Reproduction of analysis correlating gut microbiota and clinical outcomes in HBV-related hepatocellular carcinoma

Daniel Zalewski

Abstract

This paper is a reproduction of the study conducted by Huang et al. (2020)[1], titled "Integrated analysis of microbiome and host transcriptome reveals correlations between gut microbiota and clinical outcomes in HBV-related hepatocellular carcinoma.". The study aims to find correlations between gut microbiota and clinical outcomes and to develop a predictive model for detecting cases in medical data. The research utilized raw microbiome data from 15 patients, specifically 16S rRNA sequencing, along with 10 tumor samples and 10 non-tumor samples. Statistical analysis was performed on these samples, correlations were calculated, and a RandomForest model was designed. The results demonstrate a dominant presence of Proteobacteria in fecal samples, varied correlations between the examined OTUs and genes in tumor and non-tumor samples, and the most statistically significant genes as biomarkers impacting the prediction of the RandomForest model. More detailed information and code are avaliable via https://github.com/SleepDealler/Modeling-of-complex-biological-systems/Project.

Introduction

Hepatocellular carcinoma (HCC) is a dangerous and commonly occurring liver cancer. In 2020, nearly 900,000 new cases of HCC were reported worldwide, along with 830,000 deaths, making HCC the 6th most frequently diagnosed cancer and the 3rd leading cause of cancer deaths globally[2]. Through very detailed analysis, it has been observed that nearly 78% of global HCC cases are caused by chronic HBV or HCV infection[3]. There are several effective diagnostic methods, such as serological markers, ultrasound, computed tomography, or magnetic resonance imaging, which are commonly used methods, but a significant limitation of these methods is the difficulty in detecting small tumors[4].

In recent years, gut microbiota biomarkers have increasingly been used for the diagnosis of HCC. The method of microbiological biomarkers has great potential for detecting small tumors and diagnosing HCC at an early stage of development[5]. By observing changes in the gut microbiome between healthy individuals and HCC patients, specific microbiological profiles can be identified. Based on these profiles, microorganisms that become markers of the patient's disease state are defined. Moreover, known pathogenic mechanisms of the microbiome, such as modulation of the immune system and induction of inflammatory states, facilitate the definition of biomarkers. The greatest advantage of this diagnostic method is its low invasiveness (fecal sample collection)[6]. Unfortunately, despite the continuous development of the diagnostic method using microbiome biomarkers, there is still an urgent need to search for additional biomarkers[7].

This study is a reproduction of the biomarker search approach proposed by Huang et al. (2020)[1] and at the same time a new perspective on the search for biomarkers aiding in the diagnosis of HCC.

Methods

Data acquisition

The data used in the 16S rRNA analysis come from studies conducted in 2013 on 281 patients, including 150 HCC patients and 131 healthy controls[8]. The sequencing data were deposited and are available in the ENA_EMBL_EBI database under PRJEB8708. In 2019, from the same HCC patient group, 32 patients were selected. RNA-seq sequencing was performed on these samples using Illumina HiSeq 2500 and deposited in the NCBI Gene Expression Omnibus (GSE138485/PRJNA576155) as tumor and non-tumor cells[1].

16S rRNA analysis

In the conducted study, 15 paired sequences in FASTQ format from an available database were used. The 'dada' library was utilized for the 16S rRNA analysis. Data were filtered and trimmed using the filterAndTrim function with the following settings: trimming read lengths (240 nucleotides for forward reads and 160 nucleotides for reverse reads), discarding reads with unknown nucleotides, allowing a maximum of two expected errors for both forward and reverse sequences, and trimming reads with quality scores below 2. These settings ensure the retention of high-quality reads suitable for further analysis. Subsequently, the learnErrors function was used to model errors in the sequence reads, and dereplication was performed using derepFastq to reduce data redundancy. The DADA2 algorithm was then employed to cluster the reads into unique sequences (OTUs). The obtained reads were paired using the mergePairs function, and an OTU table was created and decontaminated from chimeras using the makeSequenceTable and removeBimeraDenovo functions. Taxonomic assignment was performed on the prepared table using the assignTaxonomy function with the SILVA nr 99 v138.1 database[1].

For the prepared data, diversity and taxonomic analyses were conducted using the 'phyloseq' and 'vegan' libraries[9]. Diversity analysis included calculating alpha diversity metrics using the estimate_richness function to assess the evenness and species richness within individual samples. The indices calculated were Shannon, Simpson, and InvSimpson. To assess the similarity between samples and identify clustering patterns, NMDS analysis with the Bray-Curtis distance metric was used to calculate distances between samples based on their species composition. Taxonomic analysis included the Phylum and Genus levels. The analysis used the tax_glom function with data transformation and filtering to enhance data visualization and analysis[1].

