

AI Driven Analysis of Canola Rhizosphere Microbiome: Diversity, Core Microbiome, and Machine Learning Classification

Zariab Ahmed

zariab389@gmail.com

University of Gujrat

Amina Shoukat Amina Shoukat

University of Gujrat

Saqib Hussain Hadri

University of Gujrat

Abdullah Tariq Abdullah Tariq

University of Gujrat

Research Article

Keywords: Canola Microbiome, Rhizosphere, 16S rRNA, Random Forest, Core Microbiome, Machine Learning

Posted Date: July 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10510709/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Rhizosphere microbiome is an important factor in maintaining plant health and development of sustainable agriculture. In this study, bacterial community composition in the canola (*Brassica napus*) rhizosphere microbiome was analyzed by 16S rRNA amplicon sequencing based on 8,701 samples with 2,663 amplicon sequence variants (ASVs). Analysis of alpha diversity showed that the number of microorganisms per sample equals 61.32 ASVs, while Shannon and Simpson diversity indices are equal to 2.35 and 0.78, correspondingly. Six ASVs were identified as core microbiome components because they were observed in at least 10% of all samples. A microbiome-based health index was developed based on ASV relative abundance and number of ASVs in each sample. A classification model (Random Forest classifier) was built in order to predict two sample groups with different microbiome-based health status. The accuracy of this classifier equals 93.97%, its AUC equals 0.9931, and five-fold cross-validation accuracy equals $93.55 \pm 0.59\%$. Ten ASVs with highest feature importance were identified by feature importance analysis. Thus, the application of machine learning algorithms allowed identifying biologically meaningful patterns in the dataset of the large canola rhizosphere microbiome.

1 INTRODUCTION

Rhizosphere is defined as a small area of soil surrounding the roots of plants.(Berendsen et al., 2012) It consists of diverse microbial communities which contribute to the health, development, and productivity of the plants.(Mendes et al., 2013) Canola is a major oilseed crop on a global scale. It is necessary to study the canola rhizosphere to develop more sustainable agricultural practices.

Advancements in high-throughput sequencing technology allow us to have a much better understanding of the plant microbiomes.(Caporaso et al., 2012) 16S rRNA sequencing is the current standard to evaluate the composition of bacteria.(Callahan et al., 2016) There are many studies performed to study the rhizosphere microbiome of canola.

Machine learning techniques are powerful tools when it comes to analysis of complex microbiome datasets. Random Forest classifiers were successfully used to perform various classification tasks in the field of microbiology.(Breiman, 2001) Machine learning techniques are useful to identify microbial biomarkers.(Pasolli et al., 2016)

Core microbiome contains those microorganisms that are constantly associated with the host.(Shade & Handelsman, 2012) To develop agricultural interventions based on microbiomes, it is necessary to find core microorganisms.(Lundberg et al., 2012)

The goal of this study is to characterize the canola rhizosphere microbiome using a huge dataset (8,701 samples). The specific aims are to assess the diversity patterns, identify the core microbiome, build a Random Forest classifier, and to identify ASVs which affect the performance of the classifier.

2 MATERIALS AND METHODS

2.1 Data Collection

The 16S rRNA amplicon sequencing dataset used in this analysis was obtained from a freely available canola (*Brassica napus*) rhizosphere microbiome database available at the Zenodo website. The dataset had a total number of 8,701 samples with 2,663 ASVs. The abundance table of ASVs was used as the main dataset for all analyses.

2.2 Data Preprocessing

ASV abundance data was then imported to Python for preprocessing. Samples that have zero abundance were dropped from the dataset. In cases where there were missing values, these were filled in with zeros. Species richness was estimated based on the total count of ASVs observed in each sample, while relative abundance was estimated by taking the quotient of each ASV count by its total abundance. Preprocessed abundance data was used for diversity analysis, community composition analysis, core microbiome estimation, and machine learning. (AUC). (McMurdie & Holmes, 2014) (McKinney, 2010)

2.3 Analysis of Diversity

The indices used to calculate alpha diversity were species richness, Shannon diversity index, and Simpson diversity index..(Shannon, 1948) .The species richness was calculated by counting the number of ASVs in each sample, while the Shannon and Simpson diversity indices were calculated based on the relative abundance of ASVs. In order to determine the community composition, the mean relative abundance of each ASV was determined, and the top twenty ASVs were listed. Graphs showing alpha diversity, community composition, and core microbiome were prepared using matplotlib library. (Waskom, 2021)

