Evaluation of the impact of bioinformatic pipeline, normalization methods, and sequence data characteristics on beta diversity metrics for 16rRNA marker-gene survey data.

1 Abstract

2 Introduction

- 16S metagenomics and ecological diversity analysis
- sequence data characteristics
  - differences in library size
  - error rate
- Bioinformatic pipelines
  - differentiating sequencing errors from biological variation
  - grouping sequences into biologically informative units
- Normalization methods
  - rarefaction
  - numeric methods
    - microarray
    - marker-gene survey
- Diversity metrics
  - phylogenetic v. taxonomic
  - weighted v. unweighted metrics
  - metric interpretation
- Previous diversity assessments
  - MBQC
  - McMurdie and Holmes 2013
  - Weiss et al. 2017

3 Methods

3.1 Methods

- data set
  - mixtures from Olson et al. in-prep
  - four sequencing runs
    - libraries prepared at independent laboratories using the same protocol (ILLUMINA) with the sample 16S PCR as input.
    - libraries were sequenced twice at each laboratory.
  - For the first laboratory (JHU) the base quality scores were lower than expected and the instrument was re-calibrated before the second run resulting in improved quality scores.
  - For the second laboratory (NIST) the total run throughput was lower than expected, the pool library was re-optimized for resulting in increased throughput and lower sample to sample read count variability.

- Seq data characterization - R packages for calculating summary values
  - illumina quality control output (Calder 2015)
  - (Souza and Carvalho 2017)

3.2 Bioinformatic Pipelines

Data from the four sequencing runs was processed using 6 bioinformatic pipelines including the QIIME open reference, closed reference, de novo, and deblur pipelines, as well as the Mothur de novo and DADA2 sequence inference pipelines. Code used to run the bioinformatic pipelines is available at https://github.com/nate-d-olson/mgtst/_pipelines/, on the multirun branch. The Mothur pipeline uses the OptiClust algorithm for de novo clustering (Westcott and Schloss 2017). Preprocessing includes merging and quality filtering paired-end reads followed by aligning sequences to the SILVA reference alignment (Schloss et al. 2009). Taxonomic classification was performed using the Mothur implementation of the RDP bayesian classifier (Wang et al. 2007). The phylogenetic tree was constructed in Mothur using the clearcut algorithm (Sheneman, Evans, and Foster 2006). Mothur version 1.39.3 (https://www.mothur.org) and SILVA release version 119 reference alignment and RDP the mothur formatted version of the RDP 16S rRNA database release version 10 (Cole et al. 2014).

The DADA2 big data protocol for DADA2 versions 1.4 or later was followed (https://benjjneb.github.io/dada2/bigdata.html), except for read length trimming parameters and primer trimming. The forward and reverse reads were trimmed to 260 and 200 bp respectively. Using the values from the online protocol resulted in total abundance values around 5000. Forward and reverse primers were trimmed using cutadapt version 1.14 (https://cutadapt.readthedocs.io/en/stable/) (Martin 2011). DADA2 version 1.6.0 (B. J. Callahan et al. 2016) and reference database info. Taxonomic classification was performed using the DADA2 implementation of the RDP bayesian classifier (Wang et al. 2007). The phylogenetic tree was generated following methods in (B. Callahan et al. 2016) using the DECIPHER R package version for multiple sequence alignment (Wright 2016) and the phangorn R package for tree construction (Schliep et al. 2017). For the QIIME pipelines all used the same input merged paired-end, quality filtered set of sequences (Caporaso et al. 2010). Both open and closed reference pipelines used the Greengenes 97% similiarity database for reference clustering. UCLUST alogrithm (version v1.2.22q) was used for clustering and taxonomic assignment against the Greengenes database version 13.8 97% similarity OTUs (Edgar 2010; McDonald et al. 2012). The phylogenetic tree was constructed using FastTree and a multiple sequence alignment generated using pyNAST and the Greengenes reference alignment (Greengenes info) (Caporaso et al. 2010; Price, Dehal, and Arkin 2010). Additionally, sequence variants were inferred from the QIIME merged and quality filtered sequences using the Deblur sequence inference clustering method (version 1.0.3) (Amir et al. 2017). The same taxonomic classification and phylogenetic tree construction methods used for the other QIIME pipelines were also used for the Deblur clustered sequence data.

3.3 Normalization Methods and Beta-Diversity Metrics

Normalization methods are used to account for differences in sampling depth, number of sequences generated per sample, across samples. Rarefaction, subsampling counts without replacement to an even abundance is a commonly used method in macro-ecology and 16S rRNA marker-gene surveys (Gotelli and Colwell 2001; Hughes and Hellmann 2005). Samples were rarified to four level; 2000, 5000, and 10000 total abundance per sample, and to the total abundance of the 15th percentile. Rarefaction levels were selected based on values commonly used in published studies (Thompson et al. 2017), other comparison studies (Weiss et al. 2017; McMurdie and Holmes 2014). Rarified count data was analyzed using both weighted and unweighted Beta-diversity metrics. Other normalization methods were only analyzed for weighted metrics as these methods would not impact unweighted metric results. Other normalization methods include those previously developed for normalizing microarray and RNAseq data that are commonly used to normalize 16S rRNA marker-gene survey including upperquartile (UQ), trimmed mean of M values (TMM), and relative log expression (Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012). Cumulative sum scaling (CSS) (Paulson et al. 2013) a normalization method developed specifically for 16S rRNA marker-gene survey data and total sum scaling (proportions, TSS) were also included in our weighted Beta-diversity metric assessment.

Weighted and unweighted phylogenetic and taxonomic beta diversity metrics were compared. Beta diversity metrics were calculated using phyloseq version 1.22.3 (McMurdie and Holmes 2013). Weighted and Unweighted UniFrac phylogenetic Beta-diversity metrics were calculated using the phyloseq implementation of FastUniFrac (McMurdie and Holmes 2013; Hamady, Lozupone, and Knight 2010). For our feature-level Beta-diversity assessment the Bray-Curtis weighted and Jaccard unweighted metrics were used (Bray and Curtis 1957; Jaccard 1912).

3.4 Beta-Diversity Assessment

- technical artifacts
  To quantify the contribution of biological and technical variability to total variability the distribution of beta diversity dissimilarity metrics were compared between individuals, within individual between conditions (pre- and post-exposure), and different types of technical replicates.
  ADD STATS

We assessed the impact of bioinformatic pipeline and normalization methods on beta-diversity metrics using a similar approach used in (Reynolds et al. 2006; McMurdie and Holmes 2014). To summarize performance across titration comparisons, area under the curve was calculated using the trapazoid method (**R package ref**). This assessment evaluated the ability to differentiate titrations and post-exposure samples from pre-exposure samples.

4 Results

The beta diversity assessment framework includes three components; (1) evaluating beta diversity between PCR replicates for sequencing runs with different error rates and library sizes, (2) difference in beta diversity between biological and technical replicates, and (3) ability to differentiate between PCR replicates from unmixed pre-exposure samples and post-exposure samples as well as the unmixed post-exposure samples.
For our assessment we used a dataset consisting of two-sample titrations of DNA extracts from five vaccine trial participant stool samples collected before and after exposure to pathogenic *E. coli* (**Experimental Design Figure**).

