An alternative method to processing sequences for taxonomic classification

Kaitlyn Hobbs

August 2018

4483E

Supervisor: Dr. Gloor

Contents

1 Abbreviations

Term	Definition
ASV	Amplicon Sequence Variant
OTU	Operational Taxonomic Unit
Feature	Either an OTU or ASV depending on the pipeline applied
Technical variant/technical error	Sequencing error introduced by PCR or Illumina
Read/sequence variant	Sequence obtained from HTS instrument
MS	Multiple sclerosis

2 Abstract

Microbial communities influence host health. To profile these communities and deduce dysbiosis, the 16S ribosomal RNA gene is sequenced using high-throughput technology. Nonetheless, technical error introduced during sequencing coupled with amplicon processing and data analysis can influence results. Deducing amplicon sequence variants is alternative to generating operational taxonomic units for microbial profiling. In this study, an open-access 16S rRNA dataset was used to compare and assess the robustness of the ASV standard operative procedure against OTU methods. Additionally, a dataset comparing gut microbiomes between multiple sclerosis patients and healthy individuals with unknown compositions was used in attempt to validate published findings.? Ultimately, the ASV-based approach proved to generate more reliable results by excluding technical errors from its output and evading problems associated with clustering. In contrast to published results, no significant differences were found between MS patients and healthy individuals; however, vague reporting of computational methods prevented exact replication of the study.

3 Introduction

3.1 Sequencing Microbiomes

50% of cells comprising humans belong to microbes, most which reside in the gastrointestinal tract.[?,?] Microbial communities are distinguished by their environment, or "biome", and are known to partake in symbiotic or commensal relationships with their host.[?] Profiling these communities can help in identifying dysbiosis: an imbalance in a microbiome, which often directly impacts host health. For example, individuals with irritable bowel syndrome (IBS) have a different gut microbiota relative to healthy individuals.[?] This discovery provides insight to the potential role of gut microbes in IBS pathogenesis and opens the door for microbiota-directed therapies.

Typically, microbiota are profiled through sequencing of all or part of the 16S ribosomal RNA (rRNA) gene found natively in micro-organisms. It contains nine variable regions that differentiate with taxa differentiation and is used to associate microbes to taxonomic classes. Sequences are most frequently obtained through Next Generation Sequencing (NGS), processed using a computational pipeline, and referenced to an external database for taxonomic classification.[?] Though the high-throughput, low cost, and low false positive generation of Illumina instruments is preferred, technical errors can be introduced through sequence-by-synthesis.[?,?,?] DNA polymerases may introduce errors during PCR amplification and Illumina primer extension and amplification. Read processing pipelines output a count table, which include the number of instances that a sequence was observed by the pipeline in each sample. From there, data can be analyzed to evaluate the relative abundance of a microbe in a biome. Errors introduced in the preliminary cycles of amplification can be propagated, ultimately, misrepresenting the identity and proportions of microbes.[?,?] Accounting for this error depends on how the sequences are processed.

3.2 Methods to Processing Sequences

There are two main approaches to processing reads: by clustering sequences into operational taxonomic units (OTUs), or by deducing amplicon sequence variants

(ASVs). OTUs are clusters of reads that share an arbitrary threshold of dissimilarity; for example, if a dissimilarity threshold was chosen at 3% then sequences that are 97% identical would be grouped together to form one OTU. This OTU is aligned with sequences housed in an external database that also share 97% sequence identity and assigned the corresponding taxonomy.[?] This processing method does not consider technical error introduced by DNA polymerase in PCR and Illumina instruments. As a result, sequence variation within an OTU cluster may generate both false positive and negative results by including technical errors and natural variation, respectively.[?] Incorrectly grouped technical errors misrepresent a sequence from one taxonomy for another. False negatives arise if a sequence containing natural variation is inappropriately clustered with sequence variants belonging to a distinct taxonomy. Consequently, data obtained by processing reads using OTU clustering are less reproducible and less ideal for inter-dataset comparisons.[?,?,?]

Alternatively, inclusion of technical errors can be mediated by inferring the error rate for each base in each position using an ASV approach. Erroneous reads can be identified by their low abundance relative to true biological sequences and low beta-association with other sequences.[?,?] Relative abundance provide an inference of the total count of a sequence in an environment.[?,?,?] ASV methodolgy allows for single nucleotide resolution, distinction between natural and technical variation, as well as evaluates relative abundances.[?] Together, interpretations made from generating ASVs are more robust, reproducible, and better suited for inter-dataset comparisons. Though the ASV-based approach is specific to Illumina, analogous methods exist for other NGS procedures.[?]