2.4 Identification of Core Microbiome

The abundance of each ASV was quantified by determining its presence among samples and expressing this information as the percentage of samples that contain the ASV.(Neu et al., 2021) The ASVs that were found in at least 10% of all samples were designated as core ASVs. The core ASVs were ordered based on their abundance values

2.5 Machine Learning Model

Microbiome health index was determined by summing the normalized richness score and the average relative abundance of the top 20 abundant ASVs. The samples were divided into Healthy and Diseased classes according to the median value of the health index. Random Forest classifier was built using the Python library scikit-learn using 80 percent of samples for training and the rest 20 percent for validation of the model.(Pedregosa et al., 2011) The performance of the classifier was estimated through classification accuracy and area under receiver operating characteristic curve (AUC).(Fawcett, 2006) Hyperparameters optimization was conducted via 5-fold cross validation.(Bergstra & Bengio, 2012) Feature importance was analyzed.

2.6 Statistical Analysis

All analysis was carried out in Python.(McKinney, 2010) Data manipulation and preprocessing were performed by using the pandas and NumPy libraries, and machine learning analyses were performed using scikit-learn. Visualization of the results was performed using the matplotlib library. (Waskom, 2021)

3 RESULTS

3.1 Dataset Description

A dataset that had 8,701 samples containing 2,663 ASVs was available. The samples included the rhizosphere microbiomes of canola plants.

3.2 Alpha Diversity Analysis

The alpha diversity analysis gave information on the distribution of richness of ASVs in samples. The mean richness in samples was 61.32 ASVs. The mean Shannon diversity index was 2.35 and the Simpson diversity index was 0.78, Which indicates variations in canola rhizosphere samples

3.3 Core Microbiome

The core microbiome consisted of 6 ASVs found in more than 10% of the samples. The occurrence frequency of these ASVs was ranging from 10.34% to 12.15%, showing that only a very small number of bacterial taxa that consistently detected across the canola rhizosphere microbiome.

Table 1: Core microbiome ASVs with prevalence Ranges 10%-12.15%

ASV	Prevalence (%)
CHCK2CS1311NC02001M	12.15
ORIG2CS1311NC02000M	12.09
ORIG1CS1308NW03000a	11.14
ORIG2CS1315NC02000S	10.54
CHCK2CS1114NC02002S	10.45
ORIG2CS1315NC01000S	10.34

3.4 Abundant ASVs

Mean relative abundance of each ASV was determined in all the 8,701 samples to determine the most abundant microorganisms in the canola rhizosphere microbiome. ASVs with the top twenty mean relative abundances were chosen for subsequent analysis. The ASV with the highest mean relative abundance was ORIG1CS1207NW07000a, whose mean relative abundance was 0.009812, followed by ORIG1CS1116NW05000a, with a mean relative abundance of 0.006638 and ORIG1CS1237NW10000a with (0.004671).

3.5 Machine Learning Classification

The Random Forest classifier was able to classify the Healthy and Diseased groups using the health index value derived from the microbiome. This model gave 93.97% accuracy of classification and ROC-AUC value of 0.9931. Through five-fold cross-validation, the model gave a mean accuracy of $93.55 \pm 0.59\%$, which shows high robustness of the model. The confusion matrix revealed that there were 865 Diseased and 771 Healthy samples that were accurately classified.

Table 2
Model performance metrics

Metric	Value
Accuracy	93.97%
ROC-AUC	0.9931
Cross validation Accuracy	93.55 ± 0.59%
Training samples	6960
Testing samples	1741

3.6 Feature Importance

ASVs contributing most to the classification of health condition using feature importance analysis for twenty most informative ASVs contributing to the classification of microbiome derived health condition. The most important function was ASV CHCK1CS1356NW05002a with an importance score of 0.0270, followed by ORIG2CS1204NC01000S (0.0259), CHCK2CS1114NC02200S (0.0255), and ORIG2CS1319NW10000L (0.0191). These ASVs contributed most strongly to the predictive performance of the Random Forest classifier.

Figure6. Top 20 most important ASVs for model classification.

