activity, the amino acids are also synthesised by gut microbiota and of great importance for the host nutrition (Ma and Ma, 2019). Contrary to previous studies, the GAGE analysis resulted with enriched biosynthesis of amino acids in the samples from the

individuals with diseased liver. Some studies showed a higher expression of genes related to vitamins, cofactors, and oxidant metabolism in cirrhosis whereas controls had a significantly higher expression of carbohydrate and amino acid metabolism (Chen *et al.*, 2014; Bajaj *et al.*, 2015). After further inspection which orthologs are connected to the biosynthesis of antibiotics, there were 10 of them significantly different between the samples from the healthy and diseased individuals (Table 3). They are also connected to the pathways which are essential for the microbial growth, such as the citrate cycle, pyruvate oxidation, glycolysis, gluconeogenesis and amino acid synthesis. It should be considered that the GAGE analysis is performed only on genes with high expressivity, which is probably why it resulted with no down-regulated pathways. Also, a functional gene diversity of gut microbial communities between the patients with the liver cirrhosis caused by the alcohol and HBV was observed (Chen *et al.*, 2014) so another approach would be to analyse them separately.

The PCA of additional biological information (Figure 17) was performed to study which physiological features are significantly different between the ill and healthy individuals. It would be expected that the values of those features are related to the grouping of the samples after their division with the RF classifier. This analysis indicates that men and women are equally healthy and ill. It also indicates that the parting of the samples based on their condition is mostly influenced by the age and the albumin concentrations. Although some studies observed differences in abundancies of species in gut microbiome based on the BMI and gender (Gao *et al.*, 2018), there were no grouping of the samples based on these features. Samples from the diseased individuals appear to have lower albumin concentrations and be of older age. They also seem to have higher BMI and total bilirubin concentrations. The Kruskal-Wallis test established that the only significant differences of the mentioned features between the healthy and diseased individuals are in albumin and total bilirubin concentrations. This is expected since the patients with cirrhotic livers usually have lower levels of serum albumin and elevated levels of bilirubin. The liver is responsible for albumin synthesis which is compromised in cirrhotic livers and lower levels of albumin are produced (Walayat *et al.*, 2017). Additionally, portal blood flow is distorted in liver cirrhosis, which is accompanied by a decrease in hepatic elimination of bilirubin (Kim Iet al., 2015).

The results imply that the samples are not grouped based on some of the features, such as age, BMI and gender when employing MELP statistics. Another interesting question is would different diets, antibiotics and probiotics affect their separation. In the original paper, authors had exclusion criteria for the samples which, amongst others, included diabetes, obesity and the use of antibiotics or probiotics within 8 weeks before enrolment. There are studies showing how the diet alters the abundances of species and their functions in the human gut microbiome (David *et al.*, 2014) and how the obesity lowers the gut microbial gene richness (Cotillard *et al.*, 2013). It is known that the probiotics can change the population of microorganisms in the gut microbiota and that they have a role in the prevention of degenerative diseases (Cesaro *et al.*, 2011; Azad *et al.*, 2018). Antibiotics alter the composition of the gut microbiome, where they not only act on bacteria that cause infections but can also affect the resident microbiota indefinitely (Willing *et al.*, 2011; Yoon and Yoon, 2018). This can result in the dysregulation of host immune homeostasis and an increased susceptibility to disease. Since all the mentioned factors can modulate the abundance of different species as well as their gene richness in gut

microbiome, they should be considered in further analysis for developing the optimal classification model.

The PCA plot labelled with the condition of the samples and the cause of the cirrhotic liver (Figure 18) illustrated that the samples from the diseased individuals with the alcohol-related cause, as well as those with the HBV-related cause, are randomly scattered among all diseased samples. This implies the lack of grouping of the samples based on the cause when the analysis is based on the MELP statistics. It would be expected that there are differences in gut microbiome between the individuals with the HBV and alcohol caused liver disease since previous studies demonstrated enrichment of different microbial communities between these two groups (Engen *et al.*, 2015). The RF classifier was used to classify the samples based on their condition and not the cause, but the similar approach could be applied only to the samples from the diseased individuals to identify the genes in which they significantly differ based on the cause.

