Title: Re-analysis of Gut Microbiota in Untreated Diffuse Large B-Cell Lymphoma Patients Using an Updated Analysis Pipeline

Author: Victoria Palecek

Abstract:

The purpose of this study was to re-analyze publicly available gut microbiome data from patients with diffuse large B-cell lymphoma (DLBCL) and healthy controls to identify whether microbial community diversity/composition differ between groups. Using an updated workflow that identified amplicon sequence variants (ASVs) instead of operational taxonomic units (OTUs), this analysis achieved higher taxonomic resolution than the original study. Alpha and beta diversity measures did not produce statistically significant differences between the control and the DLBCL patient groups. Additionally, analysis of composition of microbiomes (ANCOM), a conservative method that accounts for the composition of microbiome data, identified that no taxa were differentially abundant (Mandal et al., 2015).

In contrast, the original study reported significant results using slightly less conservative techniques such as LEfSe and PLS-DA. The differences in the results of both analysis methods highlight the improvements in metagenomic analytical pipelines in the time since the original study. The future of this work could benefit from larger sample sizes and a standardized analysis pipeline.

Introduction

Diffuse large B-cell lymphoma is the most common subtype of non-Hodgkin lymphoma and is more recently being investigated for its potential links with the gut microbiome (Li et al.,2018). The human GI tract has a diverse microbiome that plays a role in immune modulation and carcinogenesis (Liu et al., 2021). Recent research suggests that some malignancies, including lymphomas, may be associated with differences in the gut microbiota. In a 2021 study by Yuan et al., the composition of gut microbiota in patients with untreated DLBCL was compared to healthy controls. Although no significant differences in alpha diversity were identified, the study identified significant differences in beta diversity and relative abundance between the groups. The phylum Proteobacteria, and particularly the genus *Escherichia-Shigella*, was significantly more abundant in DLBCL patients. The genera *Allisonella*, *Lachnospira*, and *Roseburia* were also significantly more abundant in the DLBCL group.

The original study used QIIME1 (version 1.8.0) to process 16S rRNA gene sequencing data generated from the V3–V4 hypervariable regions. Taxonomy was assigned using the Ribosomal Database Project (RDP) Classifier following clustering into operational taxonomic units (OTUs) at a 97% similarity threshold. The study's statistical analyses were conducted using R version 3.6.0. The dataset included 51 samples, with 26 from DLBCL patients (8 of which had GI involvement) and 25 from healthy individuals with no history of gastrointestinal disease or recent antibiotic use, all from mainland China.

For this current study, a subset of the original data was reanalyzed, specifically the 8 DLBCL samples with GI involvement and 8 randomly selected control samples, using an updated

analytical pipeline. Using QIIME 2 (version 2024.10) (Bolyen et al.,2019), we replaced OTU-based clustering with amplicon sequence variants (ASV), which can provide higher-resolution taxonomic classification. I also used the SILVA 16S database (version 138.2) (Quast et al., 2013) instead of the RDP classifier due to its more frequent updates and more broad sequence coverage from GenBank. My statistical analyses were conducted using the more recent R version 4.5.0 (R Core Team, 2025).

Results from the reanalysis suggest non-significant differences in microbial diversity and composition between the two groups, which contrasts with the original study's results. This difference may be due to differences in the selected subset or the use of the newer analysis pipeline. However, Yuan et al. reported no significant beta diversity differences between DLBCL patients with and without GI involvement, suggesting the change in analysis methods may be the most likely cause. This study highlights how updated pipelines can significantly change metagenomic findings.

Methods

A total of 16 samples were selected from the dataset published by Yuan et al. (2021), consisting of 8 samples from untreated DLBCL patients with GI involvement and 8 randomly selected healthy control samples. The raw sequence data was first imported into QIIME 2 (version 2024.10) for analysis. The first processing steps were demultiplexing and quality filtering. Both QIIME 2's "demux summarize" quality plots and MultiQC reports were then reviewed to assess quality. The original study trimmed reads at 230 bases due to a drop in quality below a Phred score of 20. During re-analysis the dataset maintained quality scores above 20 up to 250 bases. Therefore, both the forward and reverse reads were trimmed at 250 bases prior to denoising. Denoising and chimera removal were performed using the DADA2 plugin (Callahan et al.,2016) in QIIME 2, which produced 1,312 unique ASVs. In order to standardize sequencing depth across samples and retain all 16 samples, all samples were rarefied to a depth of 44,000.