RNA-seq analysis

For the RNA-seq analysis, 10 tumor samples and 10 non-tumor samples were used from the NCBI Gene Expression Omnibus database. The data were obtained using the command 'prefetch {index}' in SRA format. Each read was converted to FASTQ format (forward and reverse) using the command 'fastq-dump –split-3 {index}.sra'. The next step was to obtain the human reference genome GRCh38 and build the index for the reference genome using the command 'hisat2-build hg38.fa hg38'. For the prepared index, a series of read mappings was performed using the command 'hisat2 -p 8 -x hg38 -1 {index}1.fastq -2 {index}2.fastq | samtools view -Sb -> output{sample_nr}.bam'. This command allows for direct creation of a BAM file, bypassing the SAM format, which consumes much more computer memory. Then, the reads within the files were sorted based on their starting positions on the reference genome using the command 'samtools sort output{sample_nr}.bam -o output_{sample_nr}_sorted.bam'. The final step in data preparation was to perform RNA-seq read counting for all prepared files using the command 'featureCounts -T 8 -p -a gencode.v30.annotation.gtf -o counts.txt {files}'. Gencode v30 annotation was used for read counting[1].

The data analysis was performed using edgeR library in R and began with the preparation of the count matrix. Next, a DGEList object was created to store the counts and group information. For the prepared DGEList object, normalization factors were calculated using the calcNormFactors function with the TMM method to correct for library differences between samples. Subsequently, a design matrix was created, and gene dispersion was estimated using the estimateDisp function. A linear model was fitted using quasi-likelihood, and differential expression testing was performed separately for tumor and non-tumor samples using the glmQLFit and glmQLFTest functions. The results of the differential expression test were extracted and sorted by FDR value using the topTags function. The analysis was conducted for all genes and separately for significantly differentially expressed genes, with the criteria being FDR < 0.05 and logFC > 0.8[1].

Correlation between 16S rRNA and RNA-seq and finding biomarkers

The calculation of correlations began with synchronizing the data between the tumor samples and the OTU table, and the data was re-normalized using DESeq2. To make the calculations feasible, the number of OTU samples was reduced to 10. Next, Spearman correlation calculations were performed between OTU abundances and differentially expressed genes, and the data was cleaned of NA values. The same steps were taken for the non-tumor samples[1].

The model built was a Random Forest model using the 'caret' library, where normalized RNA-seq counts from the tumor and non-tumor samples were transformed and combined into a single matrix and a vector of labels. The training and test sets were split in a 7:3 ratio due to the small number of input data. To build the Random Forest, 500 trees were used along with the heuristic of the square root of the number of all variables. The model's accuracy was assessed on the test set using predictions and the confusion matrix projection. The importance of variables as biomarkers was calculated using the importance function from the 'randomForest' library[1].

Results

OTUs alpha diversity analysis

The results of the alpha diversity analysis with the Shannon, Simpson, and Inverse Simpson indices are presented in Figure 1. The Shannon index measures species richness and the even distribution of taxa within a single sample. The range of Shannon values can vary between 1 and 3, where lower values indicate greater diversity. The Simpson index measures the probability that two randomly selected individuals belong to the same species, with values ranging from 0 to 1, where higher values indicate greater diversity. The third parameter is Inverse Simpson, which is more sensitive to the dominance of individual species[10]. The alpha diversity analysis of fecal samples shows moderate taxonomic diversity in the samples, with values ranging from 1.5 to 2.5, indicating an even distribution of taxa. The Simpson measure values range from 0.4 to 0.8, indicating a relatively high probability of obtaining the same species when randomly selecting 2 individuals. The Inverse Simpson measure shows a wide range from 2 to 6, which may indicate the presence of potentially dominant taxa in some samples.

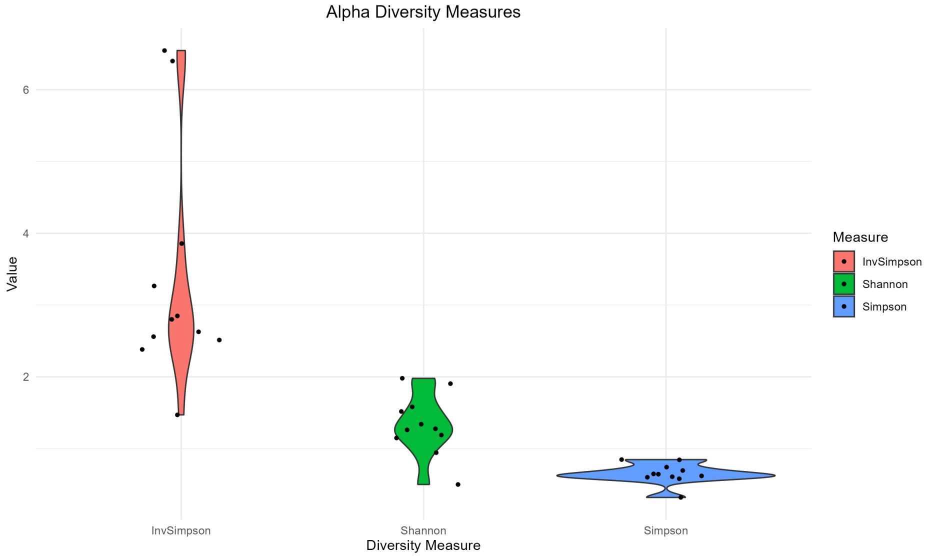

Figure 1: Alpha diversity measurements. Pink represents InvSimpson values for OTU samples, green - Shannon, and blue - Simpson.

