Metagenome Analysis of Premature Birth

Jaewoong Lee Semin Lee

Department of Biomedical Engineering
Ulsan National Institute of Science and Technology

jwlee230@unist.ac.kr

2021-05-24

Overview

Introduction

Microbiome

- Microbiota: the microorganisms which live inside & on humans (Turnbaugh et al., 2007)
- Microbiome: 10^{13} to 10^{14} microorganisms whose which collective genome (Gill et al., 2006)

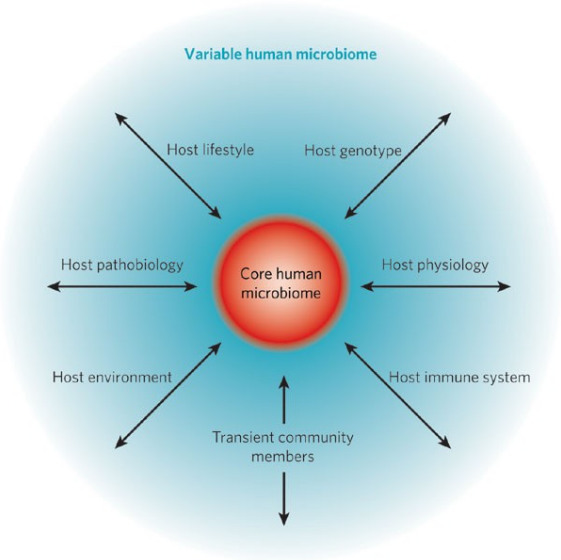


Figure: Concept of a core human microbiome (Turnbaugh et al., 2007)

rRNA

- Ribosomal RNA
- Well-known as a key to phylogeny (Olsen & Woese, 1993)

Premature Birth (Preterm Birth)

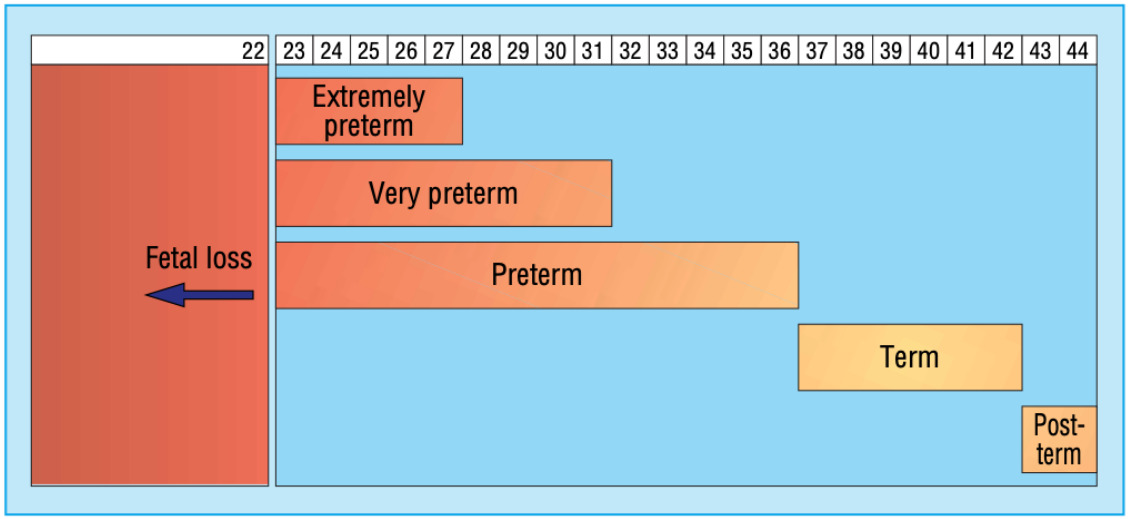


Figure: Definitions of Premature (J. Tucker & McGuire, 2004)

$\therefore$ Hence, in this study,

- Premature: < 37 weeks
- Normal: ≥ 37 weeks

Materials

16S rRNA Sequencing

16S rRNA sequencing is the *reference method* for bacterial taxonomy & identification (Mignard & Flandrois, 2006)
Three main reasons (Janda & Abbott, 2007):

- 16S rRNA exists in almost all bacteria
- Functions of the 16S rRNA has not changed over time
- 16S rRNA is large enough for bioinformatics