A phylogenetic tree was constructed using the SEPP-based SILVA 128 reference alignment (Janssen et al.,2018) to later perform an unweighted Unifrac analysis. Diversity metrics were generated using "core-metrics-phylogenetic" in QIIME 2, which includes distance matrices (UniFrac and Bray-Curtis) and diversity indices (Shannon and Faith's PD). Additional alpha diversity metrics (Simpson, Observed Diversity, and Chao1) were calculated using the "diversity alpha" plugin. PLS-DA was produced in R for comparison to the original study's results.

Prior to statistical testing, the distribution of alpha diversity data was evaluated in R. A kurtosis value of 2.17 was the result, indicating a non-normal distribution. As a result, all alpha diversity comparisons between groups were conducted using the Kruskal-Wallis test, as it is a non-parametric alternative to ANOVA. Beta diversity differences between groups were assessed using PERMANOVA (permutational multivariate analysis of variance) in QIIME 2.

For taxonomic classification, a custom Naïve Bayes classifier was trained using the SILVA 138 16S rRNA gene database, and the primers used in the original study (338F: ACTCCTACGGGAGGCAGCAG and 806R: GGACTACHVGGGTWTCTAAT). This classifier was used to assign taxonomy to the ASVs. The quantity of unique identifications found at each

level of classification was determined using the tidyverse packages in R (Wickham et al.,2019). Differential abundance analysis was performed at all taxonomic levels using the ANCOM method in QIIME 2.

Results

A total of 1,312 ASVs were identified in this study. Taxonomic classification identified 1 kingdom, 11 phyla, 17 classes, 38 orders, 68 families, 181 genera, and 182 species. In comparison, the original study identified 840 OTUs, along with 12 phyla, 22 classes, 31 orders, 56 families, 201 genera, and 43 species.

Alpha Diversity

Alpha diversity was assessed using the Kruskal-Wallis test for several alpha diversity metrics. No statistically significant differences were found between groups: Chao1 richness ($p = 0.14$), Shannon index ($p = 0.25$), Observed diversity ($p = 0.14$), Faith's Phylogenetic Diversity (PD) ($p = 0.25$), and Simpson index ($p = 0.25$). The original study reported similar results with no significant differences in alpha diversity metrics, with Chao1 $p = 0.43$, Shannon $p = 0.69$, Observed $p = 0.35$, and Faith's PD $p = 0.70$ (Figure 1).

