A Metagenomic Analysis of Healthy Mice vs. Fatty Liver Disease Induced Mice on Both Control and High Fat Diets

Nicole Ferraro

ECES 490 Final Report

Abstract

Metagenomic analysis is a relatively new bioinformatics tool that can be used to identify potential treatment therapies or diagnostic methods for various pathologies. The microbiomes of 14 murine samples were assessed to determine changes in microbial species abundances between healthy mice and mice injected with DEN to induce fatty liver disease, and within those groups, mice on a control diet vs. a high fat diet. Changes were assessed using a variety of tools and specific species were identified as potentially impactful in the development of fatty liver disease and its symptoms. These species included *Bacteroides dorei*, *Alistipes shahii*, and *Lactobacillus johnsonii*. These species have been identified as having differential abundances between healthy and FLD mice, which might indicate that this change plays a role in the development of symptoms.

Background

Non-alcoholic fatty liver disease (NAFLD) affects anywhere from 10-24% of the population around the world. It is characterized by the accumulation of fat cells in the liver and has been associated with obesity, an increasing problem, which indicates that incidences of NAFLD may also increase. The prevalence of NAFLD in obese patients increases to 57.5-74% of that population[1]. Work published by Roy, et al. suggests findings that indicate manipulation of the gut microbiota as a viable treatment or prevention plan for NAFLD, as well as a method to assess susceptibility prior to disease development[2]. Over 95% of the gut microbiome is composed of three phyla: Firmicutes, Bacteroidetes, and Actinobacteria. However, the species level displays a much higher level of diversity, and that diversity is unique to each individual[2]. Considering the microbiome contains over ten times as many genes as the human genome, it's expected that this diversity will have some effect on the host[3]. Additionally, Machado and Cortez-Pinto also propose the modulation of the gut microbiome as a potential solution for NAFLD. However, neither study mentions specific targets for this modulation, though the later study implicates increased bacterial proliferation in disease development[3]. Finally, a review published by Aron-Wisnewsky, et al. demonstrated that the diversity of bacteria is more important than the composition of bacteria when determining disease-causing factors[4]. These results are consistent with the results presented in this analysis. Aron-Wisnewsky, et al. propose several mechanisms by which the microbiota could affect NAFLD development. These include promoting obesity

through increased food energy yield, regulating gut permeability and immunity, controlling choline and bile acid metabolism, and the regulation of ethanol production by bacteria[4]. Figure 1 demonstrates these relationships.

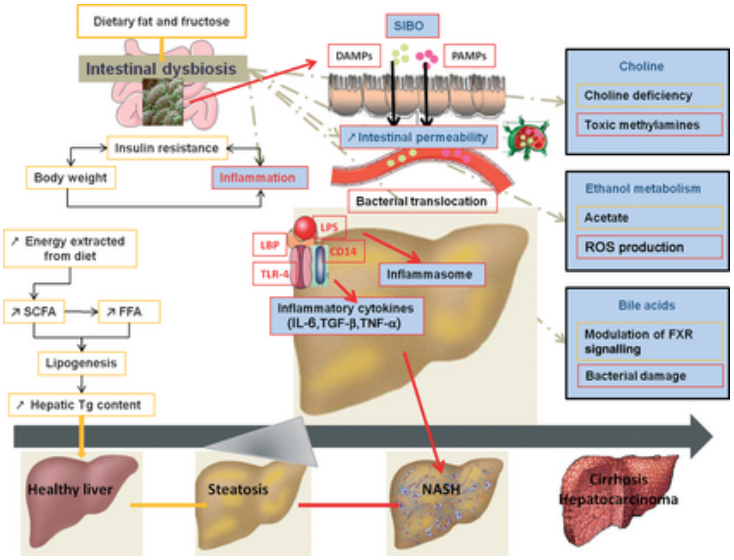


Figure 1. The above image[4] shows the relationships between the gut microbiome and functions that contribute to NAFLD and associated symptoms.

In this study, the microbiome of healthy mice vs. FLD induced mice, as well as mice on control vs. high fat diets were analyzed for species composition and diversity. Mouse models were injected with a DEN toxin that results in the development of fatty liver disease. In order to identify changes that occur in the ecosystem of the body during the development of the disease, changes in the microbiome were assessed. The objective of this analysis was to use the sequenced microbiome from each of the murine samples to assess changes in microbial abundances between the groups. The species that demonstrated significant differences in abundance between groups are then presented as potential therapeutic targets for the treatment and management of NAFLD.

Methods

The overall workflow used in the analysis is described in Figure 2 below.

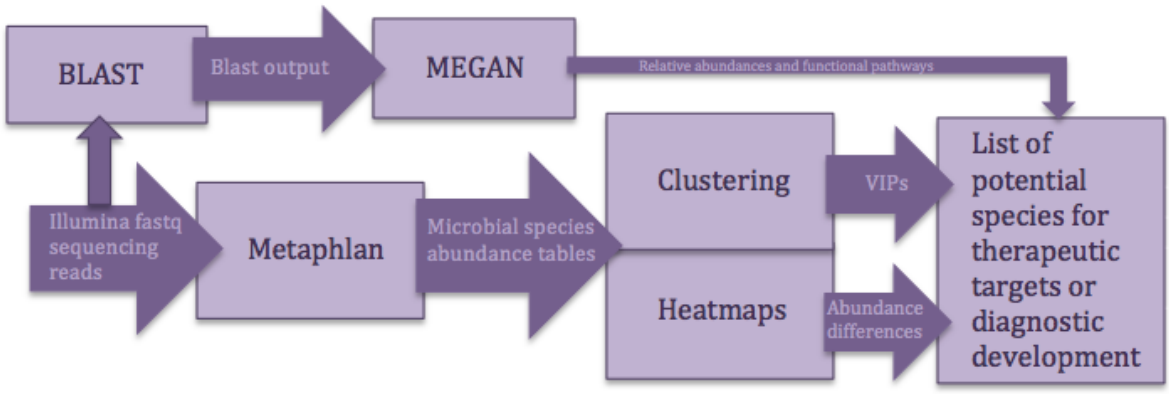


Figure 2. The above describes the input of the fastq files into BLAST and Metaphlan, and then the outputs of those functions were used in downstream analysis, eventually resulting in a target list.