OTUs beta diversity analysis

Beta diversity analysis is responsible for demonstrating taxonomic diversity between the examined samples. Figure 2 presents the distribution of points, each representing a sample. The diversity between each group is calculated using the Bray-Curtis distance, which takes into account the presence of taxa and their abundance. Bray-Curtis distance values range from 0 to 1, where higher values indicate greater differences between samples[11]. The points on the plot are not grouped due to insufficient data; however, it can be observed that the points on the plot are relatively dispersed, indicating a high diversity between the examined samples.

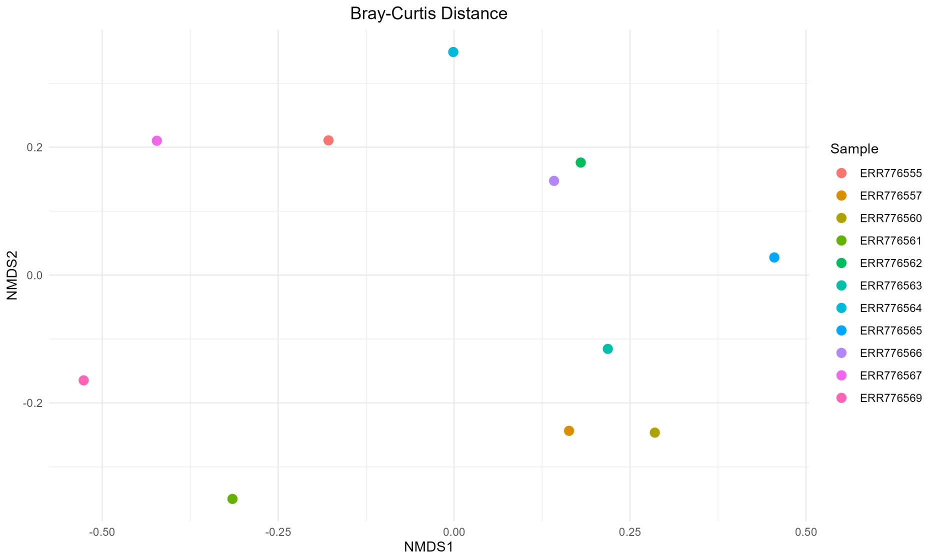

Figure 2: Beta diversity measurement. NMDS plot showing Bray-Curtis distances for OTU samples.

OTUs Phylum level analysis

In OTU analysis at the Phylum level, the composition of the microbiome of individual samples, comparison of content between other samples, and the proportions of Phylum play a key role. Such analysis allows for a more precise observation of similarities and differences between samples and the ability to identify the most frequently occurring Phylum[12]. The results of the OTU analysis at the Phylum level are presented in Figure 3. The X-axis of the plot shows the sample identifiers, while the Y-axis represents the abundance (number of sequences) of each taxon assigned to a given Phylum in a specific sample. The plot clearly shows the dominance of Proteobacteria among the presented Phyla. Proteobacteria is present in most samples and exhibits the highest number of sequences assigned to this Phylum in each of them. Another common Phylum is Bacteroidota, which is also present in most samples, but the number of sequence counts in individual samples is relatively low. In the sample with identifier ERR776565, sequences belonging to Firmicutes were detected, which were not found in the other examined samples.

OTUs Genus level analysis

The last 16S rRNA analysis conducted was the OTU analysis at the Genus level. Genus-level analysis provides the same benefits as Phylum-level analysis but allows for a more precise examination of the organisms present in the sample[13]. The Genus-level analysis assigned minimal numbers of sequences to individual Genus. The results of the analysis are presented in Figure 4. A few samples showed the presence of Bacteroides, while samples ERR776562 and ERR776567 revealed the presence of the Prevotella and Prevotella_9 Genus, respectively.

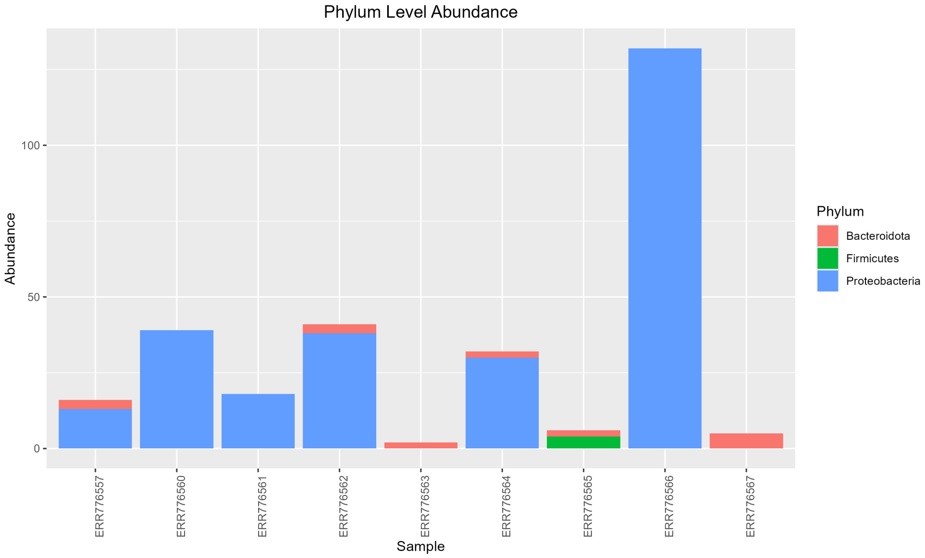

Figure 3: Phylum Level Abundance. Barplot presenting Phylum level abundance for OTU samples.