It should also be considered that the individuals from which the sample derive are of Han Chinese origin. In their research, the authors compared metagenomic species enriched in the healthy individuals from the Chinese population with the individuals from the Danish population and concluded that they were similarly correlated (Qin *et al.*, 2014). This indicates that the microbial communities of healthy individuals might be largely similar globally but needs broader research which would include more populations across the continents. Also, it lacks the same information about the individuals with the diseased liver and they are more diverse even within the same population.

As already mentioned, a great advantage of this method is that it only requires metagenomic samples. If this classification model would be slightly improved, it could be used as a valuable prediction tool in diagnosis of liver diseases. Instead of a long-term procedure which often results with liver biopsy to conclude the diagnosis, this method is non-invasive, inexpensive and does not require laborious laboratory work. There are still some drawbacks to this approach. DNA-based metagenomic analysis provides information on the gene expressivity but it lacks the information of the actual metabolic activity of microbial communities. It is focussed on the metagenomes and not the patient, which might be both an advantage and a disadvantage. It also has the potential to be used for different diseases since none of the steps in this method are specific to liver cirrhosis.

This research demonstrated that it is possible to distinguish the metagenomic samples from the healthy individuals and individuals with the cirrhotic liver based on the codon usage bias and RF classifier. Since only the genes with high expressivity were considered, perhaps including the genes with very low expressivity might improve the classification. If there were significant differences between those genes, adding them as predictors would probably improve the RF classifier's accuracy and offer additional information about the differences between the healthy and liver diseased individuals.

6. CONCLUSION

A random forest classifier for predicting whether individuals had a healthy or cirrhotic liver was built. Its misclassification error rate is 18,75%, where it mostly misclassifies the samples from the diseased individuals. After graphical analysis of the observations and training the model, the optimal subset of 4733 predictors was established based on the RF variable importances.

When the same approach was used as an exploratory analysis of the entire dataset instead of a training set, it was possible to separate the classes well using only 524 predictors. There was no grouping of the observations based on their gender, age, BMI, creatinine levels and the cause of the cirrhotic liver. The pathways which are enriched in microbial communities from the diseased individuals compared to healthy individuals were identified and characterised.

It was proven that using MELP statistics to determine the genes with high expressivity can be a useful tool for the mentioned analysis. The codon usage bias in metagenomes holds enough information to separate the samples based on their condition.

Lastly, since this approach requires only sequenced metagenomes, it has the potential to be applied to other diseases as well.

7. REFERENCES

Acharya, C. and Bajaj, J. S. (2019) 'Altered Microbiome in Patients With Cirrhosis and Complications', *Clinical Gastroenterology and Hepatology*. Elsevier, Inc, 17(2), pp. 307–321. doi: 10.1016/j.cgh.2018.08.008.

Augustyn, M., Grys, I. and Kukla, M. (2019) 'Small intestinal bacterial overgrowth and nonalcoholic fatty liver disease', *Clinical and Experimental Hepatology*, 5(1), pp. 1–10. doi: 10.5114/ceh.2019.83151.

Azad, M. A. K. *et al.* (2018) 'Probiotic Species in the Modulation of Gut Microbiota: An Overview', *BioMed Research International*, 2018. doi: 10.1155/2018/9478630.

Bajaj, J. S. *et al.* (2015) 'Salivary Microbiota Reflects Changes in Gut'. doi: 10.1002/hep.27819.

Benson, A. *et al.* (2009) 'Gut commensal bacteria direct a protective immune response against the human pathogen Toxoplasma gondii', *Cell Host Microbe*, 6(2), pp. 187–196. doi: 10.1016/j.chom.2009.06.005.Gut.

Botzman, M. and Margalit, H. (2011) 'Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles', *Genome Biology*. BioMed Central Ltd, 12(10), p. R109. doi: 10.1186/gb-2011-12-10-r109.