The fastq files were supplied for this project and consisted of 6-8 files for each of the 14 mice samples. A table summarizing the metadata for each sample is shown below.

Sample	Healthy vs. FLD	Control vs. High Fat Diet
GRWS002D	Healthy	Control
GRWS002E[1]	Healthy	Control
GRWS002F	Healthy	Control
GRWS002G[2]	Healthy	High Fat
GRWS002H[3]	Healthy	High Fat
GRWS002I[4]	Healthy	High Fat
GRWS002J[1]	FLD	Control
GRWS002K	FLD	Control
GRWS002L	FLD	Control
GRWS002M	FLD	High Fat
GRWS002N[2]	FLD	High Fat
GRWS002O[3]	FLD	High Fat
GRWS002P[4]	FLD	High Fat
GRWS002Q	FLD	High Fat

Table 1. For each sample, the above table lists whether it is a control or FLD sample, and whether it was fed a control or high fat diet. Samples with subscripts are paired, pre (healthy) and post (FLD) injection.

The first step in the analysis involved the use of a Biobakery tool, MetaPhlAn[5]. This tool allows the user to profile the composition of a microbial community, such as those found in microbiome environments. It takes as input sequencing read files, such as fastq files, and uses the Bowtie2 software to map the reads, and then outputs species abundance tables for each sample. A function can be used to merge the abundance tables into one table for each sample. MetaPhlAn also includes a function that will create heatmaps based on those abundance tables for comparison of species abundance across samples. These microbial species abundance tables can then be used to classify the samples. Two types of clustering methods were used in this analysis: k means, and PLS-DA, or partial least squares discriminant analysis. K-means was chosen when the identity of the samples was unknown, to see if a clear separation between healthy and FLD mice could be identified. PLS-DA was conducted on labeled samples to identify their similarity, and to calculate the variable importance in projection (VIP) for each species considered in the abundance table. The goal of a PLS-DA analysis is to relate a set of response variables to a set of predictor variables, and it is used to maximize the variance between samples, and so it often will reveal differences that may not be seen in a traditional Principle Components Analysis (PCA). The VIP calculation explains the influence of the variable on the response, summed over all components, relative to the total sum of squares of the model[6]. The formula for calculating VIP (though it was done here using an R package) can be seen below in figure 3.

$$VIP_j = \left\{ p \sum_{h=1}^{m} \sum_{k} R^2(y_k, t_h)\, w_{hj}^2 \Big/ \sum_{h=1}^{m} \sum_{k} R^2(y_k, t_h) \right\}^{1/2}$$

Figure 3. Above is the calculation to determine the variable importance in projection for each variable in a PLS-DA analysis. In this case, the variables are the different bacterial species.

Packages available in the statistical programming language R were utilized to perform the clustering analysis. The packages used included mixOmics[7] and RVAideMemoire[8] for PLS-DA and cluster, fpc and ade4 for k-means.

The Illumina fastq files were also used as input to a translated BLAST search, using the blastx command. A translated blast search translates nucleotide sequences and searches them against protein databases to identify hits that could indicate the species that would produce those sequences. The blastx results were visualized using MEGAN5, a metagenome analyzer. This program allows for viewing of BLAST results as well as the incorporation of functional annotation.

Results

The first step in the analysis resulted in the creation of microbial species abundance tables. In each sample, a table was created listing all species found in that sample, and their relative abundance to the rest of the sample. The merged table is too large to include here, but refer to https://github.com/nmferraro5/ECES490_Metagenomics for those results. The input fastq files were also used in a translated blast analysis, the results of which also could not be included here, but can be found at the same link. After the species abundance tables were generated, a heatmap was created through a merged abundance table that included the averages of all files for each sample. This can be seen in figure 4 below.

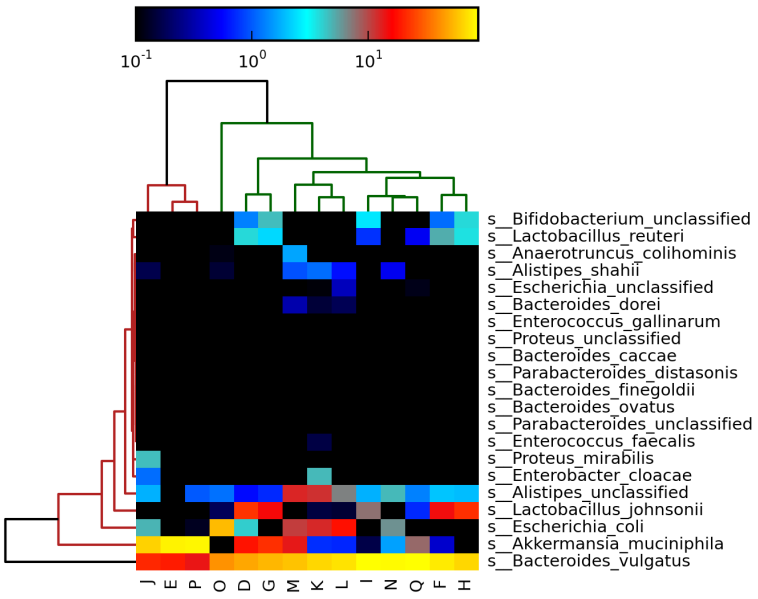


Figure 4. The heatmap to the left summarizes the relative species abundances between samples. Species of interest would be those that show variation in abundance across samples. One sample that shows differentiation in several of the analyses conducted is *Lactobacillus johnsonii*. Additionally, based on this image, two other species showing variation include *E.coli* and *Akkermansia muciniphila.* The function of these variable species will be discussed later in the discussion.