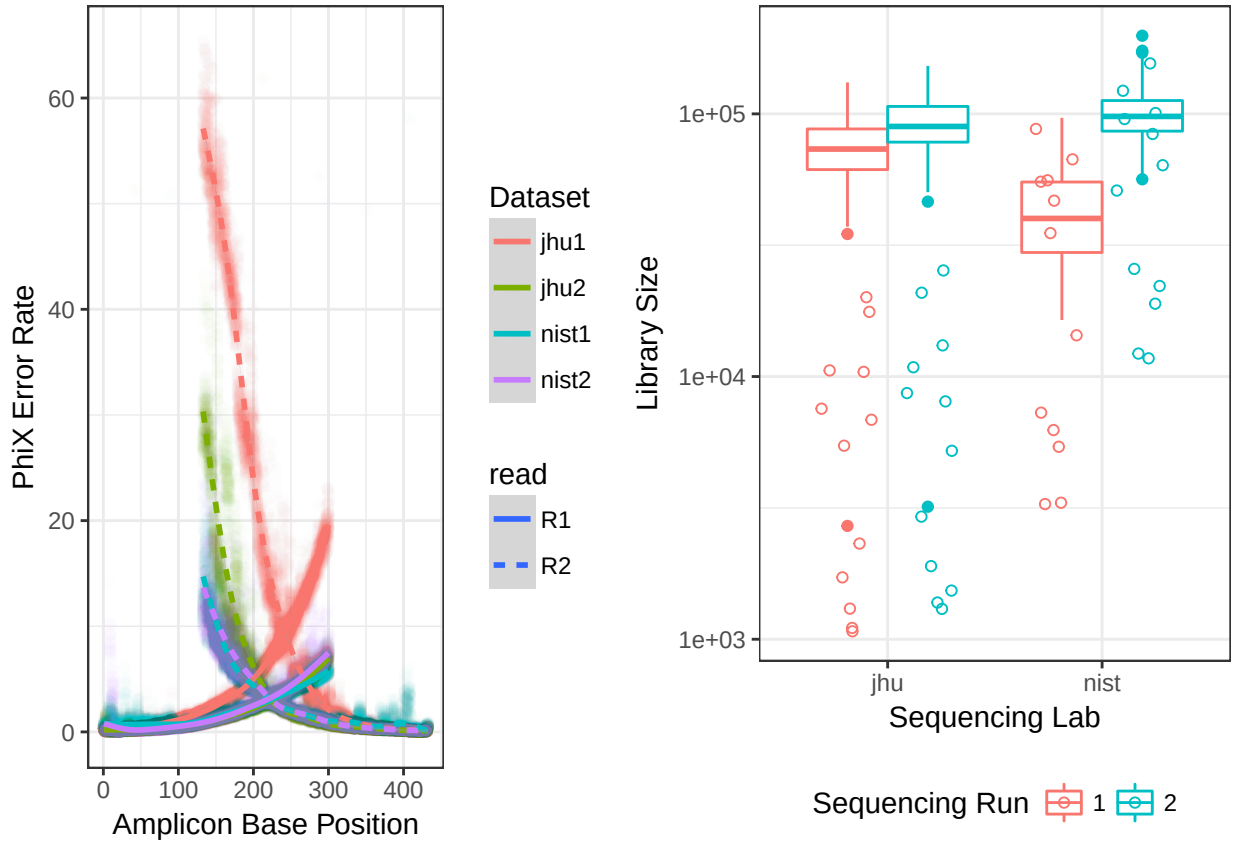

Figure 1: A. PhiX error rate relative to 16S rRNA amplicon base position for initial and the four sequencing runs. B. Distribution in the number of reads per barcoded sample (Library Size) by sequncing laboratory and sequencing run. Negative control library size is indicated by hollow points. Negative controls were not included in the boxplots.

Table 1: Summary statistics for the different bioinformatic pipeliens. Four pipelines, de novo, open reference, closed reference, and deblur (sequence inference), used the sample sequence pre-processing methods. DADA2 is a denoising sequence inference pipeline and mothur is a de-novo clustering pipeline. No template controls were excluded from summary statistics. Sparsity is the proportion of 0's in the count table. Features is the total number of OTUs (QIIME and mothur) or SVs (DADA2) in the count. Sample coverage is the median and range (minimum - maximum) per sample total feature abundance. Filter rate is the proportion of reads that were removed while processing the sequencing data for each bioinformatic pipeline.

Pipelines	Features	Samples	Sparsity	Total Abundance	Pass Rate
dada	25247	768	0.991	52356 (141585-181)	0.76 (0.87-0.01)
deblur	3711	576	0.940	9135 (30423-4)	0.14 (0.24-0)
mothur	38367	765	0.992	13312 (42954-171)	0.2 (0.45-0.02)
qiimeClosedRef	6184	754	0.929	24938 (111765-1)	0.36 (0.73-0)
qiimeDeNovo	180834	766	0.994	26250 (118767-4)	0.37 (0.75-0)
qiimeOpenRef	45663	766	0.981	26373 (118421-3)	0.37 (0.75-0)