Brady, S. F. and Clardy, J. (2004) 'Palmitoylputrescine, an Antibiotic Isolated from the Heterologous Expression of DNA Extracted from Bromeliad Tank Water', pp. 1283–1286.

Breiman, L. (1996) 'Bagging Predictors', *Machine Learning*, 24, pp. 123–140.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45, pp. 5–32.

Cesaro, C. *et al.* (2011) 'Gut microbiota and probiotics in chronic liver diseases', *Digestive and Liver Disease*. Editrice Gastroenterologica Italiana, 43(6), pp. 431–438. doi: 10.1016/j.dld.2010.10.015.

Chen, Y. *et al.* (2014) 'Functional gene arrays-based analysis of fecal microbiomes in patients with liver cirrhosis', pp. 1–13.

Cotillard, A. *et al.* (2013) 'Dietary intervention impact on gut microbial gene richness', *Nature*, 500(7464), pp. 585–588. doi: 10.1038/nature12480.

David, L. A. *et al.* (2014) 'Diet rapidly and reproducibly alters the human gut microbiome', *Nature*. Nature Publishing Group, 505(7484), pp. 559–563. doi: 10.1038/nature12820.

Delong, E. F. *et al.* (2006) 'Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior', *Science*, 311(2006), pp. 496–503. doi: 10.1126/science.1120250.

Denisko, D. and Hoffman, M. M. (2018) 'Classification and interaction in random forests', *PNAS*, 115(8), pp. 1690–1692. doi: 10.1073/pnas.1800256115.

Diaz-Torres, M. L. *et al.* (2003) 'Novel Tetracycline Resistance Determinant from the Oral Metagenome', 47(4), pp. 1430–1432. doi: 10.1128/AAC.47.4.1430.

Dufour, D. R. *et al.* (2000) 'Diagnosis and monitoring of hepatic injury. I. Performance characteristics of laboratory tests', *Clinical Chemistry*, 46(12), pp. 2027–2049.

Engen, P. A. *et al.* (2015) 'The gastrointestinal microbiome: Alcohol effects on the composition of intestinal microbiota', *Alcohol Research: Current Reviews*, 37(2).

Entcheva, P. *et al.* (2001) 'Direct Cloning from Enrichment Cultures , a Reliable Strategy for Isolation of Complete Operons and Genes from Microbial Consortia', 67(1), pp. 89–99. doi: 10.1128/AEM.67.1.89.

Eschenfeldt, W. H. *et al.* (2001) 'DNA from Uncultured Organisms as a Source of 2 , 5-Diketo-D - Gluconic Acid Reductases', 67(9), pp. 4206–4214. doi: 10.1128/AEM.67.9.4206.

Fabijanić, M. and Vlahoviček, K. (2016) 'Big Data, Evolution, and Metagenomes: Predicting Disease', *Data Mining Techniques for the Life Sciences*, 1415, pp. 509–531. doi: 10.1007/978-1-4939-3572-7.

Fouts, D. E. *et al.* (2012) 'Bacterial translocation and changes in the intestinal microbiome in mouse models of liver disease', *Journal of Hepatology*. European Association for the Study of the Liver, 56(6), pp. 1283–1292. doi: 10.1016/j.jhep.2012.01.019.

Gao, X. *et al.* (2018) 'Body Mass Index Differences in the Gut Microbiota Are Gender Specific', 9(June), pp. 1–10. doi: 10.3389/fmicb.2018.01250.

Grantham, R. *et al.* (1980) 'Codon catalog usage and the genome hypothesis', *Nucleic Acids Research*, 8(1), pp. 49–62.

He, B. *et al.* (2007) 'Article Intestinal Bacteria Trigger T Cell-Independent Immunoglobulin A 2 Class Switching by Inducing Epithelial-Cell Secretion of the Cytokine APRIL', *Immunity*, 26, pp. 812–826. doi: 10.1016/j.immuni.2007.04.014.

Hooper, L. V *et al.* (2001) 'Molecular Analysis of Commensal Host-Microbial Relationships in the Intestine', *Science*, 291(February), pp. 881–885.