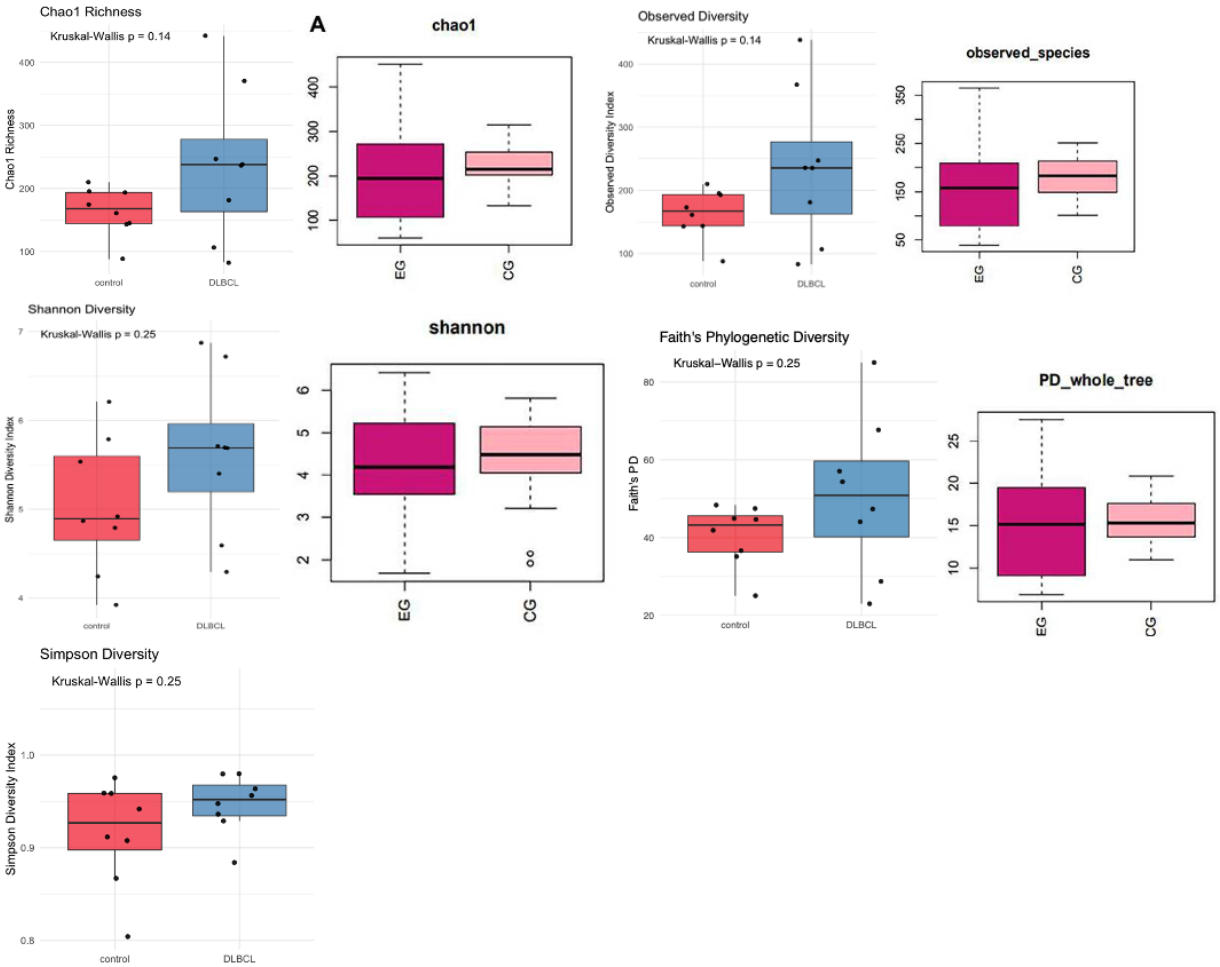

Figure 1: Box diagrams indicating alpha diversity methods for both control (CG) and DLBCL (EG) groups. Results for the current study appear in red and blue while results from the previous study appear in shades of pink.

Beta Diversity

Beta diversity analysis using PERMANOVA also returned results that were non-significant for both Bray-Curtis dissimilarity (p = 0.15)and unweighted UniFrac distance (p = 0.17) (Figure 3). The original study reported a significant result using PLS-DA(permutation test p = 0.002). A re-analysis using the same method produced a permutation test p-value of 0.572 (Figure 2).