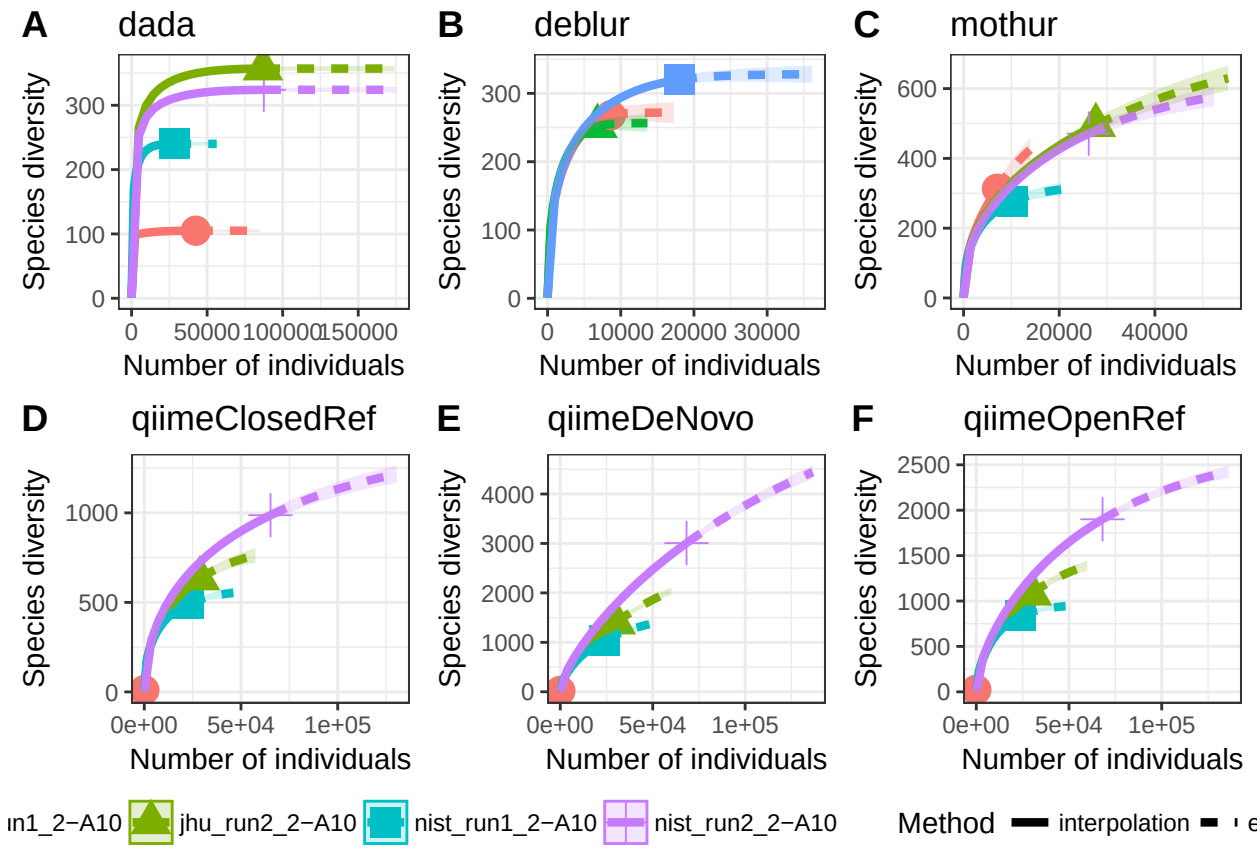

Figure 2: Rarefaction curves for example sample across pipelines and sequencing runs.

4.1 Dataset Characteristics

Bioinformatic pipelines and normalization methods are used reduce the impact of noise in marker gene sequencing data due to sequencing errors and differences in the library size between samples. Sequencing data for the two-sample titration dataset was obtained from four replicate sequencing runs with different sequence quality and library size variability (Fig. 1, *Supplemental DADA2 qual plot*). The first JHU run had higher PhiX error rates compared to the other sequencing runs especially for the reverse reads (Fig. 1A). NIST runs had lower error rates compared to the JHU runs but greater variabiltiy in library size (Fig. 1B). Good separation between sample and no template control library size for JHU but not NIST samples. However, total abundance is lower for samples compared to no template controls for most sequencing runs and samples. Though a few no template controls have values within the sample range. (see pipe characterization total abundance and pass rate plots).

The sequence data was processed using six different bioinformatic pipelines, DADA2, mothur, Deblur, and QIIME - *de novo*, open-reference, closed-reference. Pipelines employ different pre-processing, clustering, and quality filtering methods, as a result the features and count tables generated by the pipelines exhibit different characteristics in terms of the number of features, total abundance, and proportion of sequences passing quality control (Table 1). Rarefaction curves are in ecology to determine how well a community has been sampled (**REF**). Measurement methods prone to errors, such as marker-gene sequencing, will never reach the asympotote if errors are not appropriately accounted for in sample processing (**REF Chao**). Sequence inference methods have lower species diversity estimates and reach asympotote, whereas *de novo*, open-reference, and closed-reference methods do not (Fig. 2).

- De novo highest rate of artifacts (due to lack of singleton filtering)
  - De novo steepest slope in rarefaction curves
- DADA2 and Deblur plateu around the same level
- DADA2 inconsistent across sequencing runs, artificial plateau
- Mothur and Deblur consistent across sequencing runs
- qiime open ref, closed ref, and de novo richness dependent on both quality and library size

4.2 Technical Artifacts

Key Points

- JHU run 1 had lower sequence quality (boxplots) (Fig. 4).
- NIST run 1 had greater variabiltiy in sample total abundance (boxplots) (Fig. 3).
- Mean pairwise distance was greater for qiime De-novo, open and closed reference pipelines for JHU1 relative to the other sequencing runs for all metrics excluding weighted unifrac (Fig. 5).
- Qiime De novo had high pairwise distance across sequencing runs for Unifrac and low for weighted Unifrac.
- Pairwise distances varies by metric.
- DADA2 pairwise distances greater for NIST1 and NIST2 compared to JHU runs, JHU2 had the lowest pairwise distance.
- Mothur and dada, consistent results for JHU runs, better able to account for sequencing errors.

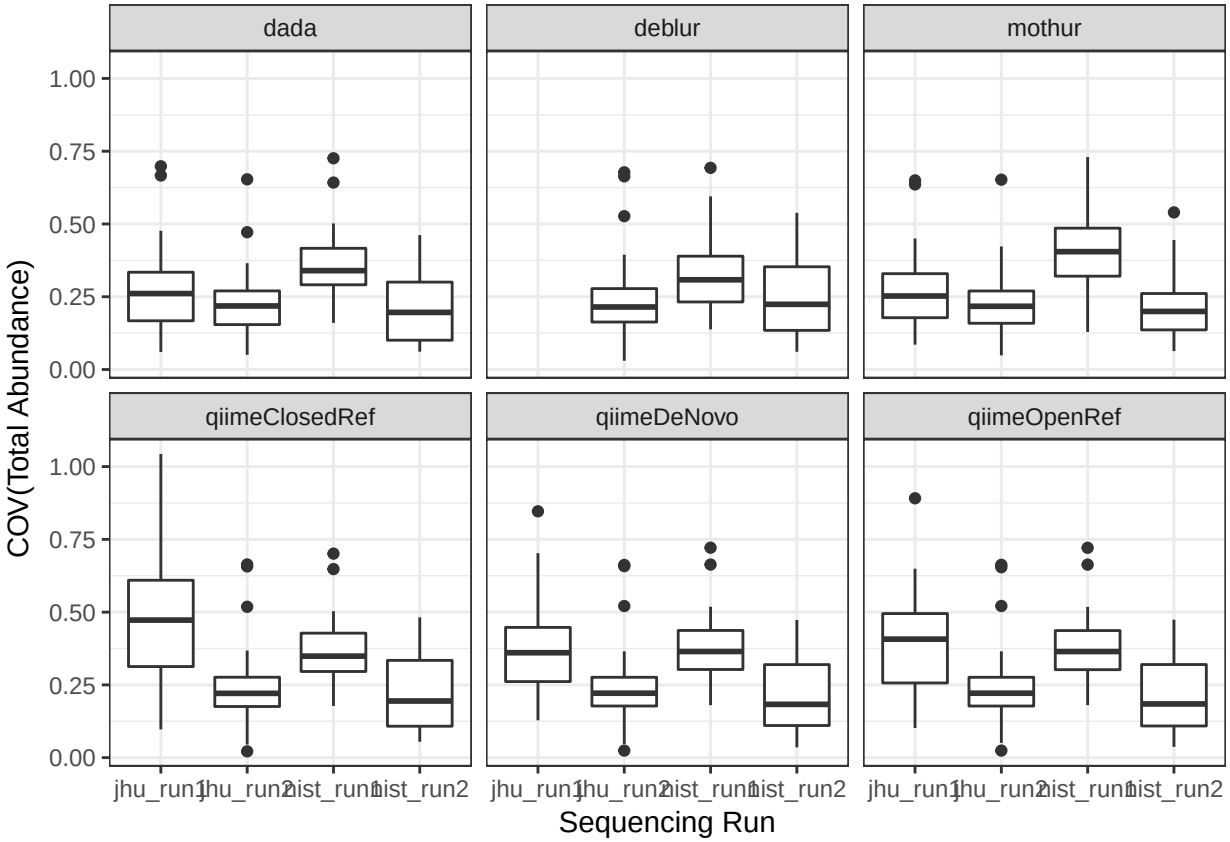

Figure 3: Distribution of the coefficienct of variation of total abundance for PCR replicates by bioinformatic pipeline and sequencing run across.