4 DISCUSSION

In this study, we conduct a detailed analysis of the microbiome in the canola (*Brassica napus*) rhizosphere through a 16S rRNA amplicon sequencing data set consisting of 8,701 samples. The microbial richness and diversity identified from the analysis indicate that the canola rhizosphere microbiome is complex and in accordance with earlier observations of diverse bacteria in agricultural soils. (Fierer et al., 2012)(Shannon, 1948)

The presence of six ASVs which were identified in at least 10% of the samples could imply that there are some bacteria present which could be considered to be part of the canola rhizosphere. These bacteria might play an important role in the ecosystem. (Neu et al., 2021)(Schlaeppli & Bulgarelli, 2015)

The research explored the rhizospheric microbiome of canola (*Brassica napus*) using the massive amplicon sequencing data of 16S rRNA. The major results achieved during this research were the calculation of mean alpha diversity, equal to 61.32 ASVs per sample, detection of the core microbiome including six ASVs found in at least 10% of the samples, and classification of the microbiome-derived health status using a Random Forest classifier, with accuracy of 93.97% and ROC-AUC of 0.9931. The results obtained prove the possibility to find patterns related to the health status by means of machine learning algorithms.

The presence of a small number of ASVs that are highly indicative implies that specific microorganisms might act as biomarkers for the structure of the canola microbiome. Even though the ASVs found in this

study were not taxonomically identified, their constant contribution to the model performance shows the correlation between the health index based on the microbiome and the presence of these ASVs. It is important for further research to identify the function and taxonomy of these ASVs by means of reference databases and metagenomics.

Declarations

Author Contribution

A . Analysis and interpretation B. data downloading and revision C. Supervision methodology finalization
D . Revisions and proof reading

Acknowledgement

I acknowledge google colabs for accession to their computer downloading data and running analysis.

Data Availability

I have used data and I have also present on Google drive a well arranged folder also on GitHub repository with code

References

1. Berendsen, R. L., Pieterse, C. M., & Bakker, P. A. (2012). The rhizosphere microbiome and plant health. *Trends in plant science*, 17(8), 478–486.
2. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).
3. Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs*, 27(4), 326–349.
4. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
5. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods*, 13(7), 581–583.
6. Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., & Bauer, M. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal*, 6(8), 1621–1624.
7. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861–874.
8. Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across

- nitrogen gradients. *The ISME journal*, 6(5), 1007–1017.
9. Knights, D., Costello, E. K., & Knight, R. (2011). Supervised classification of human microbiota. *FEMS microbiology reviews*, 35(2), 343–359.
 10. Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., & Rio, T. G. d. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature*, 488(7409), 86–90.
 11. McKinney, W. (2010). Data structures for statistical computing in Python. *scipy*, 445(1), 51–56.
 12. McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4), e1003531.
 13. Mendes, R., Garbeva, P., & Raaijmakers, J. M. (2013). The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS microbiology reviews*, 37(5), 634–663.
 14. Neu, A. T., Allen, E. E., & Roy, K. (2021). Defining and quantifying the core microbiome: challenges and prospects. *Proceedings of the National Academy of Sciences*, 118(51), e2104429118.
 15. Pasolli, E., Truong, D. T., Malik, F., Waldron, L., & Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS computational biology*, 12(7), e1004977.
 16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825–2830.
 17. Schlaeppi, K., & Bulgarelli, D. (2015). The plant microbiome at work. *Molecular Plant-microbe interactions*, 28(3), 212–217.
 18. Shade, A., & Handelsman, J. (2012). Beyond the Venn diagram: the hunt for a core microbiome. *Environmental microbiology*, 14(1), 4–12.
 19. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379–423.
 20. Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of open source software*, 6(60), 3021.