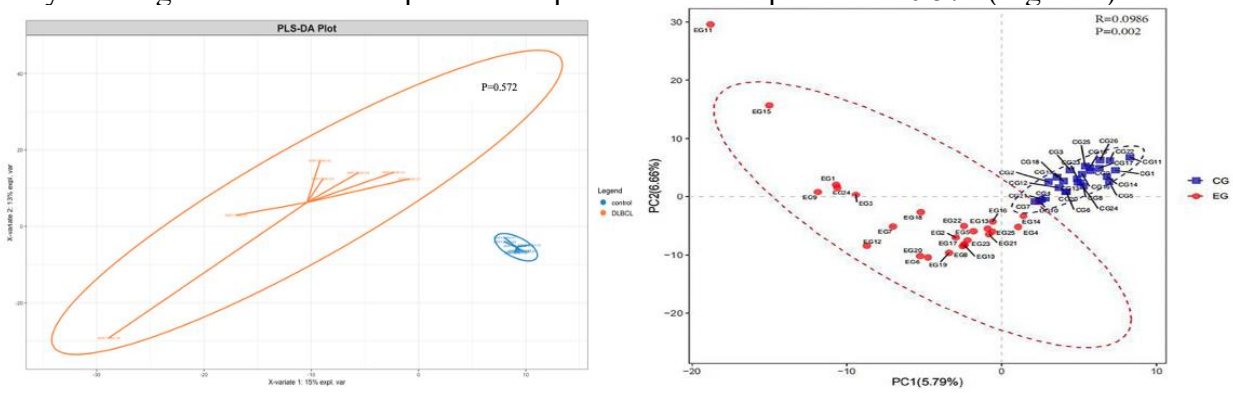

Figure 2: PLS-DA plots for the current study (left) and the original study (right)

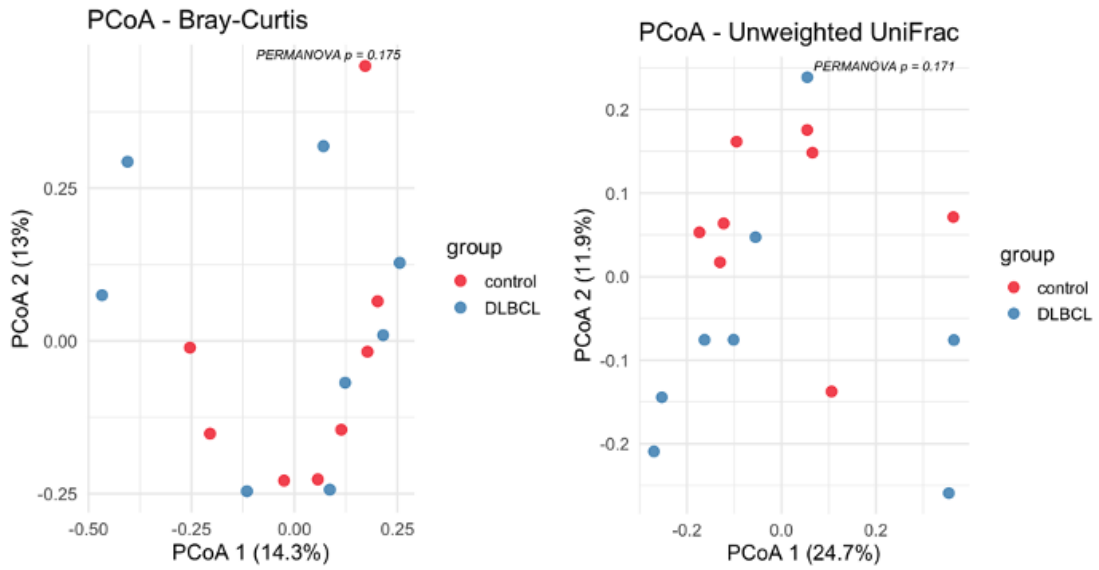

Figure 3: PCoA plots for the Bray-Curtis and Unweighted Unifrac results for the current study.

Differential Abundance

No significantly different taxa were identified at any taxonomic level using ANCOM (Figure 5). The original study used LEfSe and reported multiple differentially abundant taxa between groups. These included higher relative abundance of specific phyla (e.g., Proteobacteria), genera (e.g., *Escherichia-Shigella*, *Enterococcus*), and species (e.g., *Escherichia coli*) in the DLBCL group, while *Roseburia* and *Bacteroides fragilis* were significantly higher in the control group.

The original study also subsetted these significant taxa by classification level and created box plots showing the relative abundance of these significant taxa. Replication of these plots in R using the same taxa returned nearly identical visuals in this analysis, although no taxa were statistically significant based on ANCOM results (Figure 6).