Train/Test Data vs. Validate Data

- JBNU/Helixco data
  - First data
  - Second data
  - Stool data

Table: Sample Information

Data	Participants	Samples	Remarks
First	24	107	-
Second	35	288	-
Stool	63	126	Stool

Methods

Methods

Qiime 2 Workflow

Qiime 2 Workflow

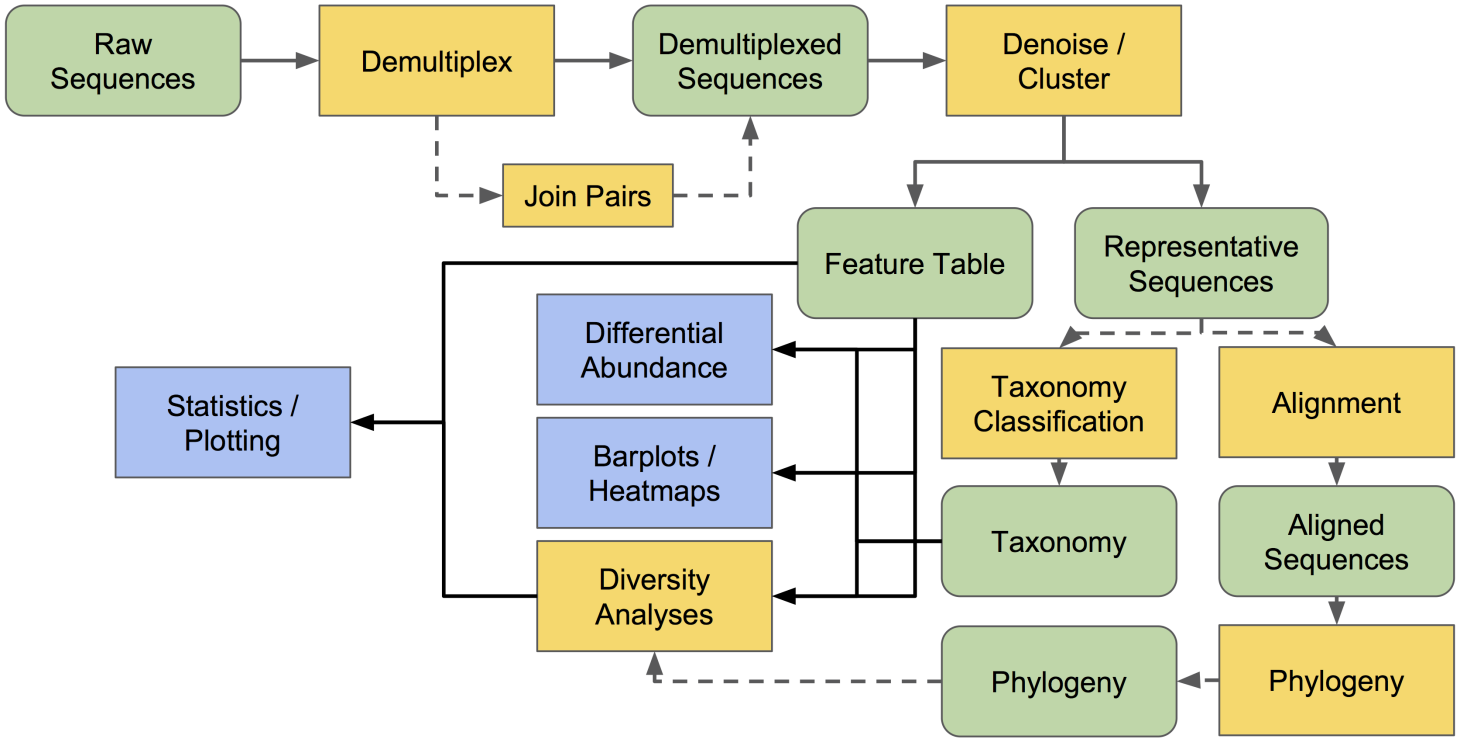


Figure: QIIME 2 workflow (Bolyen et al., 2019; Mandal, Van Treuren, White, Eggesbø, et al., 2015; McDonald et al., 2012)

Filitering with Quality Score

Drawback between:

- Longer sequence read
- Higher quality value

$\therefore$ Select the maximum length n where:

$$\begin{gathered} \forall n_i \in \{n_k | \text{MedianQualityScore} \geq 30\} \\ \exists! n \in \{n_i\} : n \geq n_i \end{gathered} \tag{1}$$