3.3 Research Premise

The bidirectional relationship between gut microbiota and the central nervous system (gut-to-brain axis) has caught the attention of researchers.[?] Relationships found between microbes and hosts have ignited the Human Microbiome Project, which aims to use high-throughput technologies to unravel host-associated microbiomes and elucidate their role in medical conditions.[?] In pursuit of these goals, it is necessary to optimize accuracy of each stage in profiling microbial communities.

Unfortunately, although many publications have drawn conclusions supporting the influence of the gut microbiome on host health, these claims has yet to be reproduced. This research study has two focuses: testing the robustness of the ASV-based processing method in effort to improve accuracy in microbiome profiling and validating the presence of dysbiosis in multiple sclerosis shown by published researches.[?]

4 Datasets

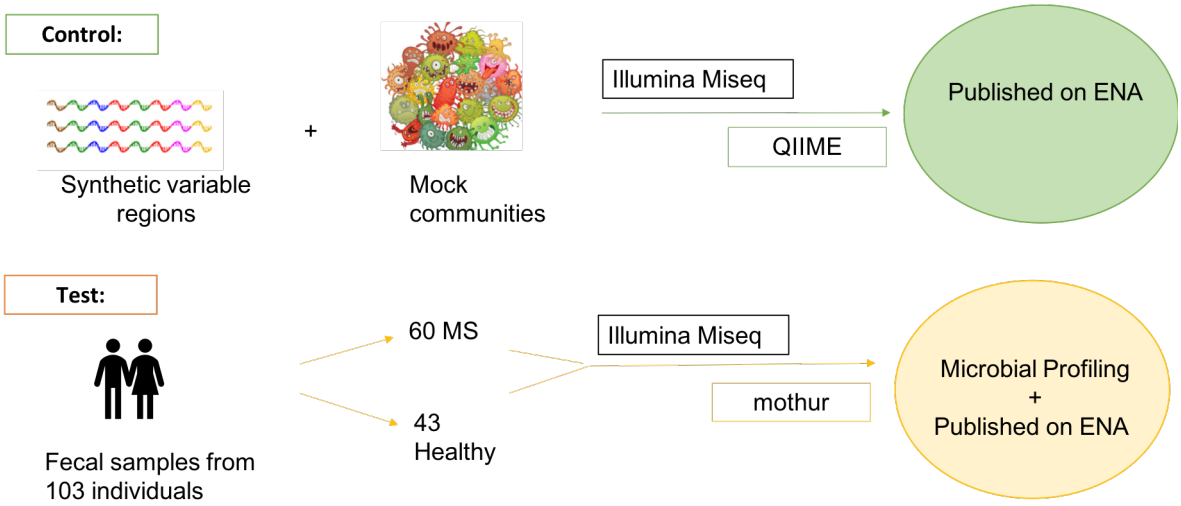

Figure 1: **Overview of control and test datasets.** Datasets were retrieved from the European Bioinformatics Institute (EBI). The control dataset is comprised of synthetic variable regions with no known microbe identity and a series of mock communities composed of 15 known species. The test dataset is composed of variable 4 regions obtained from fecal samples of 43 healthy individuals and 60 multiple sclerosis patients. For both datasets, sequences were obtained from Illumina Miseq, processed using OTU-based pipelines QIIME and mothur, respectively, and published on the European Nucleotide Archive (ENA). Both mothur and DADA2 pipelines reference the SILVA database while QIIME references GreenGene and Broad Microbiome Utilities' 16S Gold.

Jangi et al reported that individuals with MS have gut dysbiosis on the phy-

lum and genera level compared to healthy individuals. Specifically, the *Prevotella*, *Sarcina*, *Sutterella*, *Butyricimonas*, *Akkermansia*, and *Methanobrevibacter* genera differed in abundances. They also found that all MS patients (untreated or treated with either Interferon or Copaxone) differed in abundance from healthy individuals in the *Euryarcheota* and *Verrucomicrobia* phyla.[?] The researchers clustered reads at a 97% identity similarity and used DESeq2 to perform statistical tests for significant differences in taxa with a p-value of 0.05. The p-value was corrected using the Benjamini-Hochberg method with a false discovery rate threshold of 0.1.[?]