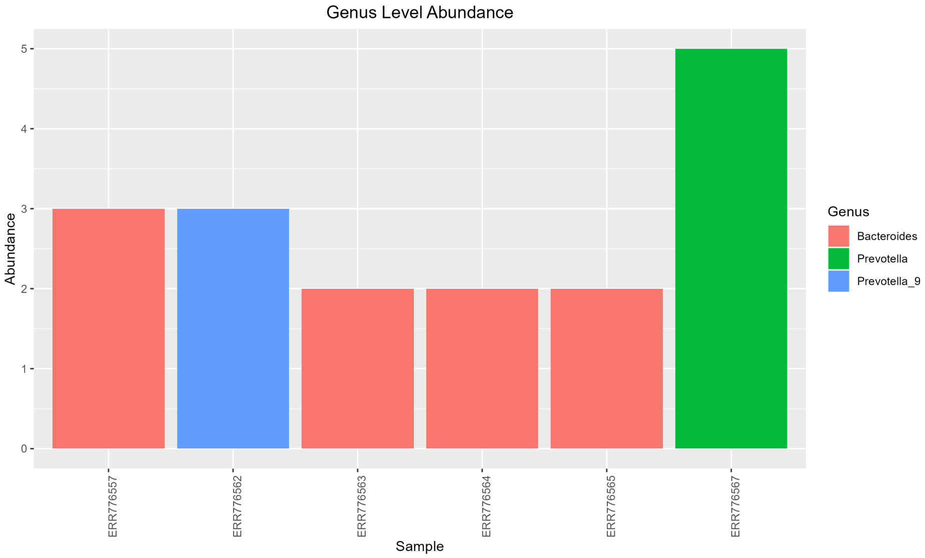

Figure 4: Genus Level Abundance. Barplot presenting Genus level abundance for OTU samples.

16S rRNA analysis - summary

In summary, the alpha and beta diversity analyses revealed moderate within-sample and between-sample diversity. In the beta analysis, due to the small number of data, the samples were not grouped, which may somewhat affect the proper inference. Furthermore, the Phylum-level analysis showed the dominance of Proteobacteria, while the Genus-level analysis indicated a much higher prevalence of Proteobacteria compared to other Genus. Due to the small number of data used in the study and relatively stringent data filtering criteria, the analysis showed gaps in the assignment of specific Phyla or Genus, resulting in a reduction of the already small number of samples that could be analyzed. The minimal number of sequence counts at the Genus level (2-5 sequences) indicates that the results obtained from the analysis cannot be considered certain.

Log2FC distribution for differentially expressed genes analysis

In the RNA-seq analysis, the most important step was identifying significant gene differences (FDR < 0.05 and logFC > 0.8) compared to all differentially expressed genes (DGE). The first measure used in the differential gene expression analysis was the log2FC measure, presented in Figure 5 and Figure 6. Log2FC is used to determine the increase or decrease in gene expression in the examined samples compared to control samples. Positive values indicate increased expression, while negative values indicate decreased expression. Changes in gene expression are interpreted exponentially, relative to the log2FC values, where the value 2 is exponent[14]. The histogram in Figure 5 shows the distribution of log2FC for all genes. The vast majority of genes did not show any changes in expression. Moreover, the histogram is symmetrical around zero, indicating a relatively even distribution and no dominance of genes with increased or decreased expression. Focusing on the extremes, there are genes with 2-fold and 4-fold decreased or increased expression. The histogram in Figure 6 shows the distribution of log2FC for significant genes. A clear difference is visible with the number of genes defined as significant, as the highest peak among all differentially expressed genes reached a count of 20,000, while in significant genes, this value is close to 80. The total number of significant genes was approximately 1,400.

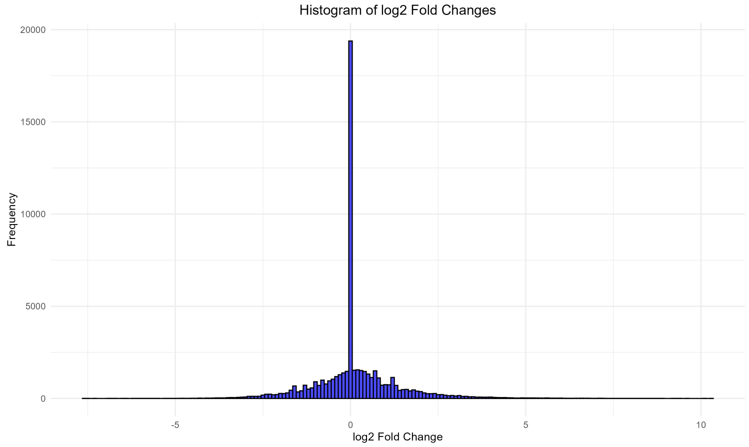

Figure 5: Histogram of log2FC. Histogram presenting log2FC for all genes.