Heatmaps were also generated for each individual sample, including all of the sequencing files for that sample. While the species abundances did not greatly vary between files of the same sample, as expected, this can be used to more closely examine differences between paired samples, in which the same microbiome is analyzed, pre and post-injection by DEN to induce FLD. Below, the heatmaps from sample H and sample O are shown in figure 5. Both of these samples were on a high fat diet, but H is the healthy, pre-injection

sample, while O demonstrates the changes that occurred after the DEN toxin was delivered. One of the same species mentioned above, *Lactobacillus johnsonii*, shows great variation in abundance between these two samples.

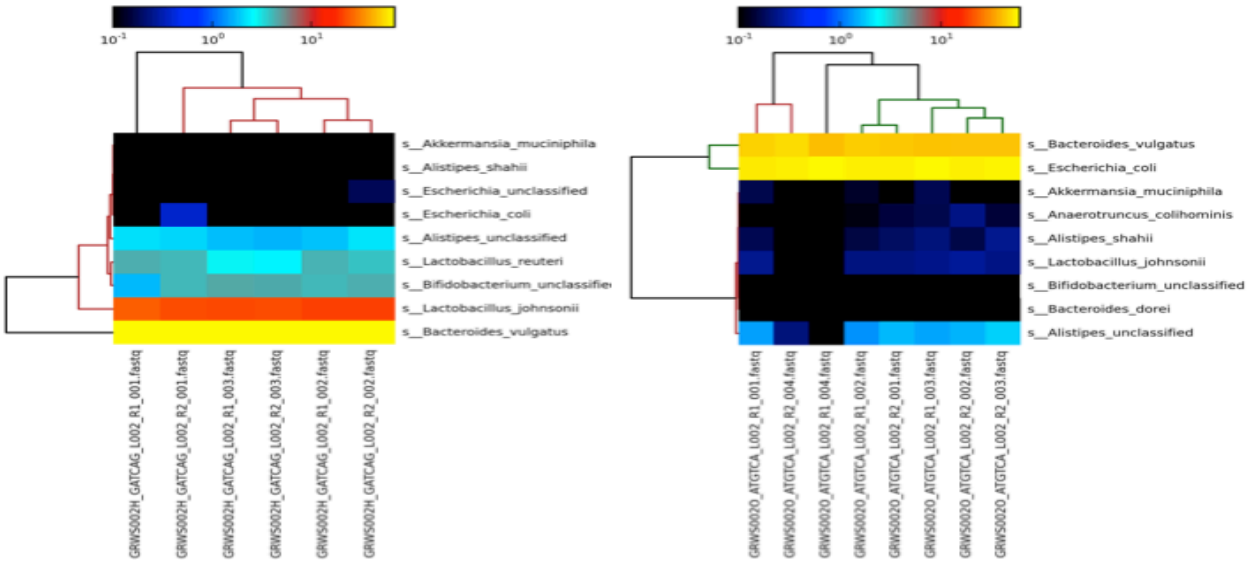


Figure 5. The above heatmaps demonstrate changes in the microbiome pre and post-injection. Notice that the species names are different on the y-axis, but looking at *Lactobacillus johnsonii* specifically in both samples, it shows much higher abundance in the pre-injection sample than in the FLD induced sample. Overall, the post-injection sample displays a decrease in diversity.

The microbial species abundance tables were also used as input for the clustering methods described above. Before labeling the samples, k-means clustering was used, with k=2, to see if a separation occurred between the samples. Those results are shown below in table 2 and figure 6.

Sample	Cluster	Sample	Cluster
GRWS002D	2	GRWS002K	2
GRWS002E	1	GRWS002L	2
GRWS002F	2	GRWS002M	2
GRWS002G	2	GRWS002N	2
GRWS002H	2	GRWS002O	2
GRWS002I	2	GRWS002P	1
GRWS002J	1	GRWS002Q	2

Table 2. Above shows the results of k-means clustering, with k=2. Of the three samples in cluster 1, two are paired (E and J).