Hooper, L. V and Macpherson, A. J. (2010) 'Immune adaptations that maintain homeostasis with the intestinal microbiota', *Nature Reviews Immunology*. Nature Publishing Group, 10(3), pp. 159–169. doi: 10.1038/nri2710.

Huttenhower, C. *et al.* (2012) 'Structure, function and diversity of the healthy human microbiome', *Nature*. Nature Publishing Group, 486(7402), pp. 207–214. doi: 10.1038/nature11234.

Ikemura, T. (1981) 'Correlation between the Abundance of Escherichia coli Transfer RNAs and the Occurrence of the Respective Codons in its Protein Genes : A Proposal for a Synonymous Codon Choice that is Optimal for the E . coli Translational System', *Journal of Molecular Biology*, 151(3), pp. 389–409.

Ivanov, I. I. *et al.* (2008) 'Specific microbiota direct the differentiation of Th17 cells in the mucosa of the small intestine', *Cell Host Microbe*, 16(4), pp. 337–349.

James, G. *et al.* (2013) *An Introduction to Statistical Learning with Applications in R*.

Kim, H. J., Lee, H. K. and Cho, J. H. (2015) 'Factor analysis of the biochemical markers related to liver cirrhosis', *Pakistan Journal of Medical Sciences*, 31(5), pp. 1043–1046. doi: 10.12669/pjms.315.8025.

Klare, I., Werner, G. and Witte, W. (2001) ‘Enterococci. Habitats, Infections, Virulence Factors, Resistances to Antibiotics, Transfer of Resistance Determinants.’, *Contrib. Microbiol.*, 8, pp. 108–122.

Ma, N. and Ma, X. (2019) ‘Dietary Amino Acids and the Gut-Microbiome-Immune Axis : Physiological Metabolism and Therapeutic Prospects’, *Comprehensive Reviews in Food Science and Food Safety*, 18. doi: 10.1111/1541-4337.12401.

Martens, E. C., Chiang, H. C. and Gordon, J. I. (2008) ‘Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont’, *Cell Host & Microbe*, 4(5), pp. 447–457. doi: 10.1016/j.chom.2008.09.007.Mucosal.

McInnes, L., Healy, J. and Melville, J. (2018) ‘UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction’. Available at: http://arxiv.org/abs/1802.03426.

Nishikawa, H. and Osaki, Y. (2015) ‘Liver Cirrhosis: Evaluation, Nutritional Status, and Prognosis’, *Mediators of Inflammation*. Hindawi Publishing Corporation, 2015. doi: 10.1155/2015/872152.

Plotkin, J. B. and Kudla, G. (2011) ‘Synonymous but not the same: the causes and consequences of codon bias’, *Nat Rev Genet.*, 12(1), pp. 32–42. doi: 10.1038/nrg2899.Synonymous.

Qin, N. *et al.* (2014) ‘Alterations of the human gut microbiome in liver cirrhosis’, *Nature*. Nature Publishing Group, 513(7516), pp. 59–64. doi: 10.1038/nature13568.

Roller, M. *et al.* (2013) ‘Environmental shaping of codon usage and functional adaptation across microbial communities’, *Nucleic Acids Research*, 41(19), pp. 8842–8852. doi: 10.1093/nar/gkt673.

Sender, R., Fuchs, S. and Milo, R. (2016) ‘Revised Estimates for the Number of Human and Bacteria Cells in the Body’, pp. 1–14. doi: 10.1371/journal.pbio.1002533.

Si-Tayeb, K., Lemaigre, F. P. and Duncan, S. A. (2010) ‘Organogenesis and Development of the Liver’, *Developmental Cell*, 18(2), pp. 175–189. doi: 10.1016/j.devcel.2010.01.011.

Šimonovský, V. (1999) ‘The diagnosis of cirrhosis by high resolution ultrasound of the liver surface’, *British Journal of Radiology*, 72(JAN.), pp. 29–34. doi: 10.1259/bjr.72.853.10341686.

Slingerland, A. E. and Stein-Thoeringer, C. K. (2018) ‘Microbiome and diseases: Neurological disorders’, *The Gut Microbiome in Health and Disease*, pp. 295–310. doi: 10.1007/978-3-319-90545-7_18.