The control dataset comprises both mock communities containing 15 known microbes and variable regions synthesized by Tourlousse et al with no known sequence identity to microbial taxonomic groups.[?]

5 Methods

R scripts used are included under Computational Methods in Supplementary Data.

6 Hypotheses

Since synthetic variable regions contain no sequence identity to known microbes, it was expected that they would not be given a taxonomic assignment after processing by DADA2.[?] Conversely, because the mock communities have a known composition, DADA2 was expected to accurately identify these sequences with high resolution and detect more sequence variants than QIIME. Finally, the number of reads from the multiple sclerosis study processed using DADA2 was thought to differ from the number of reads generated by mothur. This is because DADA2 is expected to give fewer false positives and negatives leaving only true biological variants for taxonomic assignment. Large variation within prokaryotics taxonomies makes comparing sequence variants as opposed to classifications more reliable.[?]

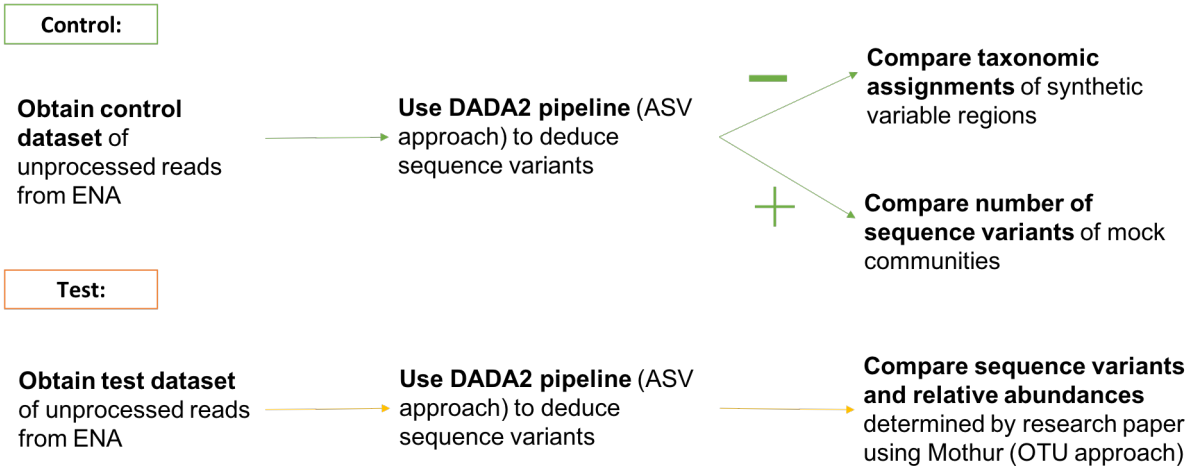

Figure 2: **An overview of methods**. Both raw Illumina Miseq reads from control and test datasets were retrieved from the European Nucleotide Archive (ENA) and processed using the ASV-based pipeline, DADA2, which references the SILVA metagenomic database to assign taxonomy. The synthetic variable regions from Tourlousse et al served as a negative control and taxonomic assignments were observed. The mock communities served as a positive control where the number of ASVs generated by DADA2 were compared to the number of OTUs found by QIIME. The MS dataset from Jangi et al was processed by DADA2 and compared to findings published by mothur?

7 Results

7.1 Positive Control

Unexpectedly, DADA2 found less than half the number of features found by QIIME. Count tables were run through the filtR function in R to identify technical variants. Technical errors are identified on the basis of low beta-association between two features (ASVs or OTUs), as represented by ρ, and low abundance of each feature across all samples, represented by the center-log ratio (clr). Results showed that DADA2 included no technical variants in its count table while QIIME included one (Table 1).

	DADA2	QIIME
No. of samples used	24	24
No. of features	985	1752
No. of technical variants*	0	1
Reference database	SILVA	Broad Microbiome Utilities' 16S Gold

Table 1: **Contrast between methodology and sequence variants produced of the DADA2 and QIIME pipelines.** 24 samples of mock communities spiked with synthetic variable regions were run through DADA2 and compared to results published by Tourlousse et al using QIIME. *Technical variants were determined using the filtR R package with a ρ cutoff of 0.7 and a clr cutoff of 5.

7.2 Negative Control

Since Illumina reads are short fragments, DADA2 generated ASVs were found in th elonger synthetic spike-in sequences by nBLAST. Taxonomic assignments of ASVs that aligned with the spike-ins were noted. Twelve spike-ins were found in 143 sequence variants within the control samples. Almost all ASVs were recognized as members of the bacteria kingdom, twenty-five were assigned a phylum, and one was given a class (Figure 3).