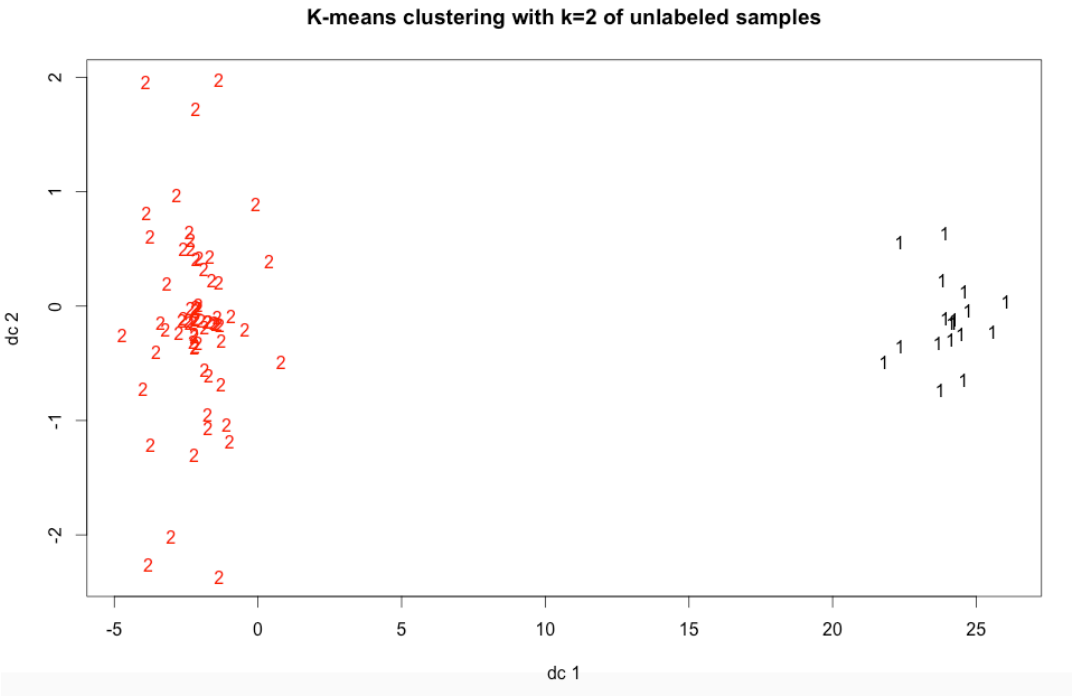


Figure 6. The k-means clustering plot results, showing clear separation between two sample groups. However, this method is not discriminant enough to show differences between known groups.

In addition to k-means clustering, PLS-DA was used to cluster labeled samples, looking at 4 different comparisons. Those results are shown below in figures 7-10.

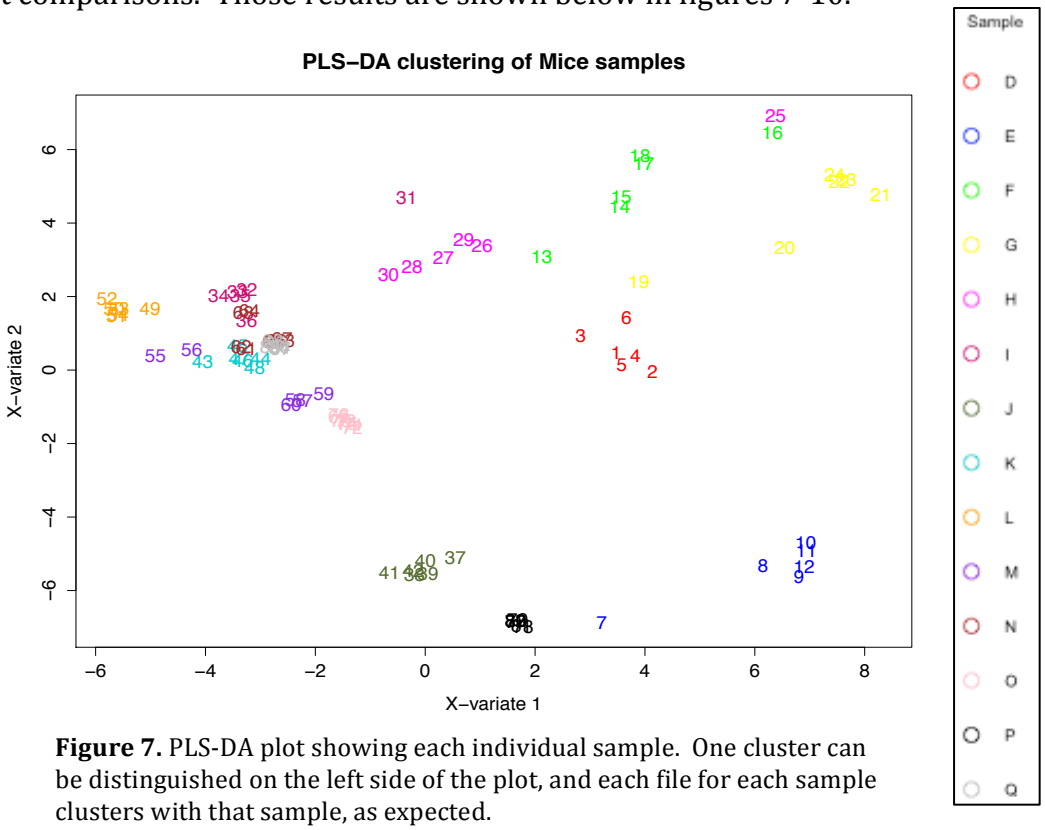


Figure 7. PLS-DA plot showing each individual sample. One cluster can be distinguished on the left side of the plot, and each file for each sample clusters with that sample, as expected.

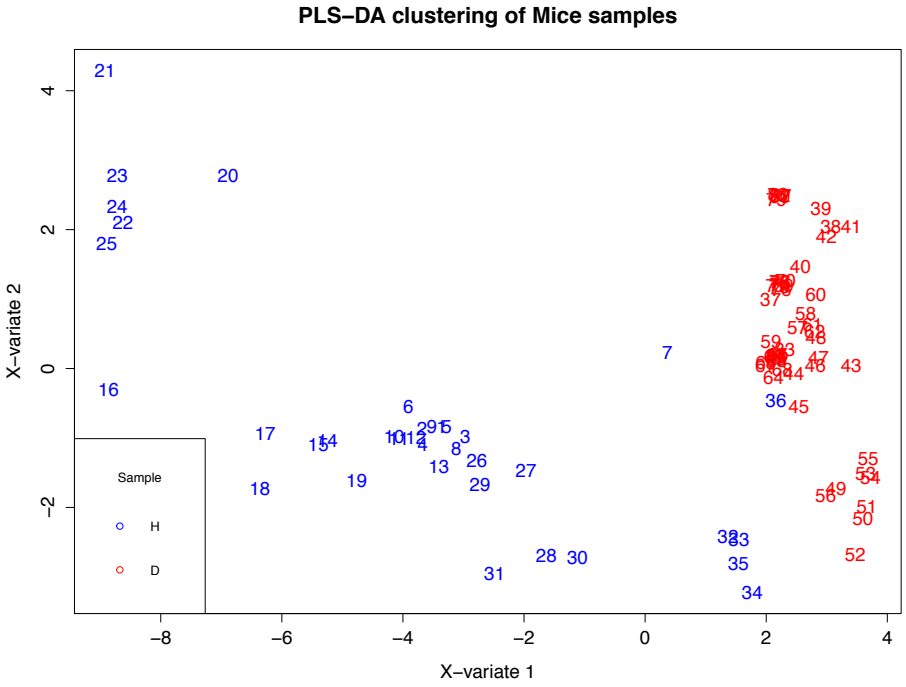


Figure 8. PLS-DA plot showing the separation between healthy and FLD induced samples. Greater variation is seen in the healthy sample group, while the FLD samples cluster more closely together, as expected based on the decrease in diversity in FLD samples.