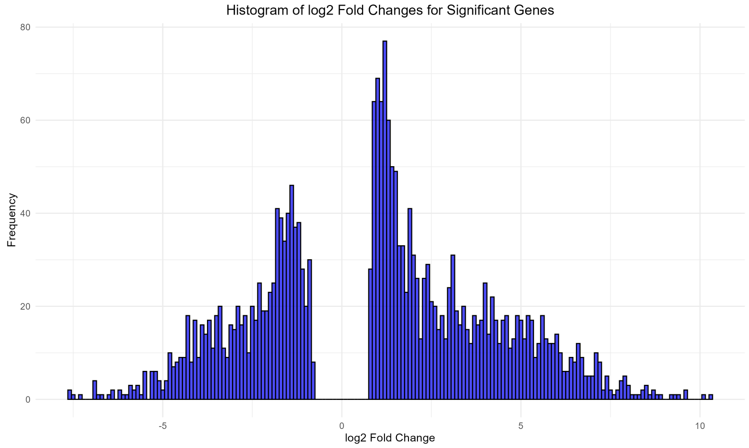

Figure 6: Histogram of log2FC. Histogram presenting log2FC for significant genes.

P-value in relation to log2FC for differentially expressed genes

The next step in the analysis of differentially expressed genes is to compare statistical significance (p-value) with the magnitude of gene expression changes (log2FC), shown in Figure 7. Comparing p-value with log2FC allows for the assessment of the biological and statistical significance of expression changes and the identification of key genes. This comparison enables the selection of genes that show statistically significant changes and biologically significant differences in expression, avoiding the misselection of significant genes[15]. The volcano plot in Figure 7 shows a high density of points (genes) that do not meet the criteria for biological or statistical significance. Points above the red line and outside the area defined by the blue lines represent genes considered significant. Genes that show high values of -log10 adjusted p-value and very low (decreased expression in tumor cells compared to control) or very high (increased expression) values of log2FC may be potential biomarkers.

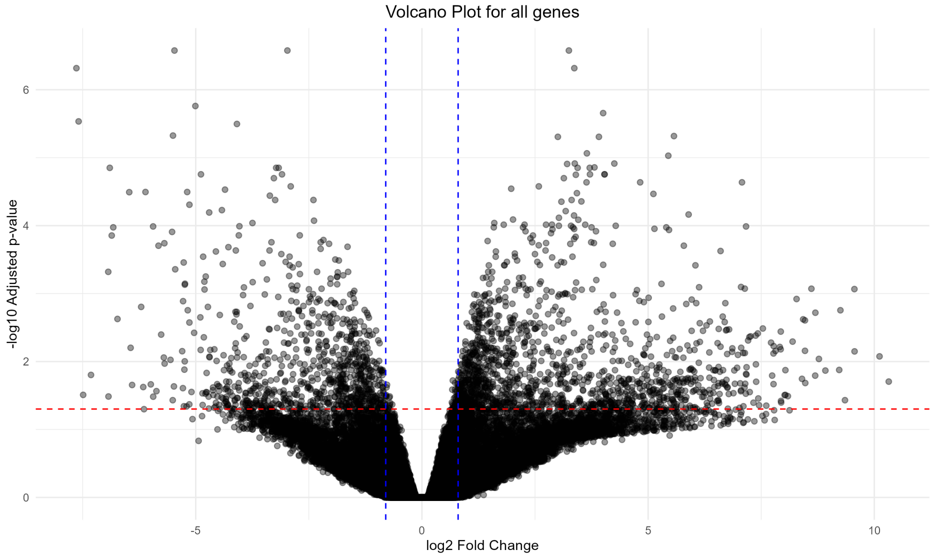

Figure 7: Volcano plot showing log2FC and -log10 adjusted p-value for all genes. Points above the red line and outside the area defined by the blue lines are significant genes.

Gene expression differences between tumor and non-tumor samples

Comparison of gene expression in tumor and non-tumor samples allows for the identification of genes that are differentially expressed between these groups, as presented in Figure 8. Furthermore, it is possible to find biomarkers that aid in the early detection of the studied disease and understanding its mechanisms[16]. The heatmap in Figure 8 shows the expression levels of the top 50 most differentially expressed genes between tumor and non-tumor samples. Each row represents a gene, while the columns represent samples. Through clustering, a clear difference in gene expression between tumor and non-tumor samples is visible. The generated groups separately contain tumor and control samples, indicating different expression patterns between tumor and non-tumor samples and very similar patterns within the groups. Genes located at the very top and bottom of the heatmap suggest that they may be diagnostic biomarkers for the tumor.

RNA-seq analysis - summary

The RNA-seq analysis enabled the identification and filtering of significant genes from differentially expressed genes that showed changes in expression (increase or decrease). Additionally, comparing p-value with log2FC revealed not only the genes that changed expression but also whether this change was statistically significant. Finally, once the significant genes were filtered, it was possible to select potential biomarkers that could aid in cancer diagnosis.