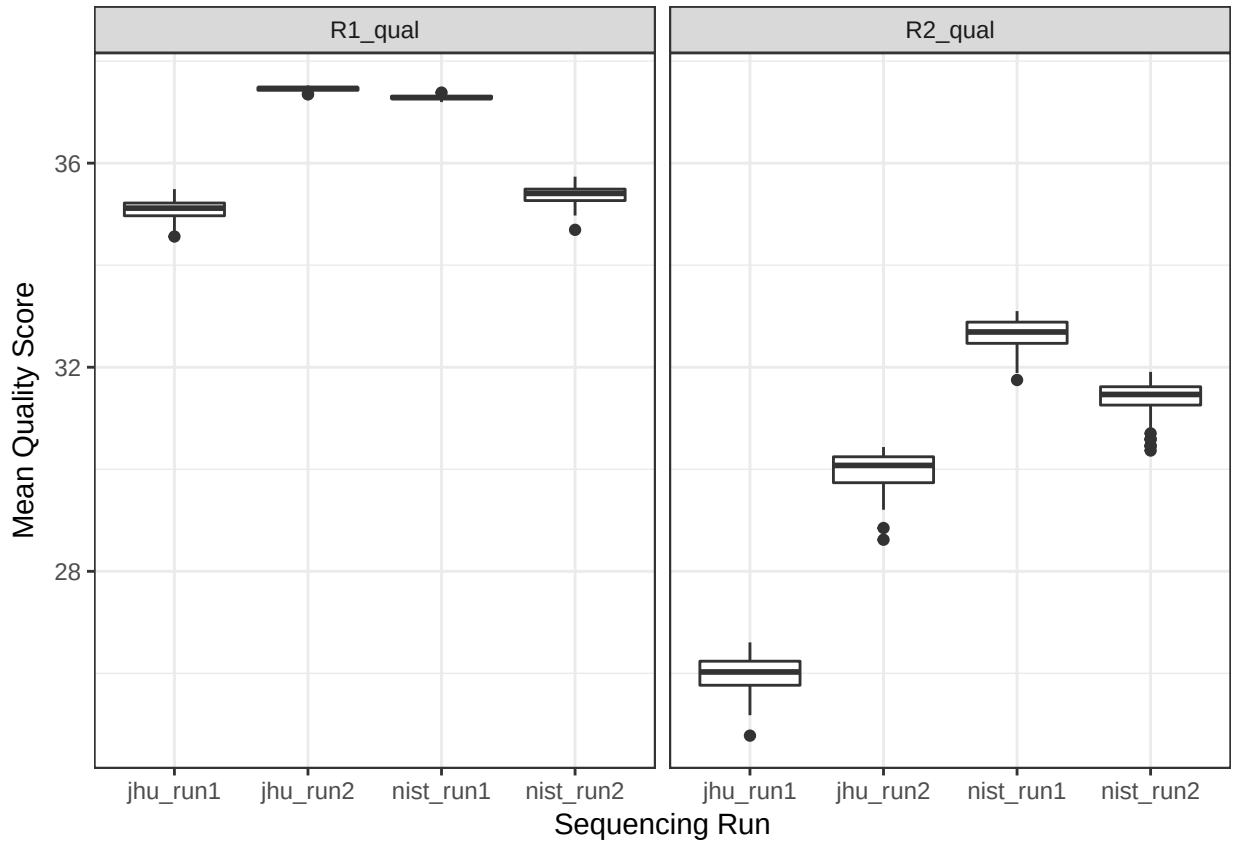

Figure 4: Mean read quality score for PCR replicates across sequencing runs for forward and reverse reads.

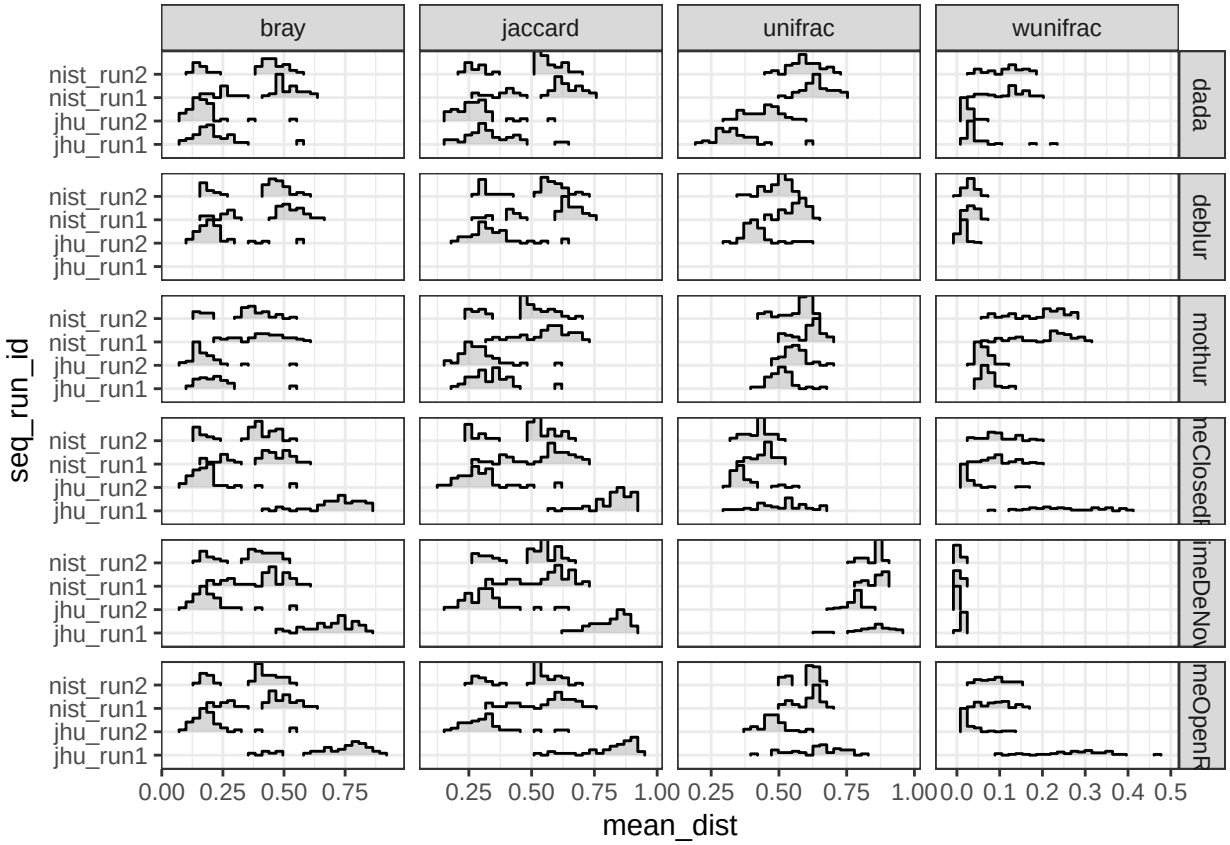

Figure 5: Distribution of mean pairwise beta diversity for PCR replicates by sequencing run and pipeline for raw count data.