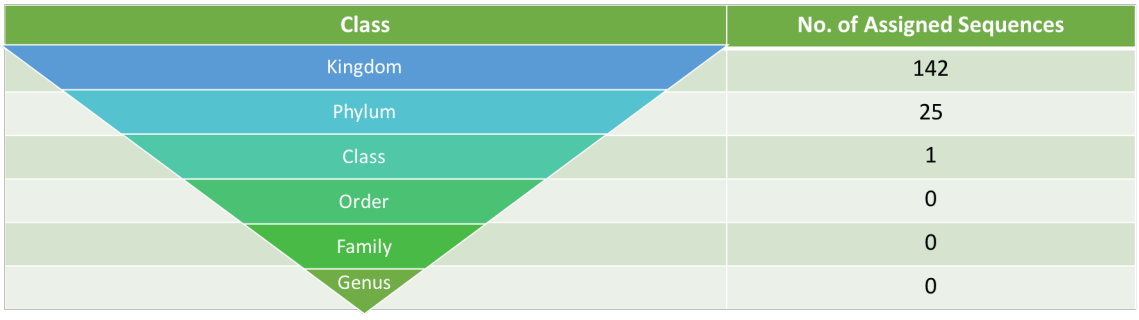

Figure 3: **Number of ASVs that were assigned a taxonomy by SILVA.** BLAST results with 100% identity and equal nucleotide length between DADA2 ASVs and the synthetic spike-in sequences confirmed the presence of the spike-ins in specific variants. If the variants were assigned a taxonomic class at a low rank, then they were considered a false positive.

7.3 Multiple Sclerosis Study

The OTU pipeline, mothur, found half the amount of features as DADA2 and included more technical errors (Table 3). In contrast to findings reported by Jangi et al, multiple analysis tools used to analyze the processed sequences did not find significant differences between gut microbiomes of MS patients and healthy individuals using both mothur and DADA2 count tables.

	DADA2	mothur
No. of samples used	105	105
No. of features	2243	1191
No. of technical variants*	0	4
Reference database	SILVA	SILVA

Table 2: **Contrast between sequence variants produced and methodology of the DADA2 and QIIME pipelines.** 105 fecal samples from 63 individuals were run through DADA2 and compared to results generated by QIIME. Features are either ASVs or OTUs depending on the pipeline used. *Technical variants were determined using the filtR R function with a ρ cutoff of 0.7 and a clr cutoff of 5.

To qualitatively observe variances between subgroups in the multiple sclerosis dataset, a compositional biplot was created using data analysis tool, CoDaSeq.[?] In contrast to results published by Jangi et al, Illumina Miseq reads processed by DADA2 lead to no significant difference between subgroups (Table 4 A). Principal component analysis (PCA) is a multivariate analysis tool that observes compositional data in n-dimensional space while the PCA biplot projects the first and second dimensions, or principal components 1 and 2, respectively, that explain the most variation.[?,?] It should be noted that the principal components are relatively low, indicating low explanatory power in the dataset (reffig:3). Low power suggests that any variation explained occurs due to random chance. Generally, the loadings biplot show any driving forces for variation observed in the sample biplot. In this dataset, the loadings are clustered around zero for both DADA2 and mothur, indicating that no features separate MS patients and healthy individuals. This is also

represented by the indistinguishability of the sample subgroups (Figure 4).

Since multivariate analysis showed no separation between subgroups, univariate analysis was pursued to confirm that no features drive variation. For all two-way analysis, subgroups were paired as follows: control and untreated, control and patients treated with Copaxone, and control and patients treated with Interferon. Effect plots generated by the differential abundance tool, ALDEx2, complimented the biplot results showing no features to be significantly different between groups. In fact, more variation within subgroups than between is evident[?,?] (Figures 6, **??**, **??**).