Steele, H. L., Jaeger, K. and Streit, W. R. (2009) ‘Advances in Recovery of Novel Biocatalysts from Metagenomes’, pp. 25–37. doi: 10.1159/000142892.

Supek, F. *et al.* (2010) ‘Translational Selection Is Ubiquitous in Prokaryotes’, *PLoS Genetics*, 6(6).

Supek, F. and Vlahoviček, K. (2005) ‘Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity’, 6(182), pp. 1–15. doi: 10.1186/1471-2105-6-182.

Thomas, T., Gilbert, J. and Meyer, F. (2012) 'Metagenomics - a guide from sampling to data analysis', *Microbial Informatics and Experimentation*. BioMed Central Ltd, 2(1), p. 3. doi: 10.1186/2042-5783-2-3.

Tringe, S. G. and Rubin, E. M. (2005) 'Metagenomics: DNA sequencing of environmental samples.', *Nature Genetics*, 6(October), pp. 805–814. doi: 10.1038/nrg1709.

Walayat, S. *et al.* (2017) 'Role of albumin in cirrhosis: from a hospitalist's perspective', *Journal of Community Hospital Internal Medicine Perspectives*. Taylor & Francis, 7(1), pp. 8–14. doi: 10.1080/20009666.2017.1302704.

Willing, B. P., Russell, S. L. and Finlay, B. B. (2011) 'Shifting the balance: Antibiotic effects on host-microbiota mutualism', *Nature Reviews Microbiology*. Nature Publishing Group, 9(4), pp. 233–243. doi: 10.1038/nrmicro2536.

Wilson, M. R. *et al.* (2019) 'Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis', *New England Journal of Medicine*, 380(24), pp. 2327–2340. doi: 10.1056/NEJMoa1803396.

Yamao, F., Andachi, Y. and Muto, A. (1991) 'Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins.', *Nucleic Acids Res.*, 19(22), pp. 6119–6122.

Yoon, M. Y. and Yoon, S. S. (2018) 'Disruption of the gut ecosystem by antibiotics', *Yonsei Medical Journal*, 59(1), pp. 4–12. doi: 10.3349/ymj.2018.59.1.4.

Website pages:

https://www.genome.jp/kegg/ko.html

SUPPLEMENT

R scripts

```
library(data.table)
library(coRdon)
library(stringr)
library(Biobase)

set.seed(17)

my_fasta_file_folder <- "C:\\Users\\Maja\\Desktop\\Eva -
diplomski\\samples\\svi"
uzorci <- readSet(my_fasta_file_folder, prepend.filenames = T)
uzorci <- codonTable(uzorci)


A function which calculates MELP values and does enrichment analysis:
enriching <- function(codon_table, sample_name){
  sample <- subset(codon_table, str_extract(coRdon::getID(codon_table),
"[HL]D\\d+") == sample_name)
  melp <- MELP(sample, ribosomal = TRUE, filtering = "hard", len.threshold
= 80, stop.rm = TRUE)
  ct <- crossTab(getKO(sample), as.numeric(melp), threshold = 1L)
  enr <- enrichment(ct)
  enr_data <- pData(enr)
  enr_data$sample <- sample_name
  return(list(enr_data))
}


sample_dt <- data.table(samples =
unique(str_extract(coRdon::getID(uzorci), "[HL]D\\d+")))
sample_dt[, enrichment := list(enriching(uzorci, samples)), by = samples]

samples <- do.call("rbind", sample_dt$enrichment)
samples$condition <- substr(samples$sample, 1, 1)

#saveRDS(samples, "C:\\Users\\Maja\\Desktop\\Eva -
diplomski\\samples\\all_samples_filtered.rds")
```