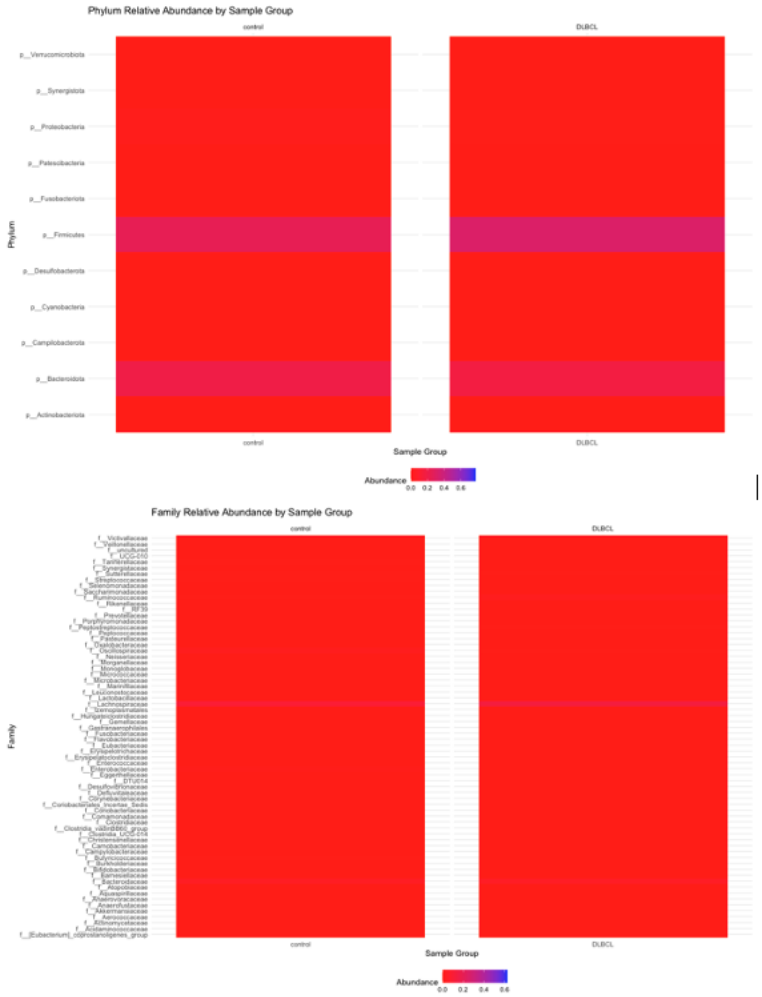

Figure 4: Heatmaps of relative abundance results for phylum (top) and family (bottom) levels of classificaiton for the current study.

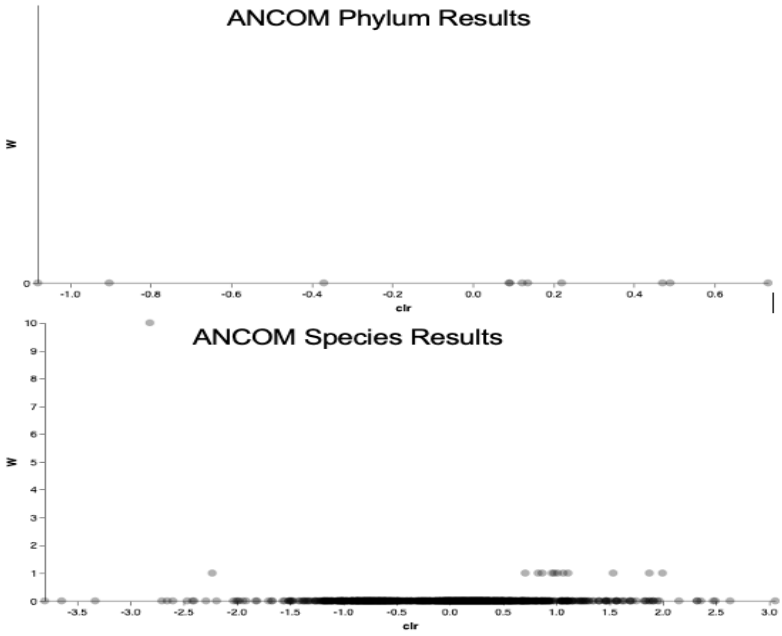

Figure 5: ANCOM results for the phylum (top) and species (bottom) levels of classification.