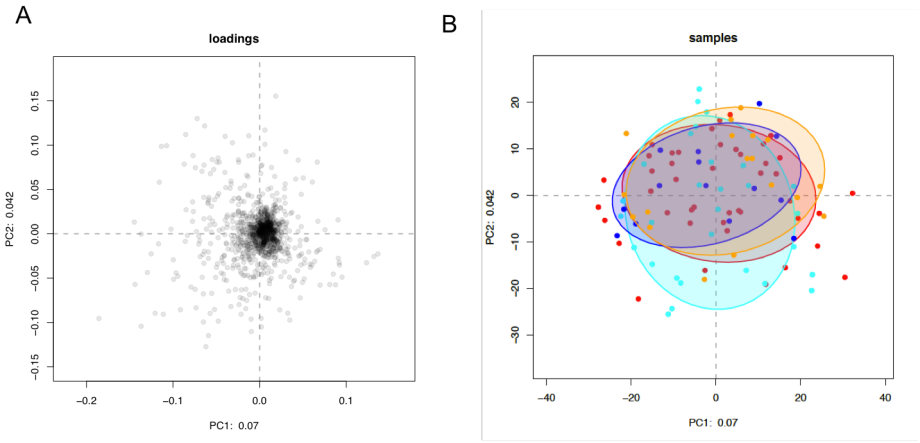

Figure 4: **Compositional biplot comparing control and untreated groups processed using DADA2.** A. A loadings biplot. Each dot represents a loading (feature) identified by the pipelines. Distinctive loadings drive variation between subgroups in the coloured biplot. B. A coloured biplot, which visually explains compositional differences between subgroups. Control group is represented by red, untreated by cyan, patients treated with Copaxone by blue, and patients treated with Interferon by orange. Principal components 1 and 2 serve as the x- and y-axis, respectively and indicate axes with the most variance explained.

Comparison of data processing protocols used in this study to those used by Jangi et al's research team showed an inconsistency in the data analysis tools used (reftable:3). While this study uses ALDEx2, Jangi et al ran DESeq2, which differs

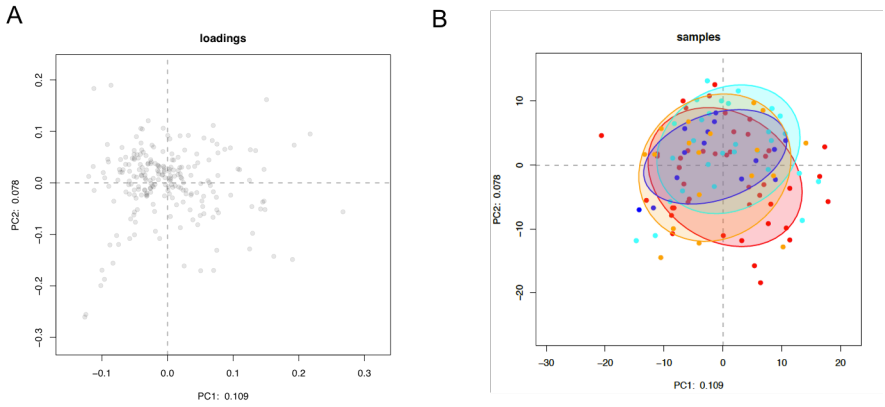

Figure 5: **Compositional biplot comparing control and untreated groups processed using DADA2.** A. A loadings biplot. Each dot represents a loading (feature) identified by the pipelines. Distinctive loadings drive variation between subgroups in the coloured biplot. B. A coloured biplot, which visually explains compositional differences between subgroups. Control group is represented by red, untreated by cyan, patients treated with Copaxone by blue, and patients treated with Interferon by orange. Principal components 1 and 2 serve as the x- and y-axis, respectively and indicate axes with the most variance explained..

in data normalization, dealing with zero counts, and mean-variance assumptions.[?] To reduce confounding variables, count tables derived from both pipelines were processed with ALDEx2 and DESeq2. The standard DESeq2 protocol was followed with default parameters as this was what was expected to be what was employed by the researchers.[?] Results found significant increases and decreases in abundances for tables generated by both pipelines (Tables **??**, **??**); however, none were observed by ALDEx2 (Tables **??**, **??**). Upon evaluation of MA plots, DESeq2 plots were abnormal and unreadable indicative of inappropriate data normalization methods (8A). Features clustered around zero, as shown in the ALDEx2-generated MA plots, represent normal data with low variation between groups and high variation within (Figure 8B).

