Identifying Pipelines for *D. Melanogaster* Microbiome Sequence Data (Final Report)

Austen Miller

August 9, 2024

Diverse communities of microorganisms, including bacteria, live within the guts of all animals. This symbiotic relationship is crucial for the health of the host by assisting in digestion and even affecting the brain (Slavin 2013; Cryan and Dinan 2012). The host’s diet can significantly impact these microscopic communities, including altering the microbial species diversity and composition. When these shifts tend towards less diversity, and more pathogenicity, they are considered dysbiotic. The consumption of alcohol affects intestinal microbiota which can exacerbate alcohol associated health problems (Engan 2015) and may even play a role in psychiatric disorders (Hillemacher 2018). Studying how the composition of these communities can be altered in model organisms can provide insights into how they might function within humans.

Full-length 16s rRNA sequencing uses the 16s ribosomal subunit sequence present in all bacteria to analyze the taxonomic composition of bacterial communities, such as the microbiomes of animals. This is possible because ribosomal sequences mutate relatively slowly due to their importance in biological functioning. However, there are 9 known variable regions in the 16s rRNA gene in which that sequence is able to distinguish bacteria at the genus and species level. The similarity of the 16s gene within different bacteria with only a few changes makes it useful as a tool to analyze community composition when sequenced.

In this particular case, we examined the effects of alcohol exposure on the microbiome of the fruit fly species *Drosophila melanogaster*. *D. melanogaster* is useful as a model organism due to its relative ease of work, short generation times, and many genes analogous to humans. The lab of Dr. Imilce Rodriguez is interested in studying how alcohol can affect the gut microbiome using *Drosophila* as a model. Using Nanopore sequencing technology, members in her lab sequenced the 16s rRNA of the bacteria communities present in the guts of young and old flies exposed to 50% ethanol vapor 0, 1 or 2 times.

Oxford Nanopore technology allows scientists to sequence long reads and thus the full-length 16s rRNA gene (~1500 bp). This is in contrast to other technologies, such as Next-Generation Sequencing (NGS) Illumina, where much shorter sequences (~500 bp) are generated. These longer reads allow for increased specificity in taxonomic classification of bacteria since all nine variable regions are sequenced. However, full-length 16S rRNA sequencing generates large amounts of data, which can be complicated to process and require substantial computational resources for analysis. Standardization across different laboratories and studies is essential but challenging to achieve. Nanopore sequencing is advantageous for various reasons, including being cheaper than Illumina, having long reads, which allows us for species-level resolution of bacteria communities, and, despite inaccuracy compared to Illumina, the company is developing new chemistry that has allowed to reduce error rate close to 1% (Zhang et al., n.d.).

Unfortunately, the existing pipelines for full-length 16s sequence data analysis are not always consistently updated. The rapid advance of technology can make the old analysis tools obsolete and difficult to replicate. The goal of my summer project was to identify the best bioinformatic pipeline to analyze Nanopore full-length 16s rRNAseq in our lab to obtain information not only about bacteria abundance and composition but also about alpha and beta diversity. I used the 16s rRNAseq data previously generated by the lab (explained above).

NanoRTax is the main pipeline we were working with. NanoRTax allows “...a nextflow-based pipeline for bacterial taxonomy classification and sample diversity analysis of nanopore full-length 16S rRNA amplicon reads” (Rodríguez-Pérez, Ciuffreda, and Flores 2022). We also used known tools for bacterial taxonomy classification and R for sample diversity analysis. The computational skills gained in the IQBIO program’s Bash and R carpentries have been extremely useful in working with these programs and understanding their results.

During the course of the summer, we were able to make some progress in the use of NanoRTax. With some help from the lab of Dr. Josue Perez at the Comprehensive Cancer Center we were able to install and run some of the dependencies of the pipeline within the Boqueron HPCF, such as Nextflow. Unfortunately, we continued to have some problems with others, such as Kraken, and were not yet able to fully run the pipeline within the HPCF.