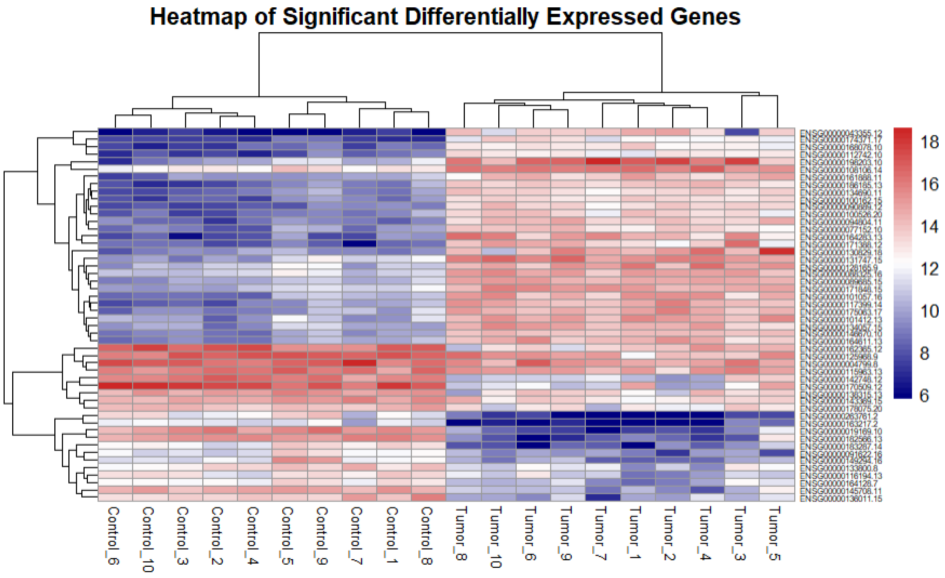

Figure 8: A heatmap displays top 50 genes that were differentially expressed between tumor and control samples. Tumor samples are shown on the right while controls are at the left side of the image. The colors represent the levels of gene expression, with red indicating high expression and blue standing for low expression levels.

OTUs and DGE correlation

The study of correlations between OTUs and DGE enables the detection of patterns of differentially expressed genes in the context of phenotypic traits of the examined samples. These genes, with strong correlations, help in identifying diagnostic and prognostic biomarkers that take into account OTU characteristics. Rows of the heatmap represent individual genes, while columns represent OTU samples. The intensity of the colors indicates the strength of the correlation between a particular gene and an OTU sample, with red indicating a low level of expression and blue indicating a high level[17]. Figure 9 shows Spearman correlation values between the examined OTU samples and DGE. The overall distribution of correlations is presented to detect the diversity of correlations. Correlation values of 1 and -1 indicate that the microbiome influences the expression of these genes, which may suggest mechanisms by which the microbiome can modulate the host's health status. This means that these pairs may be good candidates as diagnostic biomarkers when determining a patient's health status based on their gut microbiome.

Gene importance and selection of biomarkers

The significance of individual variables in the Random Forest model is a crucial step in identifying biomarkers that can be used in cancer diagnostics. This analysis allows for the determination of which genes most contribute to the classification of samples. In this approach, the importance of variables is measured based on how much the model's accuracy decreases when the values of a given variable are randomly permuted. Genes whose permutation causes the greatest decrease in accuracy are considered the most significant[18]. Figure 10 presents the genes that have the greatest significance in the Random Forest model. The genes with the highest importance values are: ENSG00000171848, ENSG00000108106, ENSG00000019169, ENSG00000178075, ENSG00000075218. These are genes encoding the proteins RRM2, UBE2S, MARCO, GRAMD1C, and GTSE1. Changes in the expression of these genes can most accurately classify whether a patient with HCC. These genes are the strongest biomarkers. Other genes (with importance > 0.05) can also be used as relative biomarkers. The genes that did not have sufficient significance to be classified as important biomarkers were: ENSG00000182134, ENSG00000147257, and ENSG00000184999.

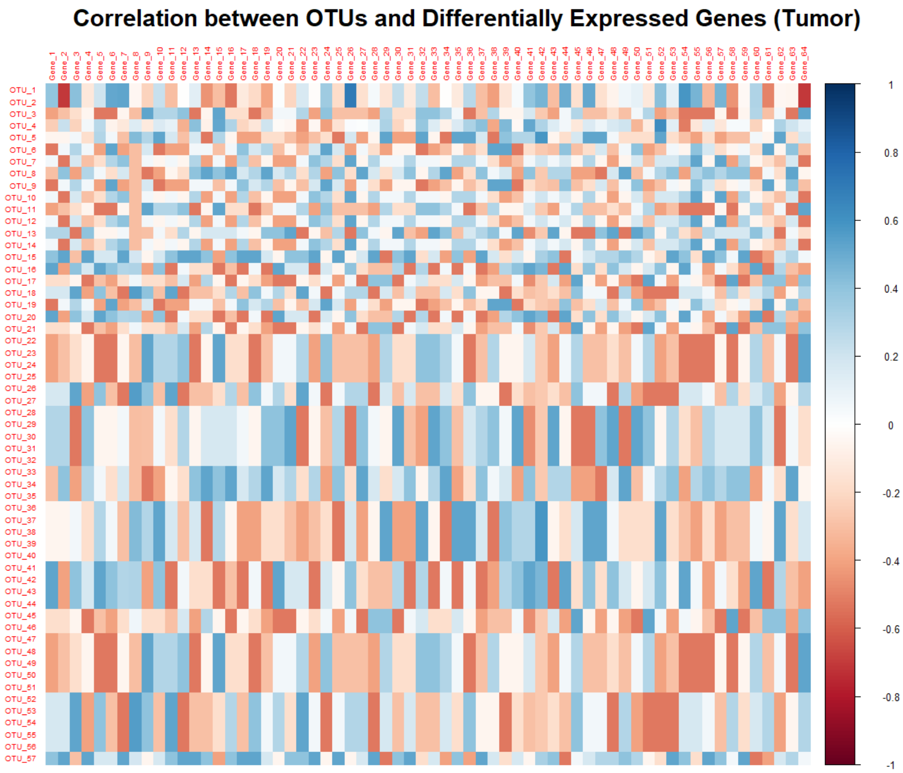

Figure 9: Heatmap showing Spearman correlation between OTUs and differentially expressed genes (tumor). Colors represent the correlation value, with red indicating a negative correlation and blue indicating a positive correlation.