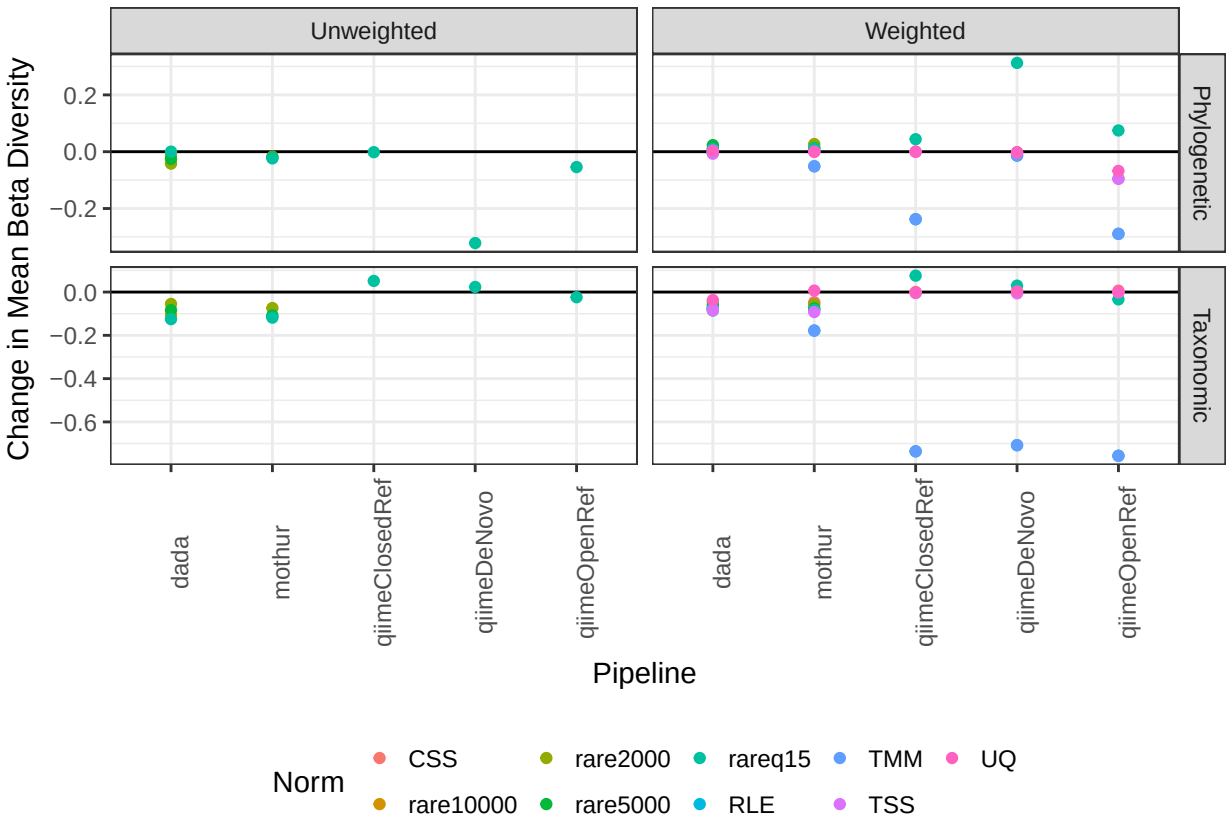

Figure 6: JHU1 - Impact of normalization method on mean beta diversity between pcr repliates.

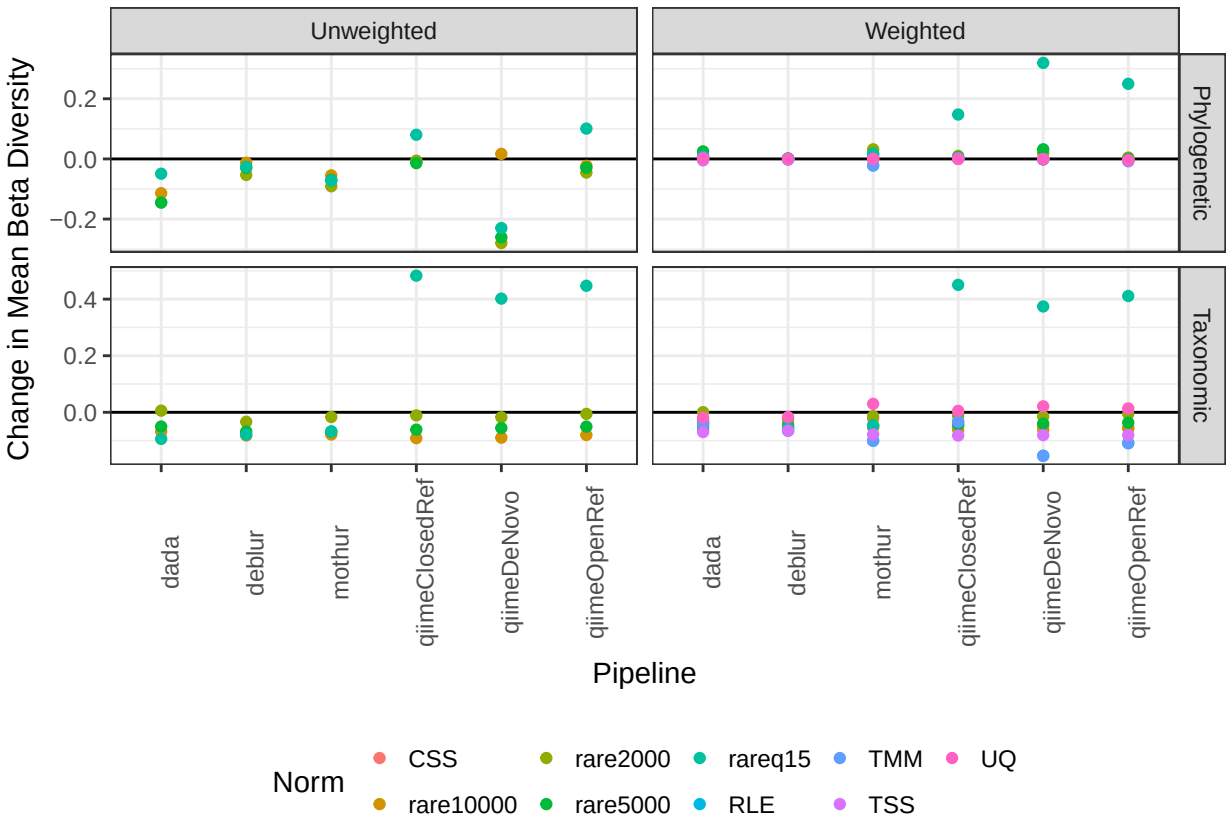

Figure 7: JHU2 - Impact of normalization method on mean beta diversity between pcr repliates.