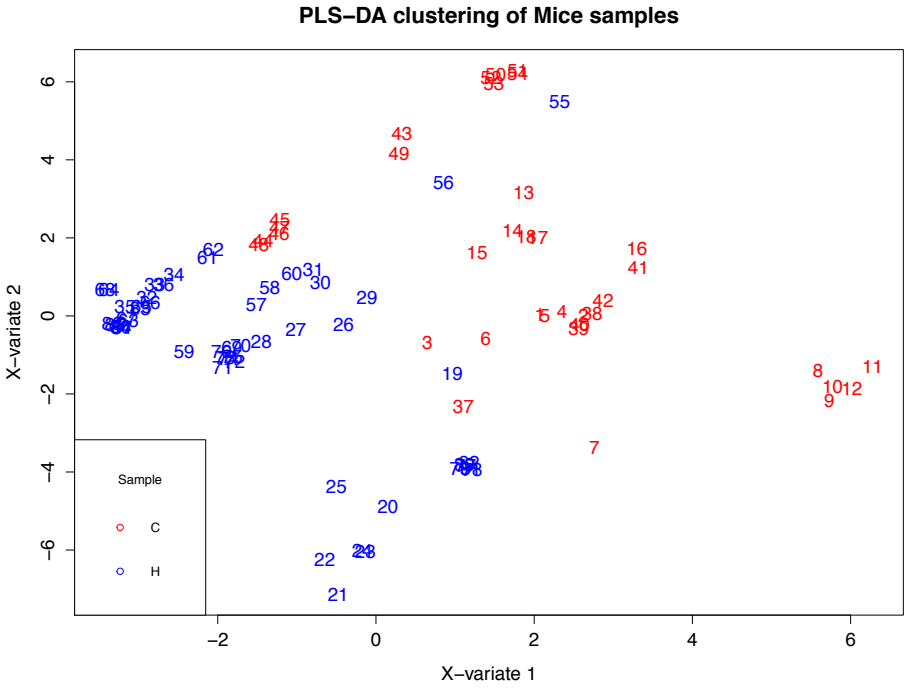


Figure 9. PLS-DA plot showing the separation between control and high fat diet samples. There is great overlap between these two samples, though some separation can be seen, showing that there may be some large-scale differences in the samples.

The greatest cluster separation is seen in the healthy vs. FLD induced samples, which is to be expected, as in the heatmaps above, there is decreased diversity in species seen in those samples vs. the healthy control samples. Diet does not seem to show a distinct difference, but one more PLS-DA plot was created, showing the four distinct groups and their separation, seen in figure 10 below.

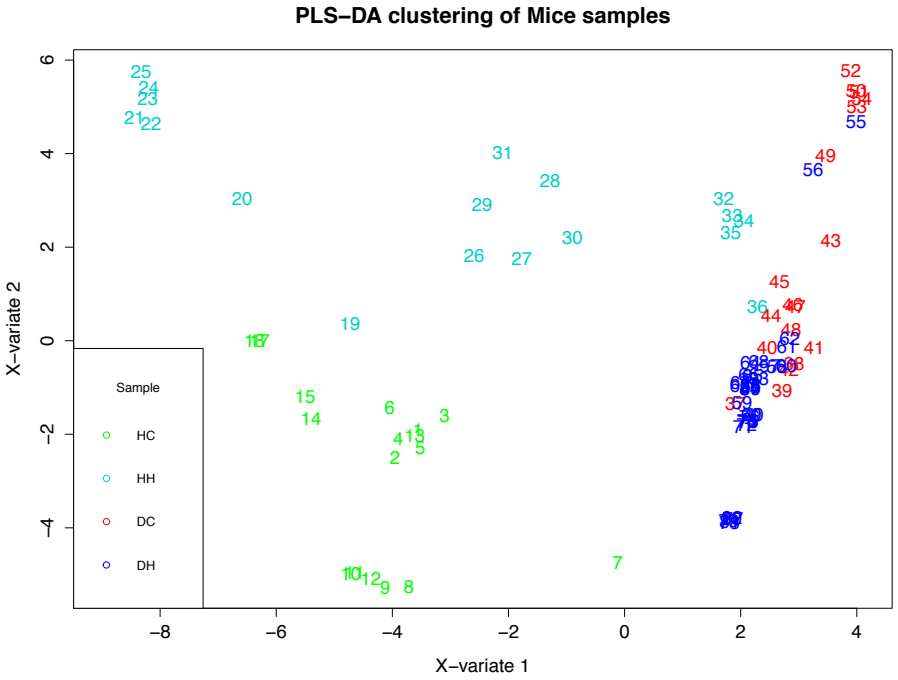


Figure 10. This PLS-DA plot differentiates between healthy and DEN injected samples, as well as control and high fat diets. As seen above, there is greater variation in the healthy samples, and in those samples, a difference is seen between control and high fat diet, while those two groups overlap more in the FLD samples.

PLS-DA was chosen in part due to the component of the analysis that calculates the VIP for each variable. The top five species influencing the differences in abundance between samples based on VIP rank are shown in table 3 below.