In the meanwhile, we have experimental data in the form of classification datasets from Epi2me. This is data from the pipeline provided by Oxford Nanopore, which also uses Nextflow and Kraken (Epi2me Labs, 2024). While not an ideal set, it was enough to be able to perform some analysis.

Using R and R Studio, I was able to interpret and visualize the Epi2me data provided by Oxford Nanopore. Unfortunately, the Epi2me pipeline does not automatically provide diversity analysis and visualizations. To rectify this, I have created an R script using the Tidyverse group of packages (Wickham et al., 2019) and the Vegan package (Oksanen et al., 2024). I used the Dplyer Tidyverse package for grouping, labeling, and modifying the structure of the data sets into community matrices (i.e. rows as barcode samples, columns as species), the required format for diversity calculations in Vegan. Vegan provided useful functions for calculating the alpha and beta diversity of such matrices, and for visualization, I used the Ggplot2 Tidyverse package.

The dataset contained reads with assigned species, but little else. On the suggestion of Dr. Perez, I eliminated singleton reads, as they are likely not accurate data, and created a metadata table with relative abundances at different taxonomic levels. At first, because the dataset only provided species names for many reads, I used the Tidyverse Stringr package to assign the genus by going through the strings of species names. After checking that all names fit Linnean binomial nomenclature (with the exception of some with subspecies), by taking the first part of the names I was able to replicate the Vegan diversity calculations at the genus level. However, I was able to find a better way to assign taxonomy based on species name using the Taxize package (Chamberlain & Szocs, 2013). This way I was able to iterate through the entire list of reads and assign them each taxonomy. This method provides each species entire taxonomy and was more reliable than using species name for genus due to a bug where the genus was being assigned incorrectly on one particular species. A full taxonomy allows for diversity analyses on higher taxonomic levels than species. This allowed me to include relative abundance calculations for every level.

Alpha and beta analyses were performed at both the species and genus levels. Visualizations were created using the Tidyverse Ggplot package, including Shannon alpha diversity index boxplots (figure 1), beta diversity NMDS (figure 2) and PCOA representations, and community composition bar plots (figures 3 & 4). These analyses can be modified relatively easily for taxa at higher levels, or within a particular group.

In the future the ideal direction that this project would take would be to be able to fully run NanoRTax and compare its output with the Epi2me data and verify that it matches. Also useful would be using the Phyloseq package (McMurdie & Holmes, 2013) to be able to run a beta diversity analysis with weighted and unweighted unifrac. Unifrac calculations use the phylogenetic distances between species and so can provide a more nuanced analysis of diversity. I did not have enough time to generate a table with pylogenetic distances, but doing so is possible with our data. I have left a script that can provide a good starting point by including code and instructions to create a phyloseq object with all component tables, except for phylogenetic distance.

Along with the R script including all of my code, I have also created copies of the taxonomy/relative abundance table, community matrix tables, and an HTML document using R markdown including copies of my code, visualizations, and instructions on how to run, or modify the script for different datasets from other experiments. Ideally, this script can continue to be of use to the lab when working with microbiome 16s sequence data when using Epi2me data, or other datasets requiring diversity analyses.