Figures

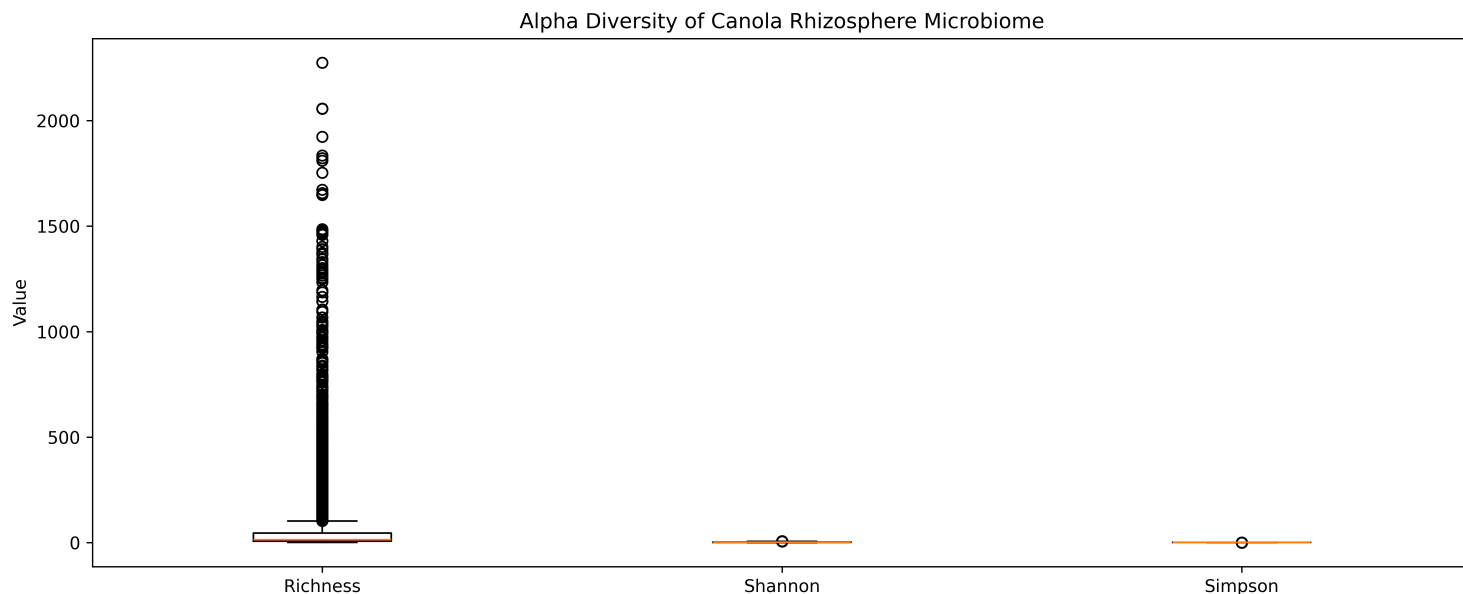


Figure 1

Alpha diversity metrics showing species richness, Shannon diversity, and Simpson diversity across canola rhizosphere samples.