Species	VIP
Bacteroides dorei	1.502143985
Alistipes shahii	1.441760421
Lactobacillus johnsonii	1.379016336
Lactobacillus reuteri	1.328495176
Bifidobacterium unclassified	1.22464876

Table 3. The top 5 species demonstrating the highest variance between samples are shown in the above table. The VIP values were used to rank the species.

In order to quantify the changes in diversity observed in the above analyses, alpha and beta diversity values were calculated. The Shannon and Simpson diversity metrics can be seen below in table 4 for each of the samples. However, the results did not correlate with what is seen above. The function to calculate alpha diversity took as input a biom table, and so the merged species abundance table was converted to a biom table, using functions in QIIME, and so it is possible that somewhere in the conversion the data was not maintained. It's also possible that the visual changes in diversity seen in the heatmaps, and within the clusters, are overstated visually, and do not actually correlate to relevant changes in the diversity within the samples. This represents a limitation of these functions, in that the input required is the biom table, which is more difficult to visualize than the tab delimited microbial species abundance table, so it is difficult to see if there is an issue in that input data.

Sample	Shannon Index	Simpson Index
GRWS002D	4.33348196876	0.937176737505
GRWS002E	3.41951582265	0.895675107521
GRWS002F	3.61369644314	0.901492627754
GRWS002G	4.21349791818	0.932162057656
GRWS002H	3.78740749745	0.912722729813
GRWS002I	3.38330568022	0.883794439029
GRWS002J	3.95093758019	0.918697883877
GRWS002K	3.73491822398	0.905908421812
GRWS002L	3.6212091164	0.901938647331
GRWS002M	4.03428503999	0.919241612793
GRWS002N	3.32998605474	0.880608043903
GRWS002O	3.72800311364	0.918742710461
GRWS002P	3.39015637989	0.88988943845
GRWS002Q	3.29256647705	0.879696120928

Table 4. The above table shows both the shannon and simpson diversity metrics for each sample in the analysis. The three least diverse samples are in bold, two of which are FLD samples.

In order to assess function of the most abundant species, a translated blast query was run for all sample fasta files. The results of that search were visualized using the MEGAN5 tool, and a KEGG analysis was done within the MEGAN5 program to understand the functional aspects of those results. Results when all samples are loaded can be seen in figures below.

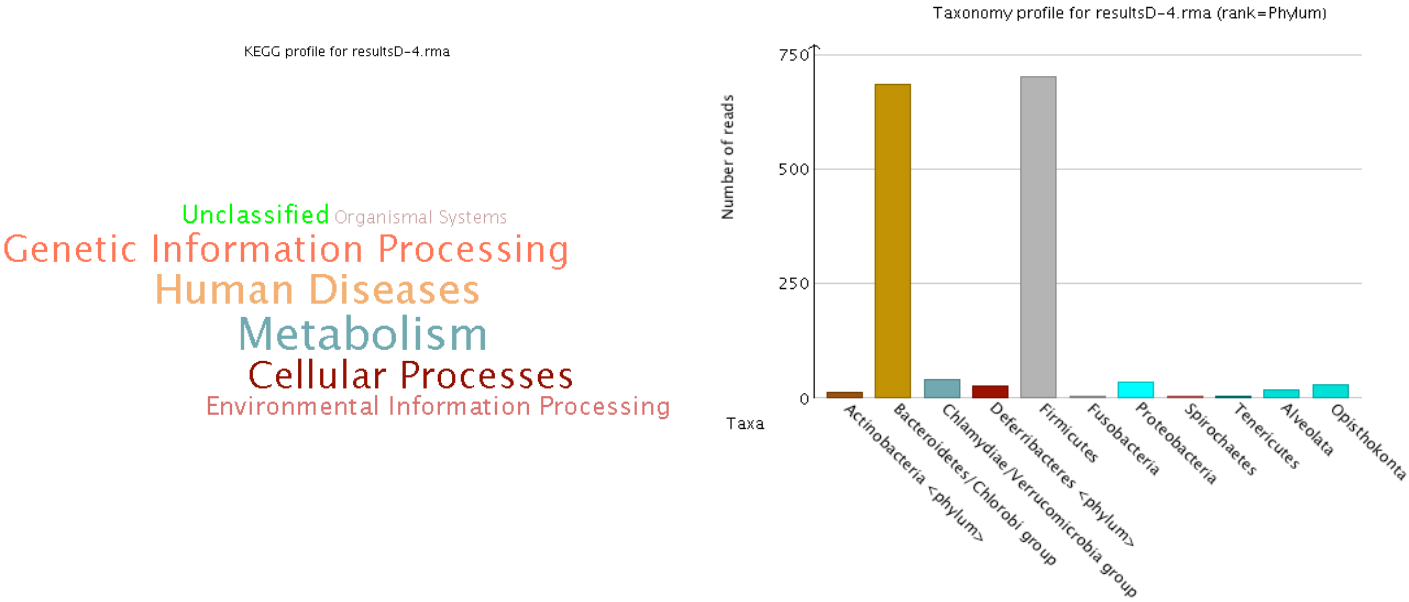
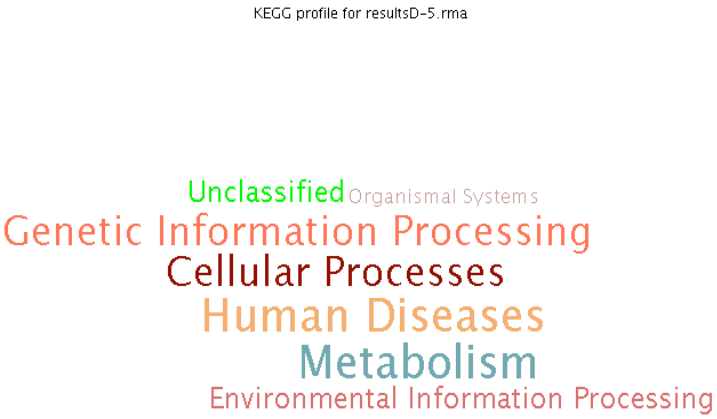


Figure 11. Above left shows the KEGG pathways enriched throughout the samples. The image to the right shows the overall abundance of species throughout the samples. As expected, Bacteroidetes and Firmicutes dominate.