Preparing the data for training the model:

```
#sampleSet <- samples[samples$gt_1 > 0, -c(7, 8, 10)]
sampleSet <- samples[samples$gt_1 > 0, c("sample", "category", "enrich",
"M")]
sampleSet <- melt(sampleSet, id.var = c("sample", "category"))
sampleSet$colnames <- paste(sampleSet$category, sampleSet$variable, sep =
"_")
sampleSet <- sampleSet[, c(1, 4, 5)]

sample_names <- unique(sampleSet$sample)
condition <- factor(substr(sample_names, 1, 1))

sampleSet <- dcast(sampleSet, sample ~ colnames, fill = 0)
#sampleSet <- sampleSet[, which(colnames(sampleSet) %in%
selected_predictors)]

sampleSet <- data.frame(scale(sampleSet, center = T, scale = T))

sampleSet$condition <- condition
rownames(sampleSet) <- sample_names


Training the model:
train_index <- createDataPartition(y = sampleSet$condition,
                                   p = 0.5,
                                   list = FALSE)

train <- sampleSet[train_index, ]
test <- sampleSet[-train_index, ]


train_ctrl <- trainControl(method = "cv",
                           number = 5,
                           returnResamp = "all",
                           verboseIter = TRUE)

RFmodel <- train(condition ~ .,
                 data = train,
                 method = 'ranger',
                 tuneLength = 50,
                 trControl = train_ctrl,
                 num.trees = 10000,
                 importance = "impurity")
```

```
RFmodel$finalModel

rf <- ranger(condition ~ ., train, num.trees = 10000, mtry =
RFmodel$finalModel$mtry,
             probability = FALSE, replace = TRUE, oob.error = TRUE,
             classification = TRUE, importance = "impurity")

RFpredict <- predict(rf, test)

mc_rate <- mean(RFpredict$predictions != test$condition)
mc_rate


Heatmaps:
prox <- extract_proximity(rf, sampleSet)
colnames(prox) <- rownames(sampleSet)
rownames(prox) <- rownames(sampleSet)

labels <- data.frame(condition = sampleSet$condition)
rownames(labels) <- rownames(sampleSet)

pheatmap(prox, cellheight = 4, cellwidth = 4, fontsize = 4,
         annotation_col = labels)


ROC analysis:
roc.curve <- function(m){
  train_index <- createDataPartition(y = sampleSet$condition,
                                     p = 0.8,
                                     list = FALSE)
  train <- sampleSet[train_index, ]
  test <- sampleSet[-train_index, ]

  rf <- ranger(condition ~ ., train, num.trees = 10000, mtry = m,
               probability = TRUE, replace = TRUE, oob.error = TRUE,
               classification = TRUE, importance = "impurity")

  RFpredict <- predict(rf, test)

  rf.roc <- roc(test$condition, RFpredict$predictions[, 2])
  plot(rf.roc, xlim = c(0,1))
  return(auc(rf.roc))
}
```

```
Feature selection based on the graphical data analysis:
library(umap)
umap_dt <- umap(sampleSet[, -ncol(sampleSet)])
umap_dt <- data.table(umap_dt$layout)
ggplot(umap_dt, aes(V1, V2, colour = condition)) + geom_point() +
  xlab("dim 1") + ylab("dim 2")

pca <- prcomp(sampleSet[, -ncol(sampleSet)], center = T, scale. = T)
ggbiplot(pca, var.axes = F, groups = sampleSet$condition, ellipse = T)




Calculating p-values for each predictor:
doitall <- function(filename, m){

  sampleSet <- readRDS(filename)

  #original
  ranger.imp <- function(){
    rf <- ranger(condition ~ ., sampleSet, num.trees = 10000, mtry = m,
probability = FALSE,
                 replace = TRUE, classification = TRUE, importance =
"impurity")
    return(list(rf$variable.importance))
  }