Correlation between OTUs and DGE and selection of the best biomarkers - summary

The analysis of the correlation between OTUs and DGE allowed for the visualization of the distribution of their correlations. Identifying pairs with the highest correlation enabled the definition of genes as biomarkers by demonstrating their significance in classification through the Random Forest model. The best biomarkers turned out to be: ENSG00000171848 (RRM2), ENSG00000108106 (UBE2S), ENSG00000019169 (MARCO), ENSG00000178075 (GRAMD1C), and ENSG00000075218 (GTSE1), which can effectively classify patients as healthy or diseased based on stool samples.

Disscussion

Numerous studies have been conducted to identify biomarkers for classifying patients with HCC based on stool samples. The study by Huang et al. (2020)[1] served as the foundation for this work, which aimed to take a renewed look at the presented topic and expand the existing knowledge.

The taxonomic diversity analysis of stool samples showed moderate alpha and beta diversity, as well as the presence of Bacteroidetes, Firmicutes, and Proteobacteria in the Phylum-level analysis. The study by Huang et al. (2020)[1] demonstrated relatively high alpha and beta diversity in OTU samples, as well as similar Phylum. The main difference is seen in the Genus-level taxonomy analysis, as this study detected only Bacteroides, Prevotella, and Prevotella_9, whereas the number of Genera in the study by Huang et al. (2020)[1] was significantly higher. The possible cause of these differences is the limited availability and smaller amount of data used in this study, which may have resulted in some Genus not being detected.

The DGE analysis study identified nearly 1,400 significant genes. The gene filtration methodology was conducted in accordance with the methodology in the study by Huang et al. (2020)[1] (FDR < 0.05 and logFC > 0.8), in which the

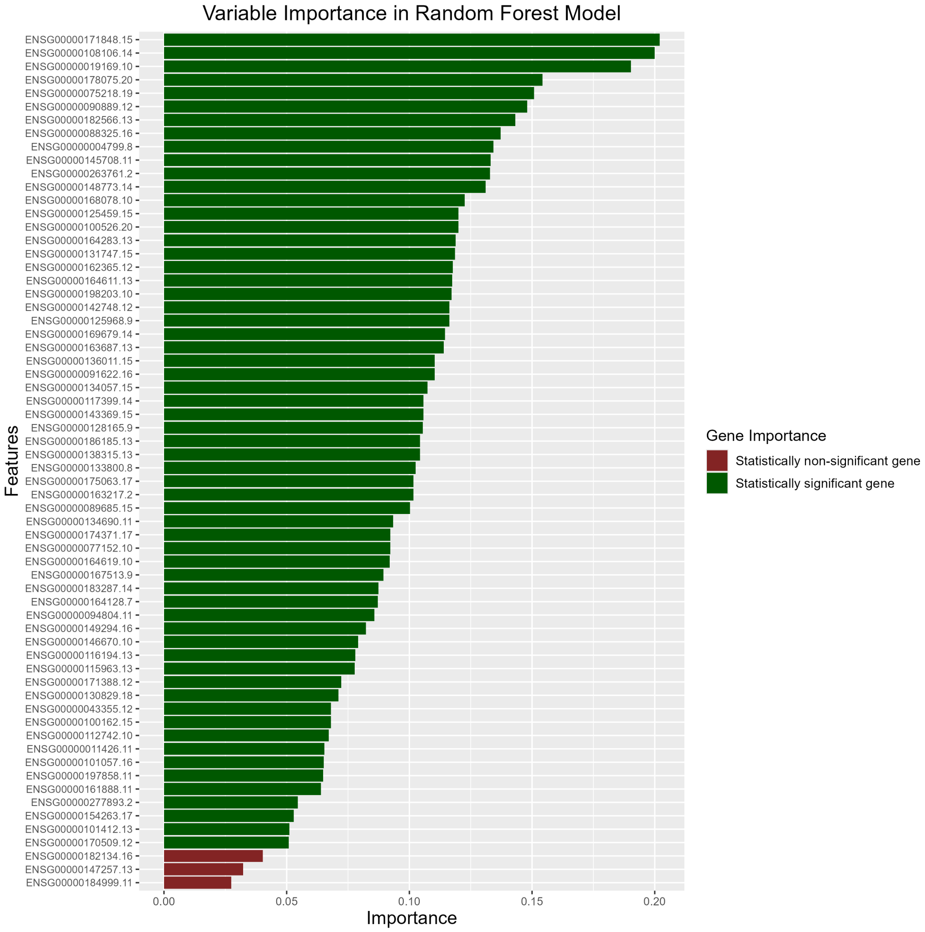

Figure 10: Barplot showing variable importance in the Random Forest model. Genes with statistically significant importance are highlighted in green, while genes with non-significant importance are highlighted in red.

total number of significant genes was 8,101, where many of them showed positive and negative correlations in gene expression. The number of significant genes identified differs substantially from the number of significant genes in the study by Huang et al. (2020)[1]. The reason for these differences may be the smaller number of RNA-seq samples used in this study.