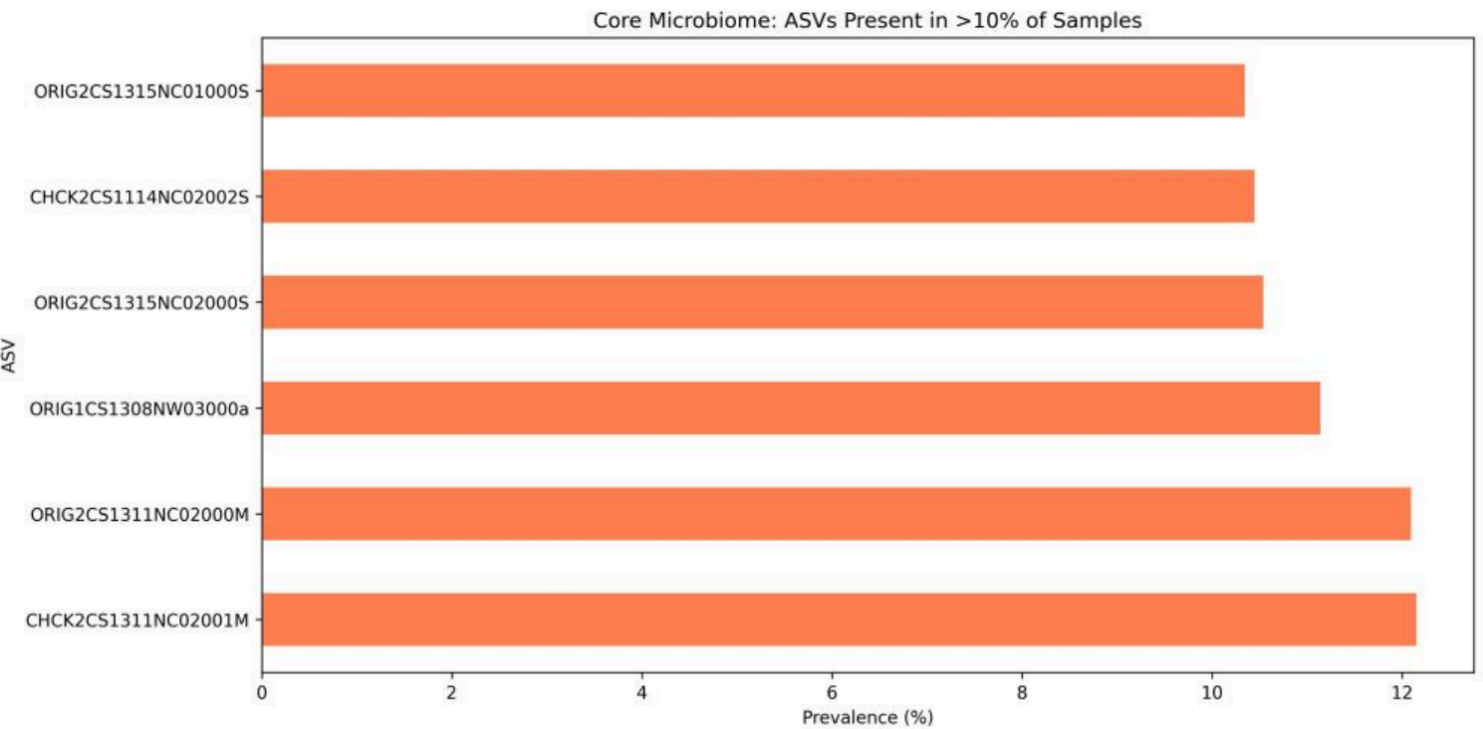


Figure 2

Core microbiome ASVs identified using a 10% prevalence threshold

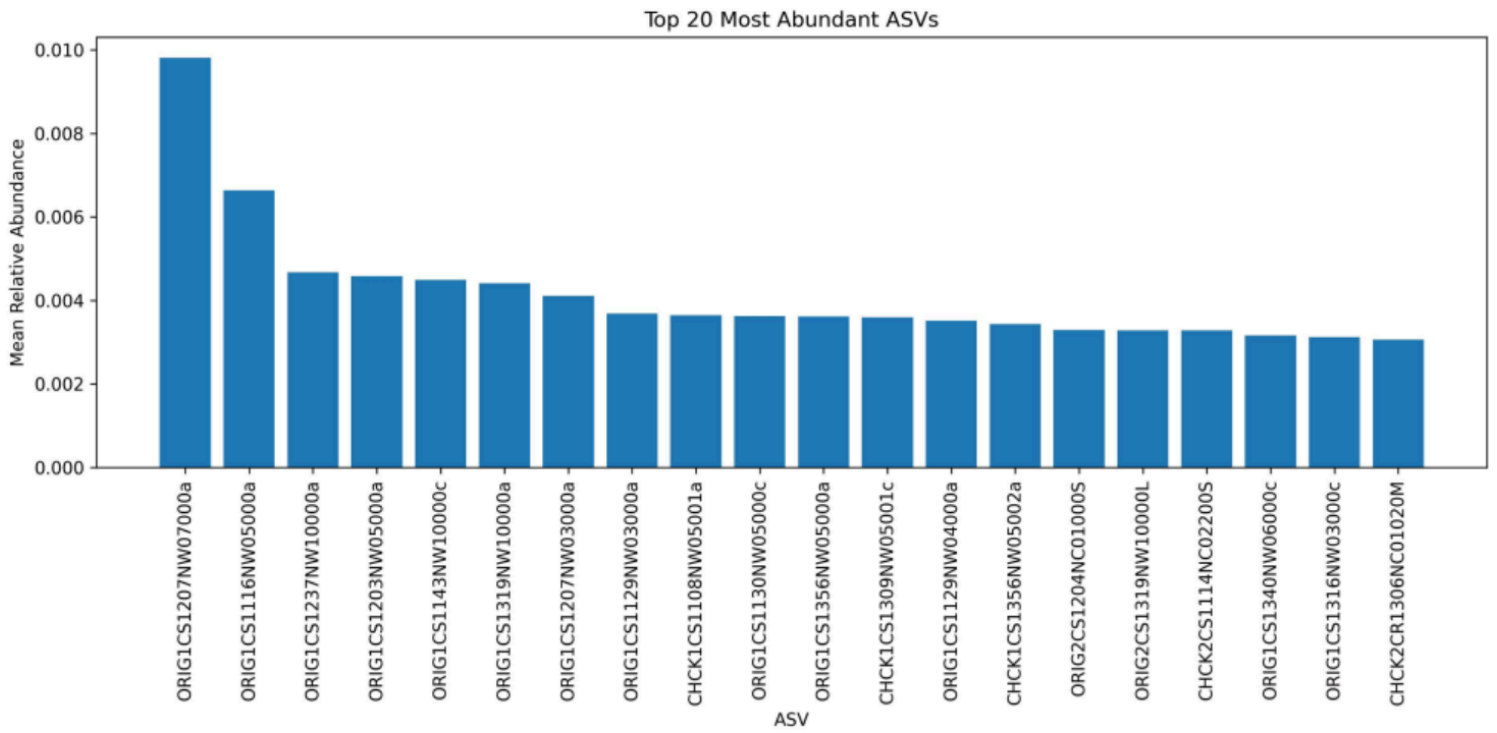