  dt <- data.table(rf = 1:1)
  dt[, var_imp := list(ranger.imp()), by = rf]

  #permutated:
  perm <- data.table(copy(sampleSet))
  perm <- perm[, lapply(.SD, sample, nrow(perm), replace = TRUE), .SDcols
= 1:(ncol(perm)-1)]
  perm[, condition := factor(substr(rownames(sampleSet), 1, 1)), by = .I]

  perm.ranger.imp <- function(){
    rf_perm <- ranger(condition ~ ., perm, num.trees = 10000, mtry = m,
probability = FALSE,
                      replace = TRUE, classification = TRUE, importance =
"impurity")
    return(list(rf_perm$variable.importance))
  }
```

```
  dt_perm <- data.table(rf = 1:1)
  dt_perm[, var_imp := list(perm.ranger.imp()), by = rf]

  saveRDS(dt, file = paste("var_imp", filename, sep = "_"))
  saveRDS(dt_perm, file = paste("perm_var_imp", filename, sep = "_"))

}

system.time(doitall(sampleSet, RFmodel$finalModel$mtry))



#original:
var_imp <- readRDS("var_imp_sampleSet.RDS")
vimp <- data.table(rf = rep(1:1000, each =
length(unlist(var_imp$var_imp[1]))),
                   gene = names(unlist(var_imp$var_imp)),
                   importance = unlist(var_imp$var_imp))
vimp <- data.table(dcast(vimp, rf ~ gene, value.var = "importance"))

#permuted:
perm_var_imp <- readRDS("perm_var_imp_sampleSet.RDS")
pvimp <- data.table(rf = rep(1:1000, each =
length(unlist(perm_var_imp$var_imp[1]))),
                    gene = names(unlist(perm_var_imp$var_imp)),
                    importance = unlist(perm_var_imp$var_imp))
pvimp <- data.table(dcast(pvimp, rf ~ gene, value.var = "importance"))


p.value <- function(predictor){
  smallest_real <- min(unlist(vimp[, ..predictor]))
  p_value <- 1 - mean(unlist(pvimp[, ..predictor]) < smallest_real)
  return(p_value)
}


#calculating p-values for each predictor:
pvalues <- data.table(predictor = unique(names(unlist(var_imp$var_imp))))
pvalues[, pvalue := p.value(predictor), by = predictor]
pvalues[, padj := p.adjust(pvalue, method = "bonferroni"), by = predictor]

selected_predictors <- pvalues[padj < 0.1, predictor]
#saveRDS(selected_predictors, "predictors_pvalues.RDS")
```

```
Wilcoxon test:
wilcox.testing <- function(predictor){
  t <- wilcox.test(unlist(sampleSet[1:80, ..predictor]),
unlist(sampleSet[81:161, ..predictor]))
  return(t$p.value)
}

wilcox_enrM <- data.table(predictor = colnames(sampleSet)[-c(1,
length(colnames(sampleSet)))])
wilcox_enrM[, p_value := wilcox.testing(predictor), by = predictor]
wilcox_enrM[, padj := p.adjust(wilcox_enrM$p_value)]
library(VennDiagram)
venn.diagram(list(RF = selected_predictors,
                  Wilcoxon = wilcox_enrM[p_value < 0.05, predictor]),
             category.names = c("RF", "Wilcoxon test"),
             filename = "venn_diagram_RF_wilcox.png")

CU analysis:
my_fasta_file_folder <- "C:\\Users\\Maja\\Desktop\\Eva -
diplomski\\samples\\svi"
codon.freq <- function(path){
  uzorci <- readSet(my_fasta_file_folder, prepend.filenames = T)
  codon_table <- codonTable(uzorci)
  codon_usage <- data.table(sample = str_extract(getID(codon_table),
"[HL]D\\d+"),
                            length = getlen(codon_table),
                            KO = getKO(codon_table),
                            codonCounts(codon_table))
  codon_usage[, 4:ncol(codon_usage)] <- codon_usage[, 4:ncol(codon_usage)]
/ codon_usage$length
  return(codon_usage)
}

codon_usage <- codon.freq(my_fasta_file_folder)
#saveRDS(codon_usage, "C:\\Users\\Maja\\Desktop\\Eva -
diplomski\\samples\\codon_usage.rds")
```

```