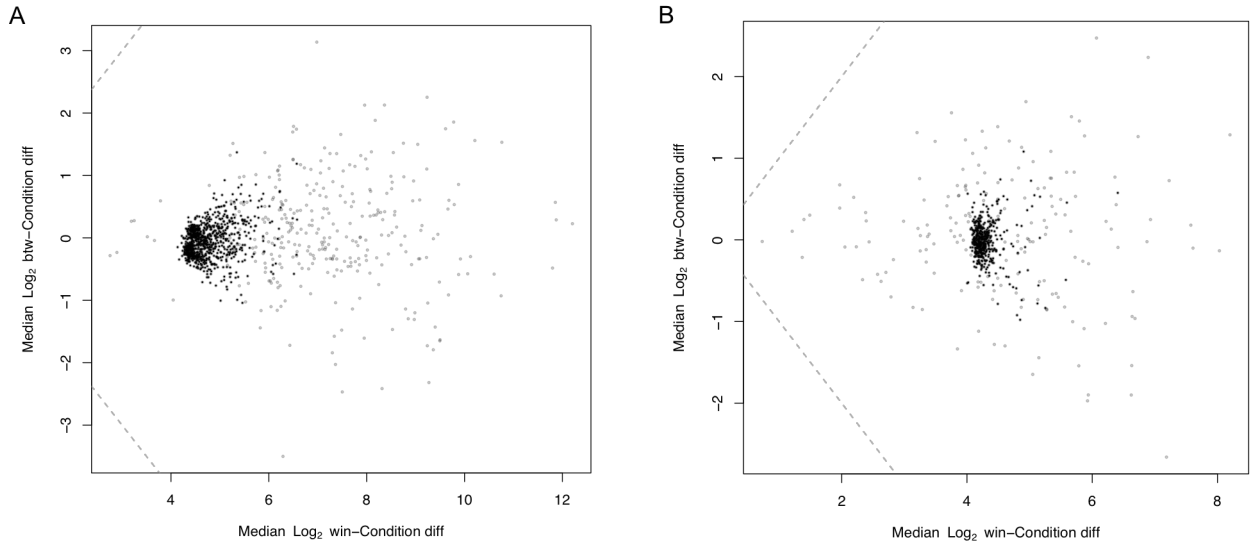

Figure 6: **A. Effect plot comparing control and untreated groups processed using DADA2. B. Effect plot comparing control and untreated groups processed using mothur.** Each point on the plot represents one feature. Effect is measured by taking a ratio of the difference between groups and the difference within. The diagonal lines represent the line of equivalence (effect is 1:1) and any points preceding the line represents a feature that has more variation between groups than within.

Method Component	Present Study	Publication Methods
Processing reads	ASV-based approach	OTU-based approach
Data analysis tool	ALDEx2	DESeq2
Multiple test correction	Benjamini-Hochberg = 0.1	Benjamini-Hochberg = 0.1
Significance test	Wilcoxon rank sum	Wilcoxon rank sum

Table 3: **Comparing data processing protocols used in the present study and by Jangi et al.**

8 Discussion and Limitations

8.1 Positive Control

Although the QIIME output had almost double the number of features as DADA2, filtR results indicated that only a small proportion are technically introduced false

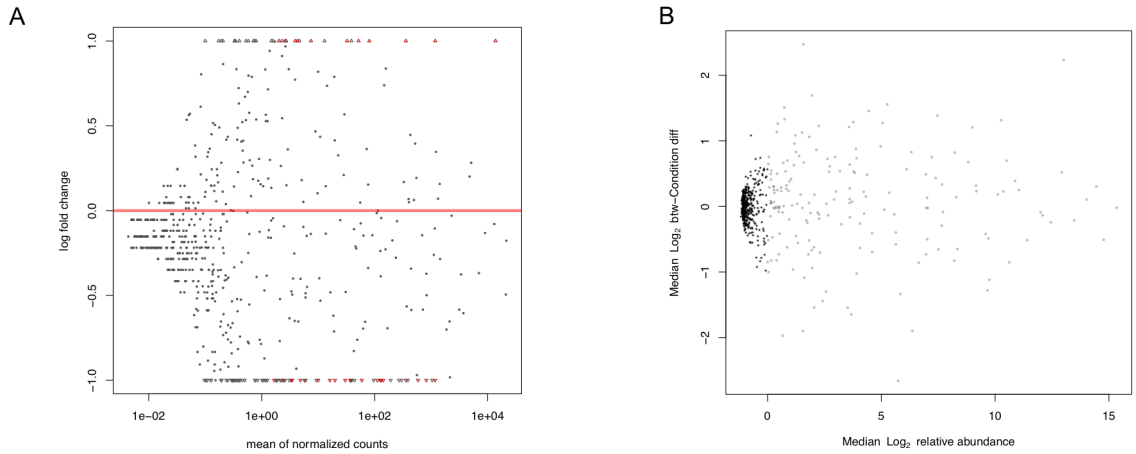

Figure 7: **A. MA plot comparing control and untreated groups processed using DADA2. B. MA plot comparing control and untreated groups processed using mothur.** MA plots plot the difference between measurements of two samples or subgroups. "M" represents the log ratio of the difference between samples on the x-axis and "A" displays the mean average along the y-axis.