Figure 3

Top 20 most abundant ASVs in canola rhizosphere

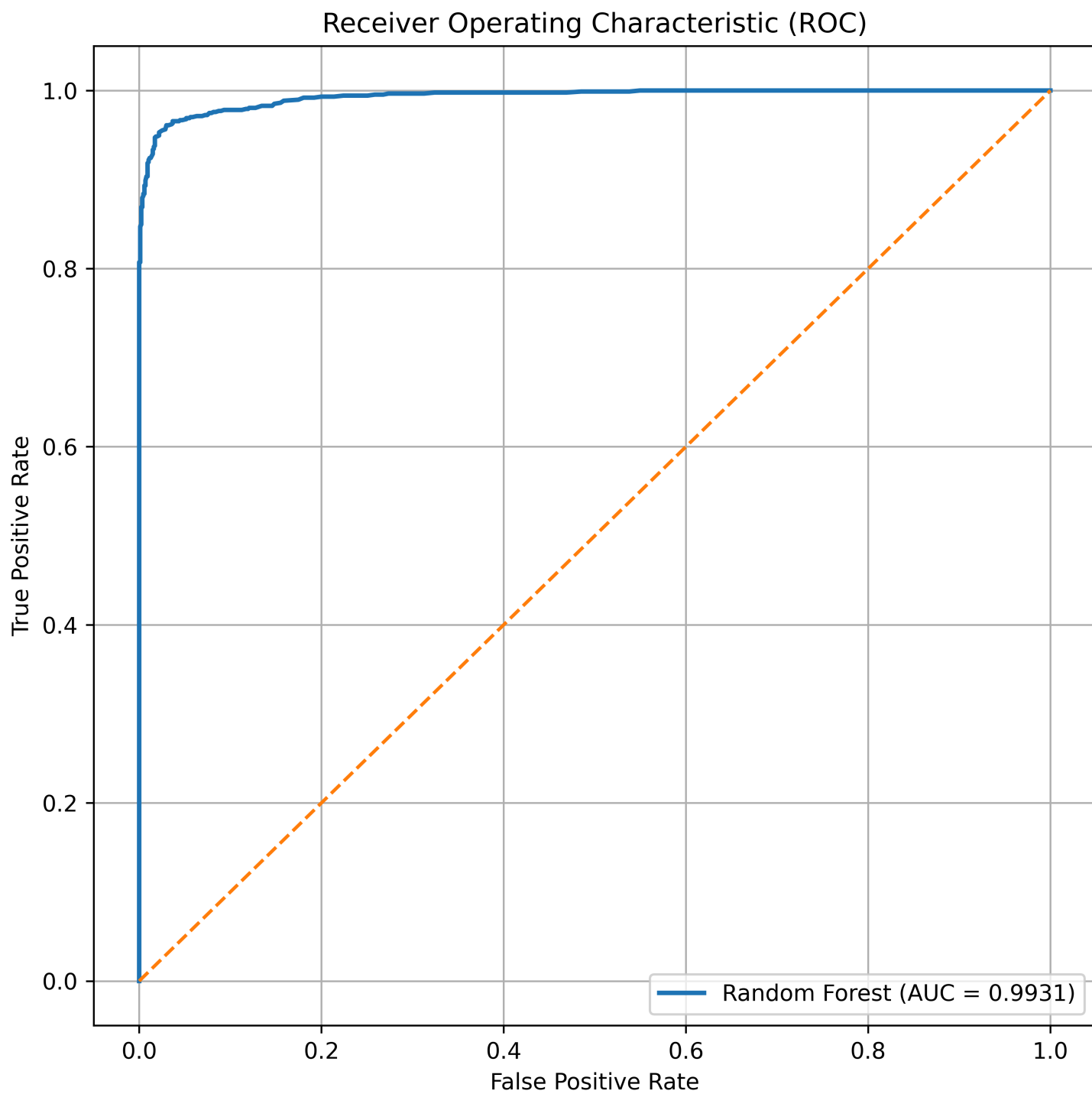


Figure 4

ROC curve showing model