Figure 1. Alpha Diversity Analysis Visualization

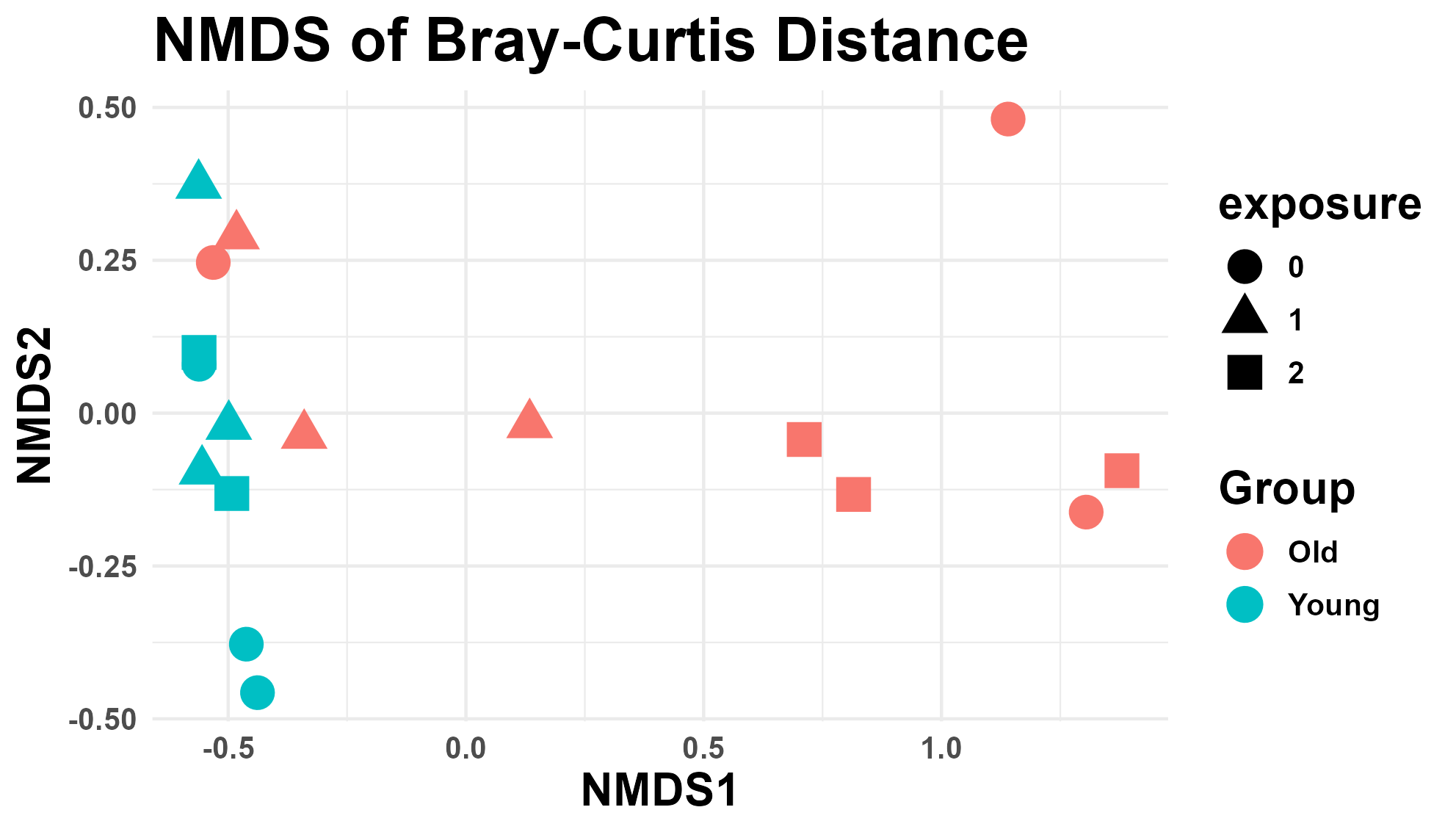


Figure 2. NMDS Beta Diversity Analysis Visualization

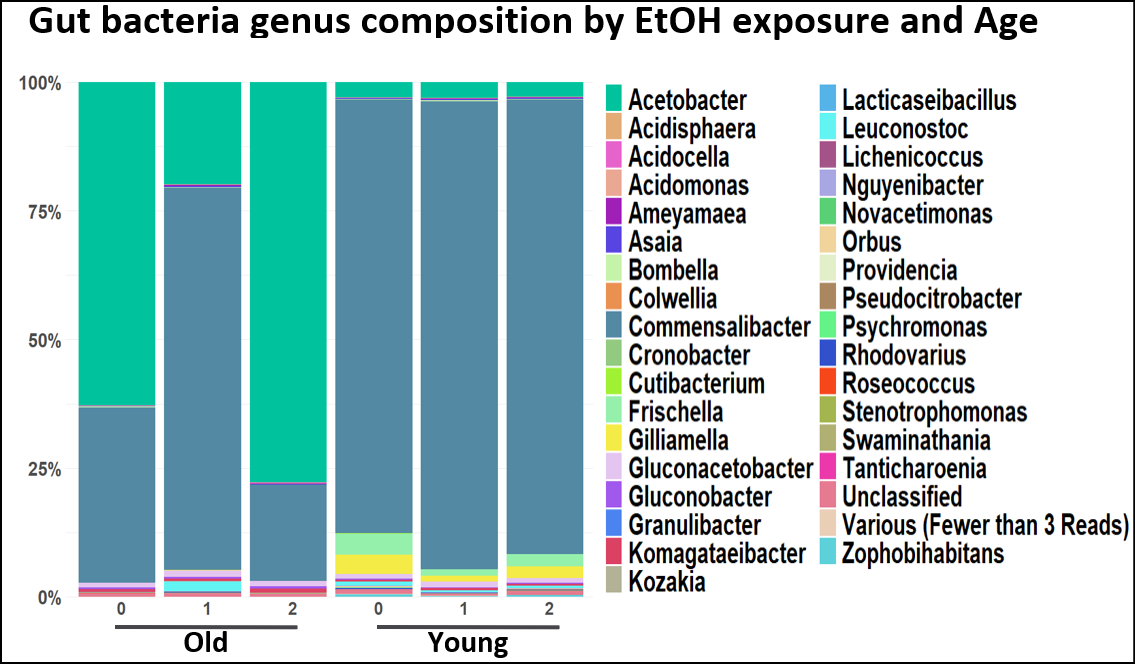


Figure 3. Community Composition (genus) Visualization

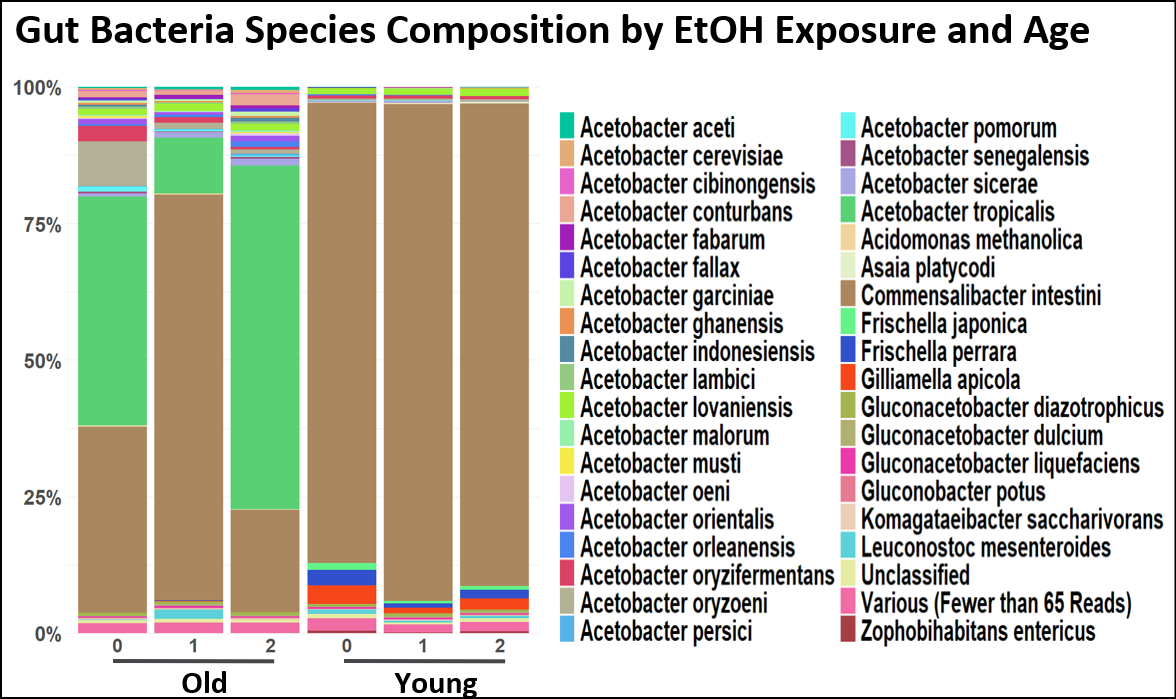


Figure 4. Community Composition (species) Visualization

References

Cryan, John F., and Timothy G. Dinan. 2012. “*Mind-Altering Microorganisms: The Impact of the Gut Microbiota on Brain and Behaviour*.” Nature Reviews. Neuroscience 13 (10): 701–12. <https://doi.org/10.1038/nrn3346>.

Slavin, Joanne. 2013. “*Fiber and Prebiotics: Mechanisms and Health Benefits*.” Nutrients 5 (4): 1417–35. <https://doi.org/10.3390/nu5041417>.

Engen, P. A., Green, S. J., Voigt, R. M., Forsyth, C. B., & Keshavarzian, A. (2015). The Gastrointestinal Microbiome: Alcohol Effects on the Composition of Intestinal Microbiota. *Alcohol research: current reviews*, *37*(2), 223–236.