positives. The remaining difference may be attributed to the clustering method employed by QIIME. QIIME is an open reference pipeline that references sequences to the external database, GreenGene. Rare variants that aren't found in the database form *de novo* clusters and are referenced to an alternative database. To cluster reads, Tourlousse et al utilized the tool USEARCH, which applies centroid clustering and references the Broad Microbiome Utilities' 16S Gold database.[?] Centroid clustering arbitrarily selects a feature to become the center of a cluster and sequentially aligns other sequences to evaluate their inclusion. Sequences are included in the cluster if they meet the chosen percent similarity to the center or average of the sequences within the cluster. The drawback of this method lies in the order of which sequences are selected as the centroid or added to the cluster. As shown in 8, if a sequence highly dissimilar to the remainder of a dataset is chosen as the center, many more clusters will be generated. This may explain the vast difference in the number of OTUs generated by QIIME compared to the number of ASVs generated by DADA2.

Since there were only 15 microbes in the mock communities and each feature corresponds to a taxonomic assignment, it was expected that approximately 15 features

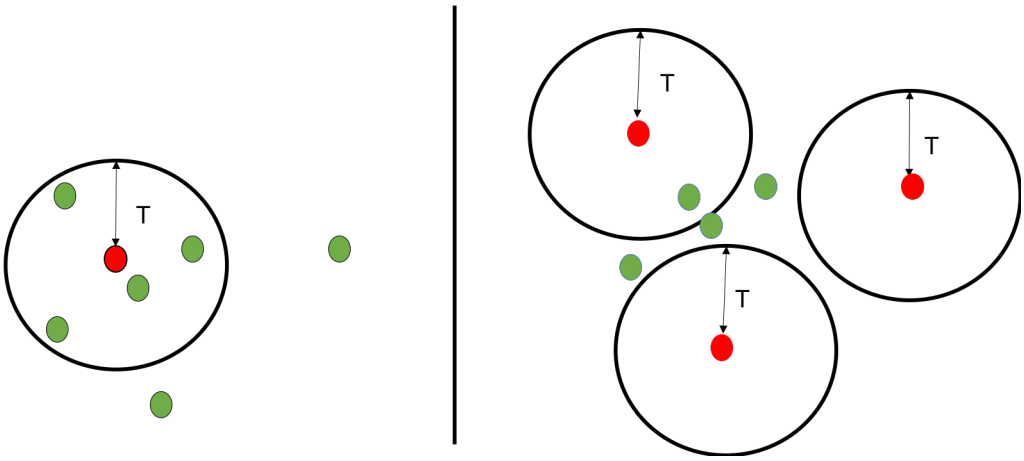

Figure 8: **Visualization of centroid clustering.** Each circle represents a sequence and proximity of sequences represents sequence similarity. Black circles represent resulting clusters with a defined radius threshold (T) signified by a black arrow. Left: clustering initiated by a sequence highly similar to the remainder of the dataset. Right: clustering initiated by sequences highly dissimilar to the remainder of the dataset.

would be detected. Nonetheless, both DADA2 and QIIME grossly overestimated the diversity in the communities. Throughout the duration of this study, updated versions of the pipelines were released and improvements are ongoing. These fixes along with more stringent filtering methods may help collapse features that represent the same microbe to reduce false positive results. Another study comparing sequence processing pipelines found that DADA2 identified more rare ASVs in a host-associated biome but at the cost of false positives.[?] The same study also found that OTU pipelines consistently identified more features than DADA2, likely as a result of the clustering method while a separate study attributed the reported inflated diversities by QIIME to inadequate error filtering.

8.2 Negative Control

The team of scientists that created the synthetic 16S rRNA variable regions intended for the sequences to resemble bacteria, therefore it is not surprising that 142 of 143 spike-in ASVs were identified as members of the bacteria kingdom when

referenced to the SILVA database.[?] A class assignment is an unusually high resolution for synthetic sequences with no reported identity and can be explained by differences in the taxonomic resolution and sequences housed of reference databases. To check the identity of the synthetic sequences, Tourlousse et al referenced NCBI's nt, est and est_human nucleotide sequence databases both before and after incorporation into 16S rRNA gene constructs. Although the NCBI databases contain more bacteria sequences, SILVA is regularly updated and quality checks sequences.[?,?] Consequently, unexpected classifications of the synthetic regions may have been a result of referencing a more recent database since the time of original publication.