performance with AUC = 0.9931

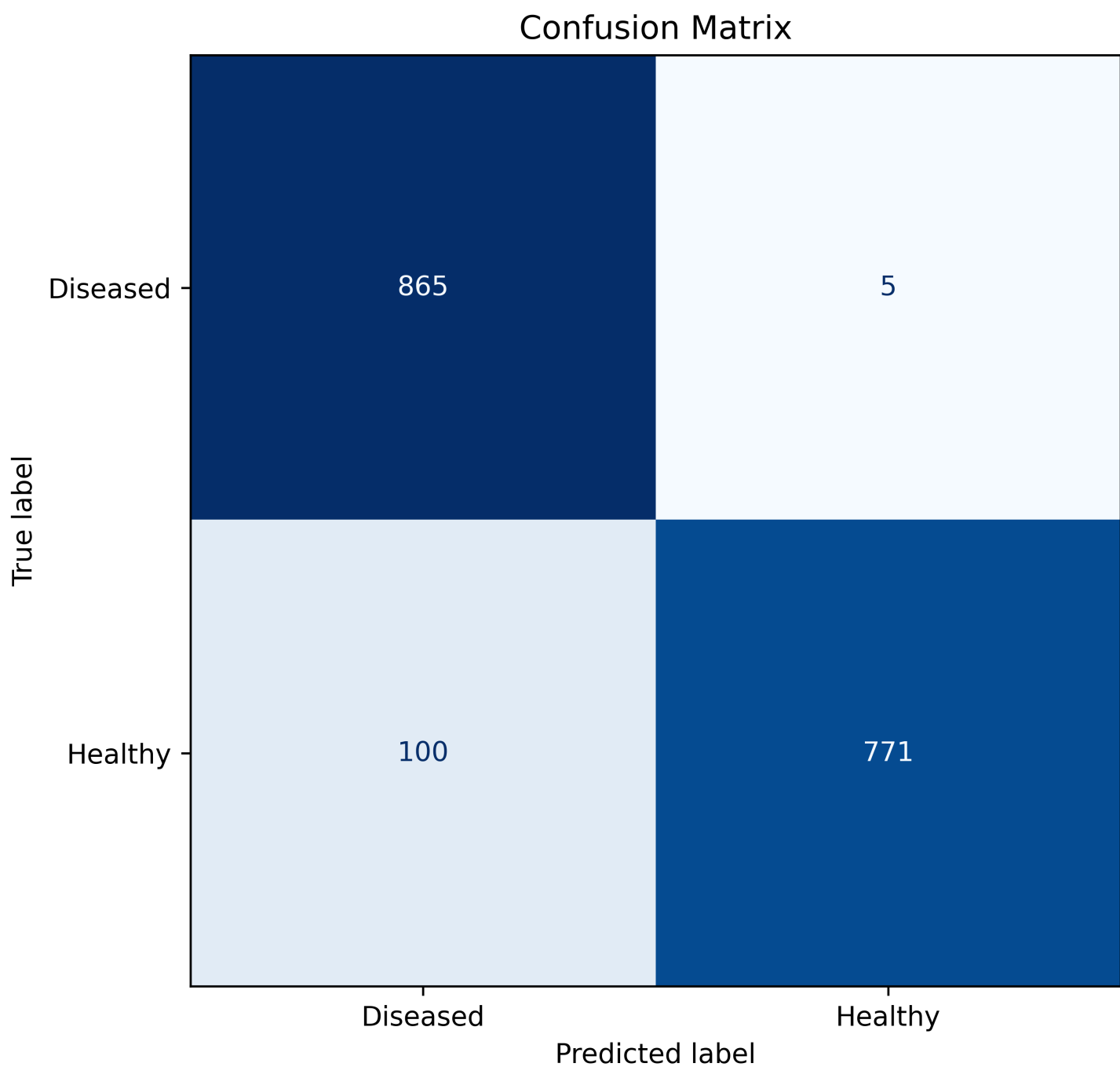


Figure 5

Confusion matrix showing model predictions versus actual labels.