The review study conducted by Li et al. (2022)[7] presented a set of studies conducted to identify microbiological biomarkers in patients with HCC. The study by Huang et al. (2020)[1] reported Bacteroides, Lachnospiracea incertae sedis, and Clostridium XIVa as biomarkers for HCC. Somewhat different results were presented in the study by Ponziani et al. (2019)[19], which identified Bacteroides, Ruminococcaceae, Enterococcus, Phascolarctobacterium, Oscillospira, and Bifidobacterium as biomarkers. In contrast, the biomarkers for HCC identified in the study by Ren et al. (2019)[20] were Bifidobacterium and butyrate-producing bacterial genera. The presented results were obtained by identifying biomarkers as Genus from OTU studies. In this study, the identified biomarkers were protein biomarkers, specifically RRM2, UBE2S, MARCO, GRAMD1C, and GTSE1. This result differs from the protein biomarkers identified in the study by Huang et al. (2020)[1], where the protein biomarkers were CD6 and MAPK10.

Conclusion

The results of the conducted study showed significant differences compared to the study by Huang et al. (2020)[1]. Many of these differences were due to limited availability or filtering of the initial data set or minor methodological changes. This study provided new insights into the search for biomarkers for HCC patients. To improve future studies, it could be beneficial to perform a batch effect analysis to prevent differences caused by processing samples in different batches or experimental conditions. Another possibility is to expand the research by using the constructed model for prediction based on clinical data. Additionally, it would be beneficial to identify microbiological biomarkers based on the correlation between OTUs and DGE.

Additional information

The code described in the methodology and the plots, are available via https://github.com/SleepDealler/Modeling-of-complex-biological-systems/Project.

References

1. Huang, H., Ren, Z., Gao, X. et al. Integrated analysis of microbiome and host transcriptome reveals correlations between gut microbiota and clinical outcomes in HBV-related hepatocellular carcinoma. Genome medicine (2020).
2. Foglia, B., Turato, C. & Cannito, S. Hepatocellular carcinoma: Latest research in pathogenesis, detection and treatment. Int. J. Mol. Sci. (2023).
3. O'Connor, S., Ward, J. W., Watson, M., Momin, B. & Richardson, L. C. Hepatocellular Carcinoma–United States, 2001-2006. MMWR Morb. Mortal. Wkly Rep. (2010).
4. Bialecki, E. S. & Di Bisceglie, A. M. Diagnosis of hepatocellular carcinoma. HPB (2005).
5. Zhang, H., Wu, J., Liu, Y. et al. Identification reproducible microbiota biomarkers for the diagnosis of cirrhosis and hepatocellular carcinoma. AMB Express (2023).
6. Trivedi, Y., Bolgarina, Z., Desai, H. N. et al. The role of gut microbiome in hepatocellular carcinoma: A systematic review. Cureus (2023).
7. Li, K., Liu, J. & Qin, X. Research progress of gut microbiota in hepatocellular carcinoma. J. Clin. Lab. Anal. (2022).
8. Pepe, M. S., Feng, Z., Janes, H., et al. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst. (2008).
9. Oksanen, J., Blanchet, F. G., Kindt, R., et al. Ordination methods, diversity analysis and other functions for community and vegetation ecologists. vegan: community ecology Package (2015).
10. Tucker, C. M., Davies, T. J. Cadotte, M. W., et al. A guide to phylogenetic metrics for conservation, community ecology and macroecology. Biol. Rev. (2017).
11. Bray, J. R. & Curtis, J. T. An ordination of the upland forest communities of southern Wisconsin. Ecol. Mongr. (1957).
12. Ramírez, C. & Romero, J. The microbiome of Seriola Ialandi of wild and aquaculture origin reveals differences in composition and potential function. Front. Microbiol. (2017).
13. Teixeira, C., Prykhodko, O., Alminger, M., Fåk Hållenius, F. & Nyman, M. Barley products of different fiber composition selectively change microbiota composition in rats. Mol. Nutr. Food Res. (2018).
14. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. (2014).
15. Li, W. Volcano plots in analyzing differential expressions with mRNA microarrays. J. Bioinform. Comput. Biol. (2012).

16. Razga, F. & Nemethova, V. Gene expression patterns as predictive biomarkers in hematology-oncology: principal hurdles on the road to the clinic. Haematologica (2017).

17. Losilla, M., Luecke, D. M. & Gallant, J. R. The transcriptional correlates of divergent electric organ discharges in Paramormyrops electric fish. BMC Evol. Biol. (2020).

18. Ram, M., Najafi, A. & Shakeri, M. T. Classification and biomarker genes selection for cancer gene expression data using random forest. Iran. J. Pathol. (2017).

19. Ponziani, F., Bhoori, S., Castelli, C., et al. Hepatocellular carcinoma is associated with gut microbiota profile and inflammation in nonalcoholic fatty liver disease. Hepatology (2019).

20. Ren, Z., Li, A., Jiang, J., et al. Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. Gut (2019).