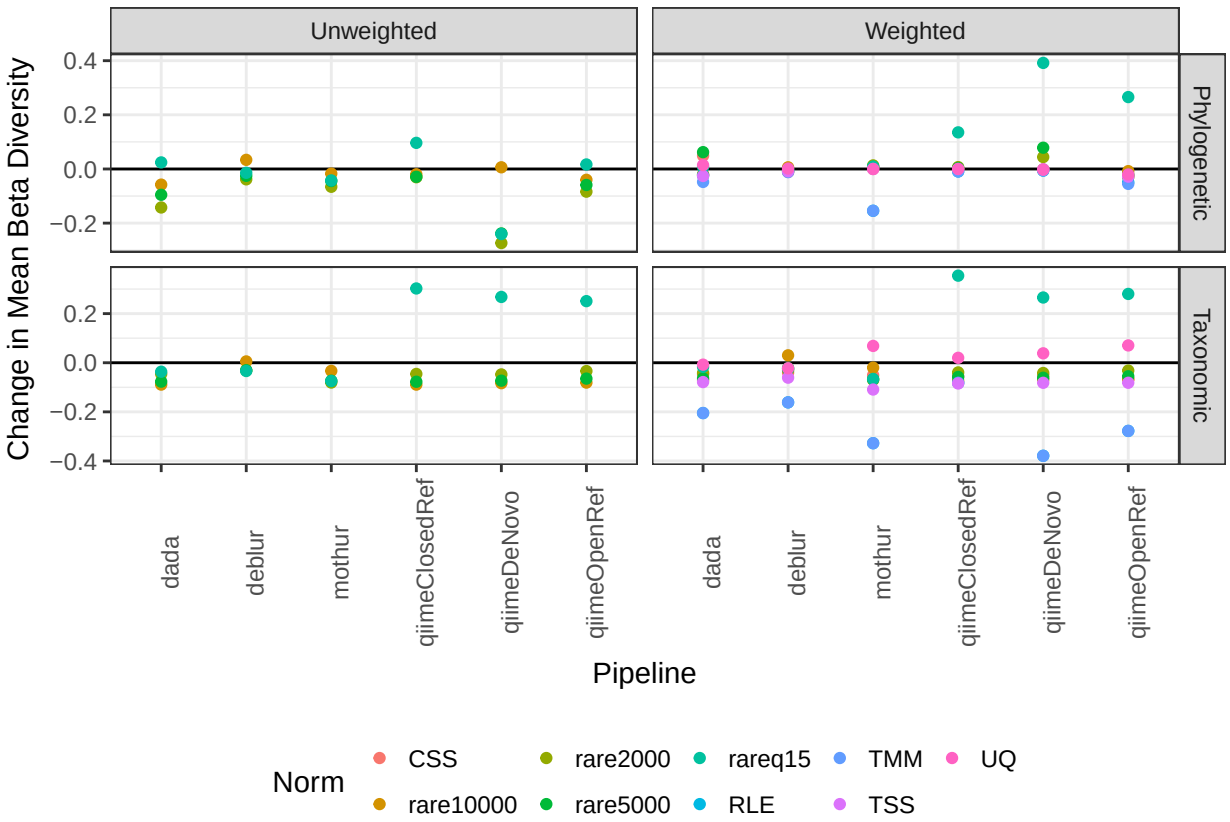

Figure 8: NIST1 - Impact of normalization method on mean beta diversity between pcr repliates.

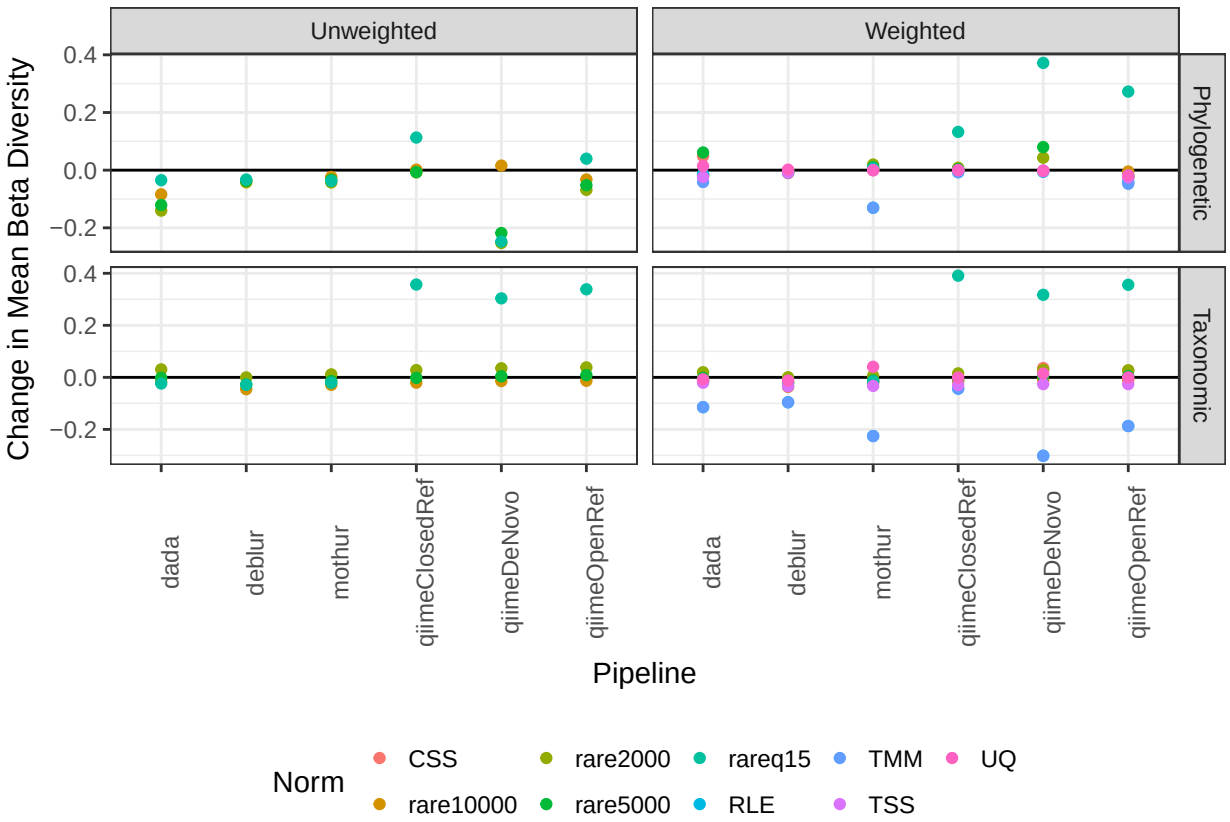

Figure 9: NIST2 - Impact of normalization method on mean beta diversity between pcr repliates.