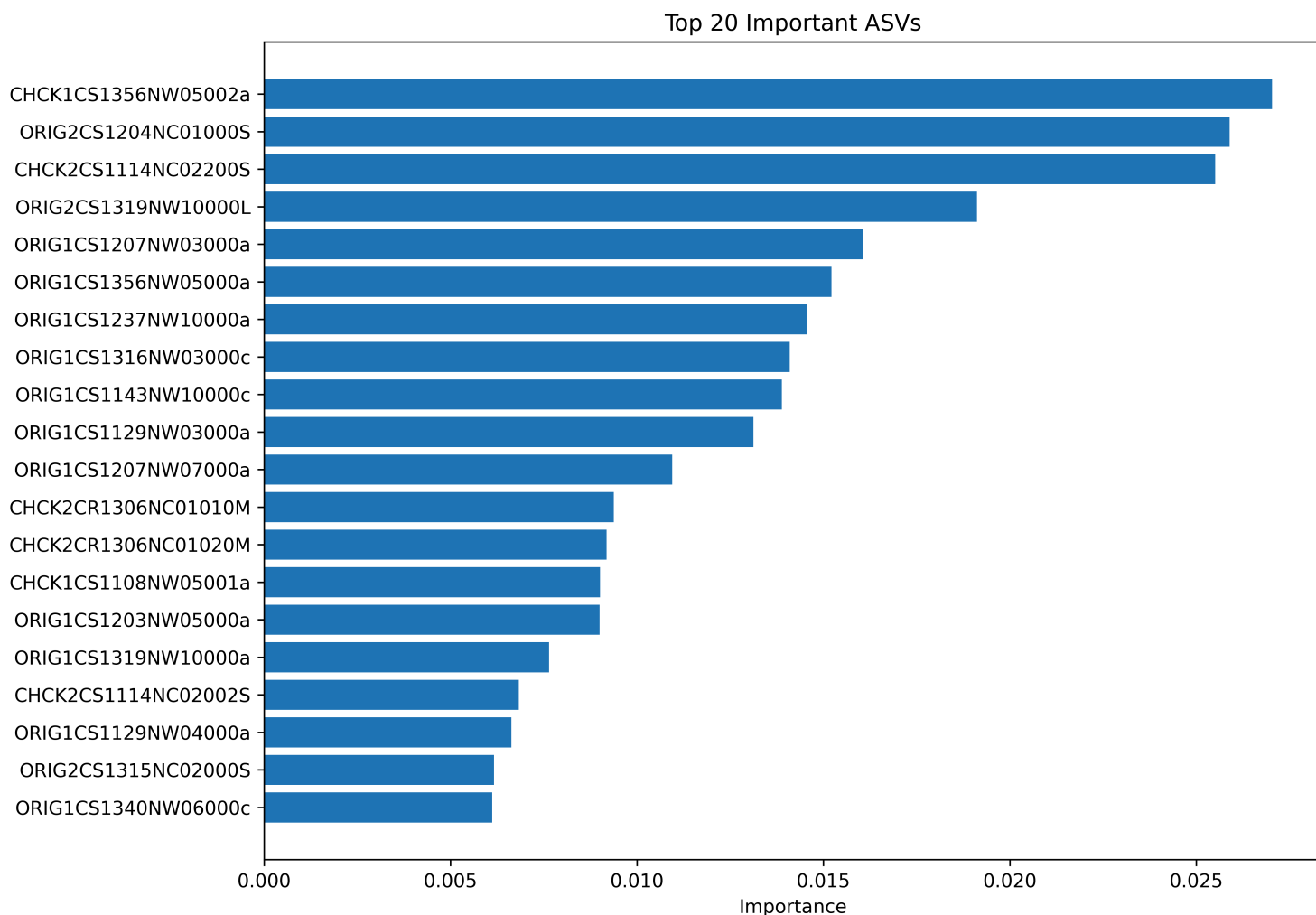


Figure 6

Top 20 most important ASVs for model classification.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ModelPerformance.csv](#)
- [coremicrobiome.csv](#)
- [confusionmatrix.csv](#)
- [featureimportance.csv](#)
- [coremicrobiomeall.csv](#)
- [Top20ASVs1.png](#)
- [communitycomposition.csv](#)
- [samplelabels.csv](#)
- [Paper320260728080220.zip](#)

- [coremicrobiome1.png](#)
- [Figure3CoreMicrobiome1.png](#)