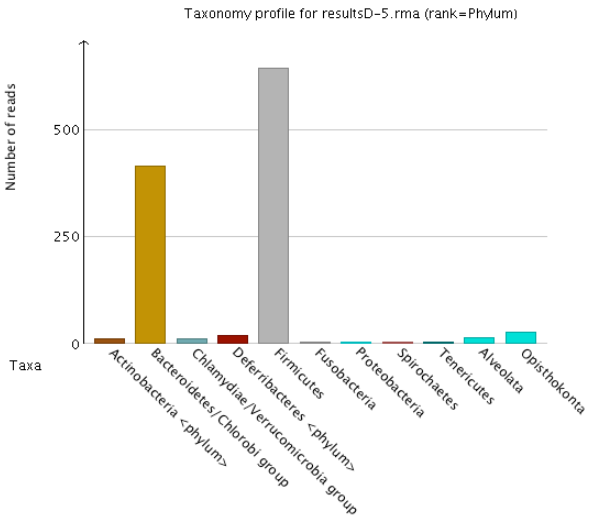


Figure 12. Above left shows the KEGG pathways enriched throughout the healthy samples. The image to the right shows the overall abundance of species throughout the healthy samples. As expected, Bacteroidetes and Firmicutes dominate. In looking at the KEGG pathways that are present in healthy samples, not much difference is seen between these samples and all samples.

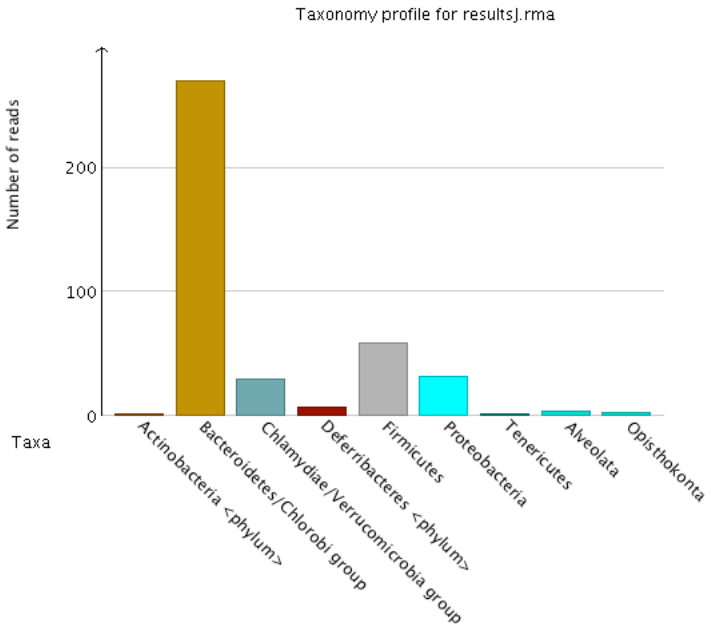


Figure 13. To the left is a bar graph showing the species abundance in FLD samples. Bacteroidetes is still highly represented, but there has been a sharp decline in Firmicutes, as compared to the healthy samples. The enriched KEGG pathways are not shown here, as only metabolism was listed as an enriched function based on these samples, further validating the hypothesis that disease and a decrease in bacterial diversity are correlated.

Discussion/Package Review

Overall, the results indicate that there is a decrease in diversity from healthy control samples to FLD induced samples. This change in bacterial diversity has been described in a number of other disease studies, including inflammatory bowel disease[9], specifically Crohn's disease[10], and Celiac's disease[11], among others, such as autism[12], obesity[12], and allergies[12]. This indicates that the microbiome may also be affecting the development of NAFLD, and could provide targets for therapy and diagnostic potential. The goal of this analysis was to understand the changes in the microbiome occurring between the different sample groups, and identify species that may be playing important roles in the disease's development and thus would be potential markers for therapy or diagnostic purposes. The

majority of analysis results indicate a decrease in species diversity, specifically in *Bacteroides dorei*, *Alistipes shahii*, and *Lactobacillus johnsonii*, as well as the Firmicutes phylum overall (of which, *L. Johnsonii* is a member). While the overall enriched pathways are shown above in the MEGAN5 KEGG results, these individual species functions are summarized below:

Species	Function
Bacteroides dorei	intestinal functioning[16]
Alistipes shahii	catalase activity[14], enzyme activities[15]
Lactobacillus johnsonii	polysaccharide and protein digestion, immunomodulation, pathogen inhibition, and epithelial cell attachment[13]

Table 5. Above table lists different functional aspects of some of the most differentially abundant species between different sample groups

Bacteroides dorei specifically has been implicated in other pathologies as well. Davis-Richardson, et al. show that the abundance of this species is increased in patients prior to development of Type 1 Diabetes (T1D)[17]. These three species represent potential avenues for disease treatment relative to FLD, and potentially other human diseases. It is also worth noting that the decrease in diversity seen in both the heatmap and clustering analysis in the FLD group has been implicated in other disease pathologies, and may represent a treatment avenue[18].