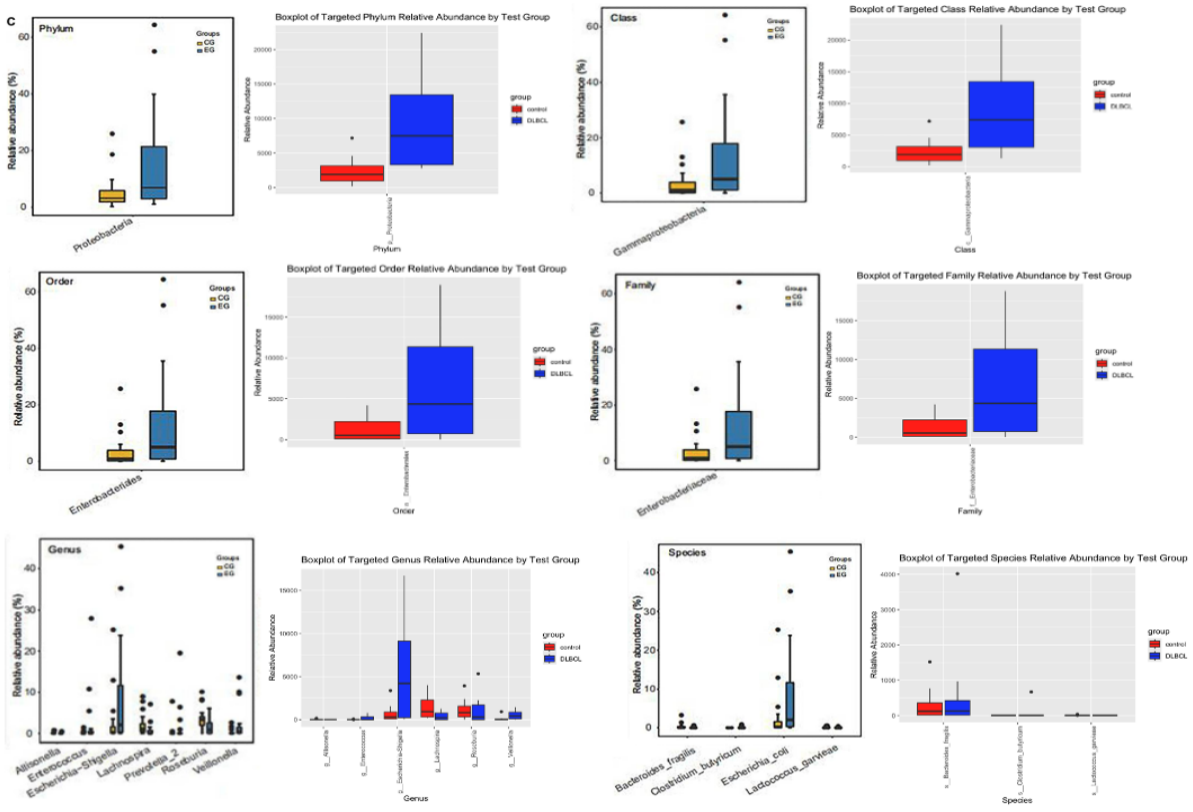

Figure 6: Box diagrams of subsetted taxa from significant LEfSe relative abundance results. The results for the original study appear in blue and yellow while the results for the current study appear un blue and red.

Discussion

This study re-analyzed publicly available microbiome data from DLBCL patients and controls to identify how the use of updated analytic tools may affect interpretations of microbial community structure and composition. During analysis we identified a greater number of features (1,312 ASVs) compared to the original study's 840 OTUs. This, as expected, shows an increase in taxonomic resolution.

Despite the taxonomic resolution being higher, no statistically significant differences in alpha or beta diversity were observed between the DLBCL and control groups using the standard alpha and beta diversity methods. Alpha diversity metrics which include Chao1 richness, Shannon index, observed diversity, Faith's phylogenetic diversity, and Simpson index, did not show a significant difference between test groups. Beta diversity analysis using Bray-Curtis and unweighted UniFrac distances also did not show a significant difference in microbial community composition between groups. These results are consistent with those of the original study, which also reported no significant differences in alpha diversity, although they did report a significant difference between test groups using Partial Least Squares Discriminant Analysis (PLS-DA).