8.3 Test

The discrepancy between gut dysbiosis reported by Jangi et al and this study's results may be attributed to numerous factors. For one, low power of the study represented by small cohorts is evident (Table **??**). Although the same hypothesis tests were used when generating MA plots to allow for a fair comparison, having a low powered study can still cause false results.[?] Abnormal MA plots generated from both mothur and DADA2 pipelines in DESeq2 are indicative of unnormalized data. The normalization method employed by this tool is encoded in the "DESeq" function without room to modify. If custom scripts were made to normalize the processed reads by Jangi et al, they were not included in the published study. Conversely, if MA plots were overlooked by the researchers the interpretations reported may be inaccurate. Moreover, while the analytical tool DESeq2 found significant differences in feature abundances for both mothur and DADA2 count tables, ALDEx2 found none. The consistency between DADA2 and mothur analysis results ultimately suggests that the discrepancy in the published article lies in the analytical tool, DESeq2, as opposed to in the mothur pipeline.

The Benjamini-hochberg correction attempts to mitigate the uprising of false positives in small sample sizes by calculating the fraction of false positives given a p-value and population size and adjusting according to an arbitrary threshold of "acceptable" false positives. A more appropriate tool that evades using the p-value significance testing entirely is the ALDEx2 effect plot.[?] Effect plots evaluate rela-

tive abundances by considering the ratio of differences between and within groups. As a result, they do not require a large sample size to accurately represent differences within and between datasets.[?] For this study, we can depend more heavily on ALDEx2 results because of its ability to eliminate low powered biases by generating effect plots based on relative abundances and provide evidence of correct data normalization. With that in mind and in contrast to results reported by Jangi et al, no significant difference was observed between MS patients and healthy individuals.

Another factor that influenced the reproducibility of the published result is the absence of detailed reporting of methods by the researchers. Custom python scripts used to filter sequences, exact primer sequences used, and the SILVA database reference file were omitted, all which impact crucial steps when running the mothur pipeline. This number of uncertain variables would undoubtedly lead to incomparable results thus replicating the employed protocol *ab initio* was abandoned.

Despite the apparent influence of analytical tools on interpreting results, it should be noted that mothur still produced technical variants detectable by filtR. Comparatively, DADA2 consistently detected different numbers of features than the OTU-based pipelines used in the test and control with no technical variants. This seems to suggest that DADA2's higher resolution permits a more thorough and robust distinction of true biological variants as well as effectively removes erroneous reads.

9 Conclusions

The ASV-based standard operative procedure is more reliable than OTU clustering because it is more durable against technical errors and avoids variable results produced by clustering methods. Yet, over-estimation of populations remains an issue for both processing methods.

A lack of detailed reporting of methodology lead to irreproducible results for the MS study. Consequently, this study could not confidently conclude a difference in gut microbiomes in MS patients relative to healthy individuals. Discrepancies in results may be attributed to the way data is analysed for interpretation after being run through pipelines since both mothur and DADA2 pipelines generated consistent results regardless of the tool used.

10 Further Directions

To establish confident conclusions on the role of the gut microbiome in multiple sclerosis, new datasets should be obtained with a larger number of samples to increase the power of the study. Currently, pipelines are being improved and latest versions should be used to process reads. Moreover, more reliable results can be ensured by using an ASV pipeline. Finally, raw data should continue to be reported and open access; however, much like adapted methods are reported in publications, revisions to computational scripts should be openly documented as well.

11 Acknowledgements

I'd like to thank the Gloor Lab for all their help and support throughout the summer. Specifically, I'd like to acknowledge Dr. Greg Gloor, Jean Macklaim, Ben Joris, and Dan Giguere for their patience with my question-asking and aid in my slow ascent over the learning curve that is coding in R. This lab has fostered a positive work environment that has considerably enhanced my knowledge and made each day in the lab enjoyable. I'd like to give a special thanks for Dr. Ball and Dr. Brandl for their feedback and encouragement in writing as well as managing the course. Finally, a big thank you to Ryan Szukalo for being my on-call Latex advisor and handyman, without whom this PDF would not be possible.