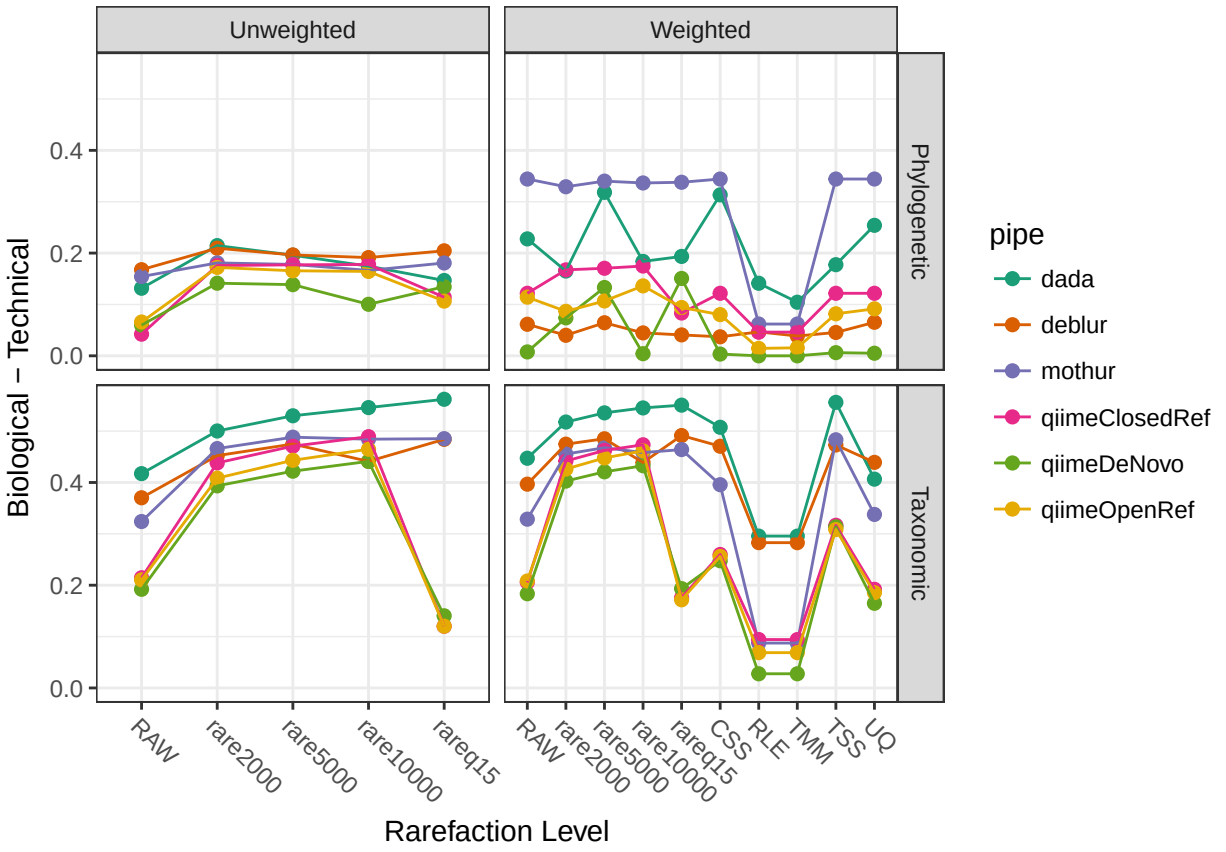

Figure 10: Biological vs. Technical Variation, y-axis is the differences between the mean biological and technical variation (pairwise distance between replicates.)

- Deblur pipeline failed for JHU1.
- Higher pairwise distances for NIST runs indicate diversity metrics impacted by larger variation in total abundance across PCR replicates.

Additional Points

- De novo weighted Unifrac low for QIIME De novo: This is potentially due to large number of singletons in weighted unifrac dataset, ~120K out of ~180K total features. These singletons are likely sequencing errors and therefore closely related to other taxa therefore minimally impact the weighted unifrac results.

4.3 Biological v. Technical Variation

Beta-diversity distances between biological and technical replicates varies by pipeline and beta-diversity metric (Fig. 10).
For weighted metrics RLE and TMM decreased the difference relative to raw counts indicating that for beta diversity analysis these normalization methods reduce the power to distingush true biological differences from technical variability or noise.
In general rareifiy count data had higher mean difference excluding when rarefying to 15th quantile for qiime de novo, closed , and open-reference, this is potentially due to sample loss.

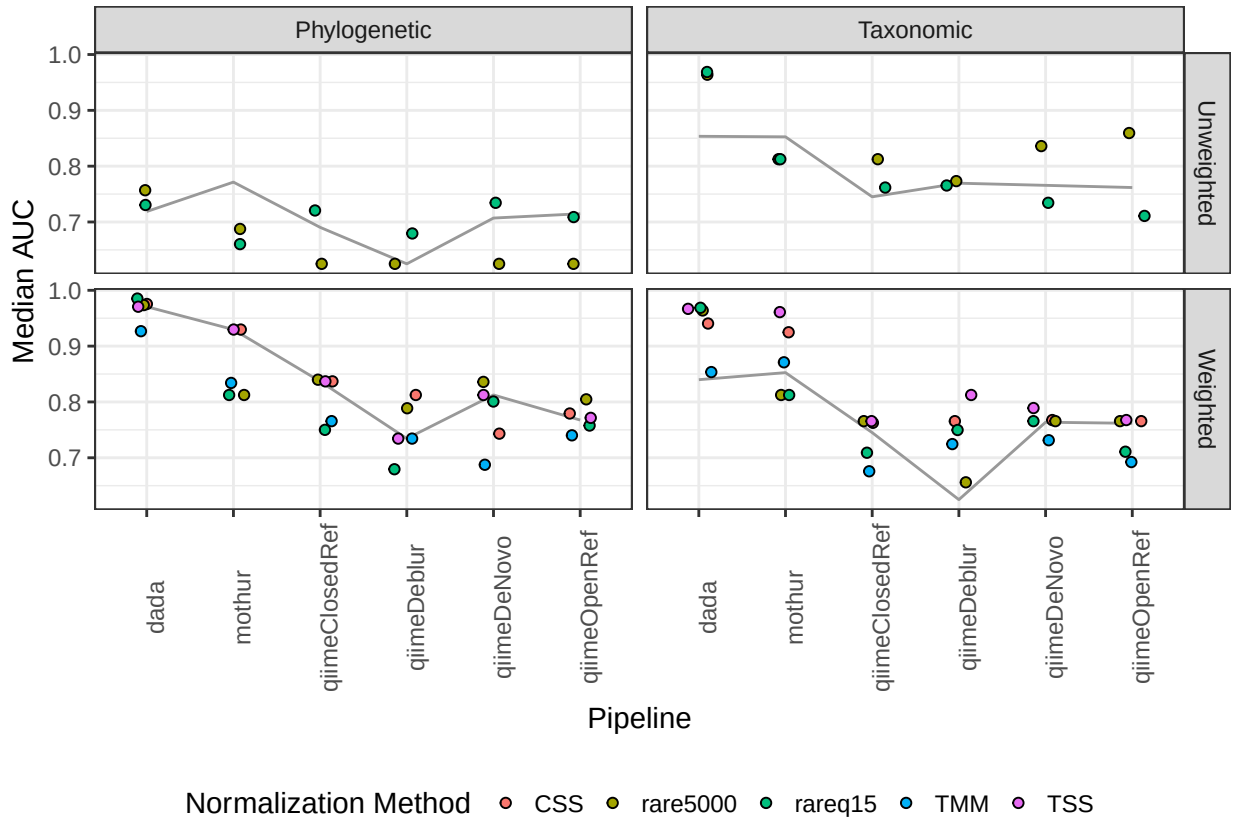

Figure 11: Comparison of median AUC for clustering results across pipelines and normalization methods for four beta diversity metrics. Grey line indicates, the median AUC for unnormalized, raw, count table values. Points above the grey line are normalization methods that improve performance and below are methods that decrease performance.

4.4 Comparison to Expectation

Performance varied by pipeline with DADA2 having consistently higher performance compared to the other pipelines (Fig. 11).
Rarefaction level had inconsistent performance relative to unnormalized data. Rareified to the 15th quantile library size improved performance relative to unnormalized data with qiime pipelines when using UniFrac but lower performance for Jaccard.
For weighted metrics, normalization method performance relative to unnormalized counts varied by pipeline, though TMM and rarefaction to 15th quantile had consistently lower performance compared to unnormalized data.