The packages and tools used in this analysis were able to produce results that had potential clinical implications. Particularly, MetaPhlAn was effective in producing the species abundances and providing input for clustering analysis. PLS-DA and the associated R packages were effective in showing separation between the different sample groups, and providing a quantifiable measurement of variable influence. Blast produced translated blast query results, which was useful as input into MEGAN5, but alone was difficult to interpret due to the length of the results. MEGAN5 allowed for easy visualization of the blast results, and provided KEGG enrichment pathways for each sample, but did not allow, at least in what was accomplished in this analysis, for the direct comparison between two sample results, and instead combined results, as in the output seen above. This was useful in comparing sample groups, but did not provide distinct sample to sample differences, though it corroborated the decrease in diversity seen in FLD samples. However, there were several tools that were attempted, but did not produce usable results. Both HuMaNn and Picrust, from the Biobakery package, were attempted many times with different input types, with no usable results. HuMaNn continued to return empty pathway output files, even when the translated blast query results were used as input, listed as a valid input metric in the HuMaNn manual. The error log file from the HuMaNn submission attempted to be included in the github for reference, since it included pathway output, but it exceeded the maximum file size (please contact if interested in seeing this file). None of those pathways appeared in the humann output files, and were not easily interpretable, unlike the KEGG analysis in MEGAN5, which is why that is presented here instead. Additionally, Picrust was run several times, with no results, as it returned the error that no OTUs were found in the database, based on the biom table that was used as input. The biom table was created from the microbial species abundance table. The QIIME pick_otus.py function was used to

attempt to generate compatible OTU tables, but despite the lack of errors returned by this job, no OTU tables were generated for use.

Future work would include working further with these functions and the type of input used for those programs to attempt to obtain usable results that would provide further insight into the functional pathways affected by the species composition of each sample. Currently, the tools used were sufficient to complete the analysis and provide potentially impactful results, and so the use of HuMaNn and Picrust is not strictly necessary, but it would provide further validation of enriched pathways, and potentially could show specifically what species are involved in each pathway and the changes in that from sample to sample in the different groups. Future work also includes looking at potential ways to target the species that are listed above as differing in abundance in a way that could have therapeutic implications in diseased patients. Microbiome transplantations are a potential therapeutic intervention that are beginning to gain traction as a viable clinical treatment, and could have beneficial effects in FLD patients, based on this analysis.

References (in order of appearance)

1. Angulo, P. (2002). Nonalcoholic fatty liver disease. *New England Journal of Medicine*, *346*(16), 1221-1231.
2. Le Roy, T., Llopis, M., Lepage, P., Bruneau, A., Rabot, S., Bevilacqua, C., ... & Gérard, P. (2013). Intestinal microbiota determines development of non-alcoholic fatty liver disease in mice. *Gut*, *62*(12), 1787-1794.
3. Machado, M. V., & Cortez-Pinto, H. (2012). Gut microbiota and nonalcoholic fatty liver disease. *Ann Hepatol*, *11*(4), 440-449.
4. Aron-Wisnewsky, J., Gaborit, B., Dutour, A., & Clement, K. (2013). Gut microbiota and non-alcoholic fatty liver disease: new insights. *Clinical Microbiology and Infection*, *19*(4), 338-348.
5. Metagenomic microbial community profiling using unique clade-specific marker genes. Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, Curtis Huttenhower. Nature Methods, 8, 811–814, 2012
6. Pérez-Enciso, M., & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human genetics*, *112*(5-6), 581-592.
7. Dejean, S., Gonzalez, I., Le Cao, K. A., Monget, P., Coquery, J., Yao, F., ... & Rohart, F. (2011). mixOmics: Omics data integration project. *URL http://CRAN. R-project. org/package= mixOmics. R package version*, 2-9.
8. Hervé, M. (2014). RVAideMemoire: diverse basic statistical and graphical functions. *R package version 0.9–32*.
9. Ott, S. J., Musfeldt, M., Wenderoth, D. F., Hampe, J., Brant, O., Fölsch, U. R., ... & Schreiber, S. (2004). Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut*, *53*(5), 685-693.
10. Fujimoto, T., Imaeda, H., Takahashi, K., Kasumi, E., Bamba, S., Fujiyama, Y., & Andoh, A. (2013). Decreased abundance of Faecalibacterium prausnitzii in the gut microbiota of Crohn's disease. *Journal of gastroenterology and hepatology*,*28*(4), 613-619.
11. Wacklin, P., Kaukinen, K., Tuovinen, E., Collin, P., Lindfors, K., Partanen, J., ... & Mättö, J. (2013). The duodenal microbiota composition of adult celiac disease patients is associated with the clinical manifestation of the disease.*Inflammatory bowel diseases*, *19*(5), 934-941.
12. Clemente, J. C., Ursell, L. K., Parfrey, L. W., & Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell*, *148*(6), 1258-1270.
13. Pridmore, R. D., Berger, B., Desiere, F., Vilanova, D., Barretto, C., Pittet, A. C., ... & Schell, M. A. (2004). The genome sequence of the probiotic intestinal bacterium Lactobacillus johnsonii NCC

533. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(8), 2512-2517.

14. Lagier, J.-C., Armougom, F., Mishra, A. K., Nguyen, T.-T., Raoult, D., & Fournier, P.-E. (2012). Non-contiguous finished genome sequence and description of *Alistipes timonensis* sp. nov. *Standards in Genomic Sciences*,*6*(3), 315–324. doi:10.4056/sigs.2685971
15. Flores, R., Shi, J., Gail, M. H., Gajer, P., Ravel, J., & Goedert, J. J. (2012). Association of fecal microbial diversity and taxonomy with selected enzymatic functions. *PloS one*, *7*(6), e39745.
16. http://microbewiki.kenyon.edu/index.php/Bacteroides_dorei_Phenomics
17. Davis-Richardson, A. G., Ardissone, A. N., Dias, R., Simell, V., Leonard, M. T., Kemppainen, K. M., … Triplett, E. W. (2014). *Bacteroides dorei* dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Frontiers in Microbiology*, *5*, 678. doi:10.3389/fmicb.2014.00678
18. Manichanh, C., Reeder, J., Gibert, P., Varela, E., Llopis, M., Antolin, M., … Guarner, F. (2010). Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. *Genome Research*, *20*(10), 1411–1419. doi:10.1101/gr.107987.110