Hillemacher, T., Bachmann, O., Kahl, K. G., & Frieling, H. (2018). Alcohol, microbiome, and their effect on psychiatric disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *85*, 105–115. <https://doi.org/10.1016/j.pnpbp.2018.04.015>

Delahaye, Clara, and Jacques Nicolas. 2021. “*Sequencing DNA with Nanopores: Troubles and Biases.*” *PLoS ONE* 16 (10): e0257521.<https://doi.org/10.1371/journal.pone.0257521>.

Zhang, Tianyuan, Hanzhou Li, Silin Ma, Jian Cao, Hao Liao, Qiaoyun Huang, and Wenli Chen. n.d. “The Newest Oxford Nanopore R10.4.1 Full-Length 16S rRNA Sequencing Enables the Accurate Resolution of Species-Level Microbial Community Profiling.” *Applied and Environmental Microbiology* 89 (10): e00605-23. <https://doi.org/10.1128/aem.00605-23>.

Rodríguez-Pérez, Héctor, Laura Ciuffreda, and Carlos Flores. 2022. “NanoRTax, a Real-Time Pipeline for Taxonomic and Diversity Analysis of Nanopore 16S rRNA Amplicon Sequencing Data.” *Computational and Structural Biotechnology Journal* 20:5350–54.<https://doi.org/10.1016/j.csbj.2022.09.024>.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open-Source Software, 4*(43), 1686. <https://doi.org/10.21105/joss.01686>

Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Solymos, P., Stevens, M., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M., Lahti, L., McGlinn, D., Ouellette, M., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C., & Weedon, J. (2024). *vegan: Community Ecology Package* (Version 2.6-6.1) [R package]. CRAN. <https://CRAN.R-project.org/package=vegan>

Epi2me Labs. (2024). *Epi2me: Tools for microbial community analysis* (Version x.x) [Software]. GitHub. <https://github.com/epi2me-labs>

Chamberlain, S., & Szocs, E. (2013). taxize - taxonomic search and retrieval in R. *F1000Research, 2*(191). <https://doi.org/10.12688/f1000research.2-191.v2>

McMurdie, P. J., & Holmes, S. (2013). *phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data*. PLoS ONE, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>