5 Discussion

Dataset

- samples

- sequence characteristics
  - Four sequencing runs vary in error rates and number of reads per sample.
- Pipelines

  - Pipelines vary in ability to differentiate true sequences from sequencing artifacts.
  - De novo rarefaction curves, singletons, and sparsity - high false positive rate

  - DADA2 - rarefaction plateu at different points for individual runs - high false negative rate

Technical artifacts

- Impact of sequence quality and variation in number of reads on diversity metric repeatability, mean beta diversity between PCR replicates, is pipeline and diversity metric dependent.
  - De novo high unweighted unifrac for all runs but low weighted unifrac, attributed to singletons, in ability to group sequencing artifacts with true biological sequences.

  - Low error rate and read number variability had consistently better repeatability.

  - Normalization methods help increase repeatablity, excluding rarefying data to 15th quantile, which decreased repeatability especially for QIIME pipelines. TMM improved weighted beta diversity repeatabilty for NIST datasets, greater varability in library size.

Bio V. Tech

- Difference in beta diverity between biological samples (individuals and exposure) and technical replicates (sequencing runs) varied by diversity metric and pipeline.

- Normalization method impact varied by pipeline and diversity metric.
- Rareifying data to 15th quantile decreased the ability to differentiate between biological and technical replicates.
- RLE and TMM similarly decreased the difference in beta-diversity between biological and technical replicates, especially for Bray Curtis (weighted taxonomic diversity metric.)

comp to exp

- Evaluate the ability to distinguish between biological samples with varying levels of similarity

- Results varied by pipeline and diversity metric, with DADA2 and mothur consistently out performing the other methods.

- For weighted phylogenetic methods normalization methods rarely improved the results, but improved the results in most cases for weighted taxonomic diversity metrics.
- Inconsistent results when using rarefyed data and unweighted metrics. Results were pipeline and diversity metric dependent.

Conclusions

- When you have data with low error rates and variability in number of reads, consistent pipeline performance.

- Pipelines vary in ability to distingush sequencing artifacts from true biological sequences.

- These differences impact the beta diversity estimate repeatabiltiy.

- Normalization can help improve repeatabiltiy, but sometimes at the cost of decreasing the difference between biological signal and technical variability.

- Mothur and dada2 are better able to handle lower quality datasets.
- Normalization methods can improve ability to detect true biological signal though normalization methods developed for gene expression methods may not be appropriate.

Other thoughts/ ideas

- Bio v. tech: only unmixed pre and between individuals

6 References

Amir, Amnon, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, James T Morton, Zhenjiang Zech Xu, Eric P Kightley, et al. 2017. "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns." *mSystems* 2 (2).

Bray, J Roger, and J T Curtis. 1957. "An Ordination of the Upland Forest Communities of Southern Wisconsin." *Ecol. Monogr.* 27 (4). Ecological Society of America:325–49.

Calder, R. Brent. 2015. *SavR: Parse and Analyze Illumina Sav Files*. https://github.com/bcalder/savR.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13:581–83. https://doi.org/10.1038/nmeth.3869.

Callahan, BJ, K Sankaran, JA Fukuyama, PJ McMurdie, and SP Holmes. 2016. "Bioconductor Workflow for Microbiome Data Analysis: From Raw Reads to Community Analyses [Version 2; Referees: 3 Approved]." *F1000Research* 5 (1492). https://doi.org/10.12688/f1000research.8986.2.

Caporaso, J. Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (April). Nature Publishing Group SN -:335 EP. http://dx.doi.org/10.1038/nmeth.f.303.

Cole, James R, Qiong Wang, Jordan A Fish, Benli Chai, Donna M McGarrell, Yanni Sun, C Titus Brown, Andrea Porras-Alfaro, Cheryl R Kuske, and James M Tiedje. 2014. "Ribosomal Database Project: Data and Tools for High Throughput rRNA Analysis." *Nucleic Acids Res.* 42 (Database issue):D633–42.

Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster Than BLAST." *Bioinformatics* 26 (19):2460–1.

Gotelli, Nicholas J, and Robert K Colwell. 2001. "Quantifying Biodiversity: Procedures and Pitfalls in the Measurement and Comparison of Species Richness." *Ecol. Lett.* 4 (4). Blackwell Science Ltd:379–91.

Hamady, Micah, Catherine Lozupone, and Rob Knight. 2010. "Fast UniFrac: Facilitating High-Throughput Phylogenetic Analyses of Microbial Communities Including Analysis of Pyrosequencing and PhyloChip Data." *ISME J.* 4 (1):17–27.

Hughes, Jennifer B, and Jessica J Hellmann. 2005. "The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity." In *Methods in Enzymology*, 397:292–308. Academic Press.

Jaccard, Paul. 1912. "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1." *New Phytol.* 11 (2). Blackwell Publishing Ltd:37–50.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal* 17 (1):10–12.

McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth. 2012. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Res.* 40 (10):4288–97.

McDonald, Daniel, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight, and Philip Hugenholtz. 2012. "An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea." *ISME J.* 6 (3):610–18.

McMurdie, Paul J, and Susan Holmes. 2013. "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PLoS One* 8 (4):e61217.

———. 2014. "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible." *PLoS Comput. Biol.* 10 (4):e1003531.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." *Nat. Methods* 10 (12):1200–1202.

Price, Morgan N, Paramvir S Dehal, and Adam P Arkin. 2010. "FastTree 2–Approximately Maximum-Likelihood Trees for Large Alignments." *PLoS One* 5 (3):e9490.

Reynolds, A P, G Richards, B de la Iglesia, and V J Rayward-Smith. 2006. "Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms." *J. Math. Model. Algorithms* 5 (4). Springer Netherlands:475–504.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1):139–40.

Schliep, Klaus, Potts, Alastair J., Morrison, David A., Grimm, and Guido W. 2017. "Intertwining Phylogenetic Trees and Networks." *Methods in Ecology and Evolution* 8 (10):1212–20. https://doi.org/10.1111/2041-210X.12760.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." *Appl. Environ. Microbiol.* 75 (23):7537–41.

Sheneman, Luke, Jason Evans, and James A Foster. 2006. "Clearcut: A Fast Implementation of Relaxed Neighbor Joining." *Bioinformatics* 22 (22):2823–4.

Souza, Welliton, and Benilton Carvalho. 2017. *Rqc: Quality Control Tool for High-Throughput Sequencing Data*. https://github.com/labbcb/Rqc.

Thompson, Luke R, Jon G Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J Locey, Robert J Prill, et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity." *Nature* 551 (7681):457–63.

Wang, Qiong, George M Garrity, James M Tiedje, and James R Cole. 2007. "Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy." *Appl. Environ. Microbiol.* 73 (16):5261–7.

Weiss, Sophie, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, et al. 2017. "Normalization and Microbial Differential Abundance Strategies Depend Upon Data Characteristics." *Microbiome* 5 (1):27.

Westcott, Sarah L, and Patrick D Schloss. 2017. "OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units." *mSphere* 2 (2).

Wright, Erik S. 2016. "Using Decipher V2.0 to Analyze Big Biological Sequence Data in R." *The R Journal* 8 (1):352–59.