A function which analyses whether HD vs LD samples are distinguished based
on codon frequencies:
codon.usage <- function(codon_usage){
  codon_usage <- codon_usage[length > 80]
  codon_usage <- codon_usage[, lapply(.SD, median), .SDcols =
4:ncol(codon_usage),
                             by = c("sample", "KO")]
  codon_usage <- codon_usage[, lapply(.SD, median), .SDcols =
3:ncol(codon_usage),
                             by = "sample"]
  codon_usage$condition <- factor(substr(codon_usage$sample, 1, 1))

  PCA <- prcomp(codon_usage[, -c("sample", "condition")])

  return(ggbiplot(PCA, ellipse = TRUE, var.axes = FALSE, labels =
codon_usage$sample,
                  groups = factor(codon_usage$condition)) +
theme_classic())
}

codon.usage(codon_usage)
```

```
Pathway analysis with GAGE:
selected_predictors_KOs <-
unique(na.omit(str_extract(unlist(str_split(selected_predictors, "_")),
"K\\d++")))

library(gage)
path.set <- kegg.gsets("ko")
ko.gs <- path.set$kg.sets
gage_dt <- samples[samples$pvals < 0.05, c("category", "enrich",
"sample")]
gage_dt <- gage_dt[gage_dt$category %in% selected_predictors_KOs, ]
gage_dt <- dcast(gage_dt, category ~ sample, value.var = "enrich", fill =
0)
gage_dt <- as.data.frame(gage_dt)
rownames(gage_dt) <- gage_dt$category
gage_dt <- gage_dt[, -1]
```

```
#a function which does enrichment analysis for all pathways
expressivity <- function(gage_run){
  kegg.sig <- sigGeneSet(gage_run)

  upregulated <- data.frame(kegg.sig$greater)
  pathways_up <- rownames(upregulated)
  pathways_up <- unlist(str_split(pathways_up, "ko[0-9]*"))
  pathways_up <- pathways_up[rep(c(FALSE, TRUE), length(pathways_up)/2)]

  downregulated <- data.frame(kegg.sig$less)
  pathways_down <- rownames(downregulated)
  pathways_down <- unlist(str_split(pathways_down, "ko[0-9]*"))
  pathways_down <- pathways_down[rep(c(FALSE, TRUE),
length(pathways_down)/2)]

  enr_pathways <- data.table(pathway = c(pathways_up, pathways_down),
                             enrichment = c(upregulated$stat.mean,
downregulated$stat.mean))
  setorder(enr_pathways, by = enrichment)
  return(enr_pathways)
}

pathwayEnrich <- gage(gage_dt, gsets = ko.gs, ref = 80, compare =
"as.grpup")
enr_pathways <- expressivity(pathwayEnrich)
ggplot(enr_pathways, aes(reorder(pathway, -enrichment), enrichment, fill =
enrichment)) +
  geom_bar(stat = "identity") + coord_flip() + xlab("pathway")
Comparison with other data:
library(readxl)
library(stringr)

metadata <- read_excel("C:\\Users\\Maja\\Desktop\\Eva -
diplomski\\R\\metadata.xlsx",
                       col_types = c("text", "text", "numeric", "numeric",
"text", "text", "text", "numeric", "numeric", "numeric", "numeric",
"numeric", "text", "numeric", "numeric", "text"))
metadata$`Sample ID` <- paste(str_extract(metadata$`Sample ID`, "[HL]D"),
                              str_extract(metadata$`Sample ID`, "\\d+"),
sep = "")

metadata <- metadata[metadata$`Sample ID` %in% sampleSet$samples, ]
colnames(metadata)[colnames(metadata) == "Sample ID"] <- "samples"
metadata <- metadata[, -c(7, 8, 12, 13, 14, 15, 16)]
```

```
data_info <- data.table(samples = rownames(sampleSet))
data_info <- merge(data_info, tablica, by = "samples")


library(ggbiplot)
PCA <- prcomp(sampleSet[, -ncol(sampleSet)], center = TRUE, scale. = TRUE)
ggbiplot(PCA, ellipse = TRUE, var.axes = FALSE, groups =
factor(data_info$condition)) +
  theme_classic()


Kruskal-Wallis test:
kruskal.test(data_info$condition ~ data_info$`TB (umol/L)`)
```