However, it is important to note that PLS-DA is not a true beta diversity test. It uses a supervised method that tries to separate groups based on groupings that you define in advance, which isn't a direct measure of dissimilarity. Because of this, PLS-DA is prone to overfitting, especially in large or complex datasets (Metabolon, n.d.). When PLS-DA was reproduced in R for this study, the permutation test at 999 permutations produced a non-significant result.

There was also a difference in results when differential abundance was analyzed. The original study found many taxa that were significantly different between groups using LEfSe, including the Proteobacteria phylum and the genus *Escherichia-Shigella*. In my repeat analysis using ANCOM, a more conservative method (Mandal et al.,2015), we did not find any significant taxa at any level of classification.

Additionally, the visualizations in the original study, which include box plots of subsets of only the significantly different taxa at different levels of classification, were effective in highlighting test group differences. However, because these plots were subset to include only statistically significant taxa from the LEfSe results, they may give the impression of stronger group separation than is supported by ANCOM. When we replicated these plots using R, the plots were nearly identical to the original study, despite not finding a statistically significant difference between these taxa during re-analysis.

Conclusion

This study re-analyzed gut microbiome data from DLBCL patients and controls to assess microbial diversity and differential abundance using updated analytic methods. A greater number of ASVs were identified compared to the original OTU-based analysis, but no statistically significant differences in alpha diversity, beta diversity, or relative abundance were observed.

These findings differ from those of the original study, which reported statistically significant differences in beta diversity and relative abundance using the methods available for use at the time of the study (2021). The differences in results show how advancements in the field of metagenomic analysis in recent years are significant enough to impact outcomes of older datasets. They also highlight how the choice of analytic method itself can impact results. Future research in the field of DLBC metagenomic study could consider applying a standardized analytic pipeline to facilitate easy meta analysis efforts. Future studies could also consider using larger or more diverse study groups.

References

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37, no. 8 (August 2019): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13, no. 7 (July 2016): 581–83. https://doi.org/10.1038/nmeth.3869.

Janssen, Stefan, Daniel McDonald, Antonio Gonzalez, Jose A. Navas-Molina, Lingjing Jiang, Zhenjiang Zech Xu, Kevin Winker, et al. "Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information." *mSystems* 3, no. 3 (2018): e00021-18. https://doi.org/10.1128/mSystems.00021-18.

Li, Shaoying, Ken H. Young, and L. Jeffrey Medeiros. "Diffuse Large B-Cell Lymphoma." *Pathology*, 50th anniversary review issue, 50, no. 1 (January 1, 2018): 74–87. https://doi.org/10.1016/j.pathol.2017.09.006.

Liu, Xinyi, Yanjie Chen, Si Zhang, and Ling Dong. "Gut Microbiota-Mediated Immunomodulation in Tumor." *Journal of Experimental & Clinical Cancer Research* 40, no. 1 (July 3, 2021): 221. https://doi.org/10.1186/s13046-021-01983-x.

Mandal, Siddhartha, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. "Analysis of Composition of Microbiomes: A Novel Method for Studying Microbial Composition." *Microbial Ecology in Health and Disease* 26 (2015): 27663. https://doi.org/10.3402/mehd.v26.27663.

Metabolon. "Partial Least Squares Discriminant Analysis (PLS-DA)." Accessed April 19, 2025. https://www.metabolon.com/bioinformatics/pls-da/.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41, no. D1 (January 1, 2013): D590–96. https://doi.org/10.1093/nar/gks1219.

"R: The R Project for Statistical Computing," 2025. https://www.r-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. "Welcome to the Tidyverse." *Journal of Open Source Software* 4, no. 43 (November 21, 2019): 1686. https://doi.org/10.21105/joss.01686.

Yuan, Li, Wei Wang, Wei Zhang, Yan Zhang, Chong Wei, Jingnan Li, and Daobin Zhou. “Gut Microbiota in Untreated Diffuse Large B Cell Lymphoma Patients.” *Frontiers in Microbiology* 12 (April 13, 2021): 646361. https://doi.org/10.3389/fmicb.2021.646361.