Denoising Techniques

- DADA2: Amplicon Sequence Variants (ASVs) (Callahan et al., 2016)
- Deblur: Operational Taxonomic Units (OTUs) (Amir et al., 2017)

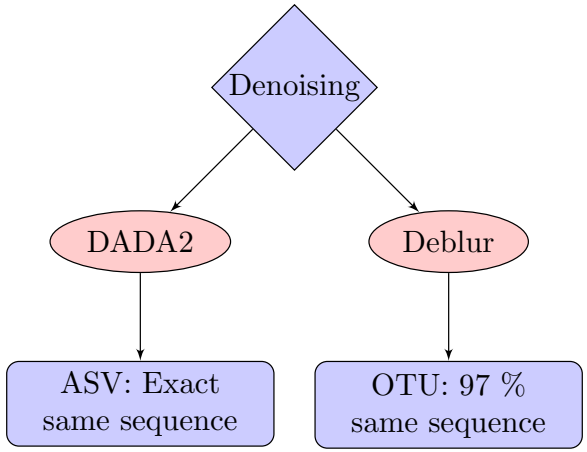


Figure: Denoising Algorithms

Taxonomy Classification

- Greengenes (GG) (DeSantis et al., 2006)
- SILVA (Pruesse et al., 2007; Quast et al., 2012)

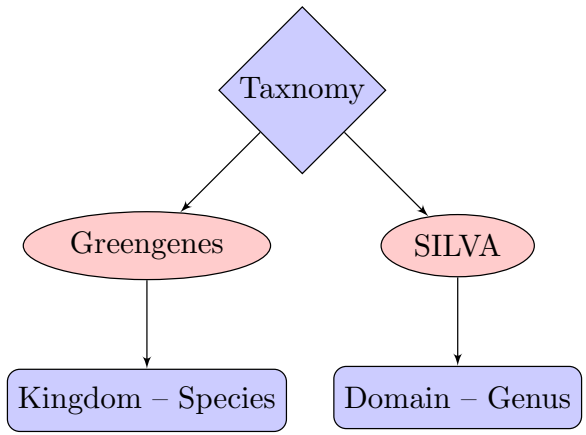


Figure: Taxonomy Classifications

"A **higher** performance at taxonomic levels above *genus level*;
but performance appears to **drop** at *species level*" (Gihawi et al., 2019)

Merging Denoising/Taxonomy

Merging multiple IDs (ASVs or OTUs) into one, which have

- Different IDs
- Identified as same taxonomy

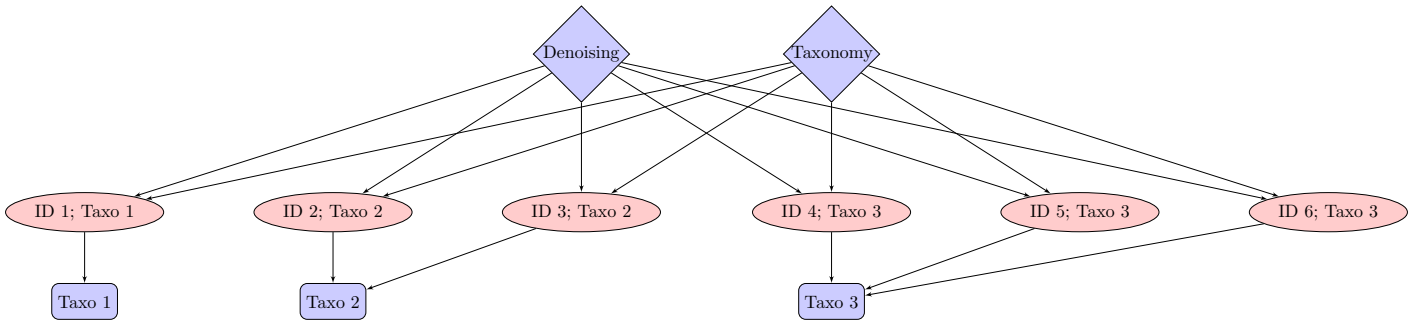


Figure: Example Diagram for Merging Denoising/Taxonomy

Methods

Abundance Test

ANCOM

- Analysis of composition of microbiome (Mandal, Van Treuren, White, Eggesbø, et al., 2015)
- ANCOM detects significantly abundant taxa, while maintain high statistical power
- Find taxa that can divide each classes

Methods

Diversity Indices

Diversity Indices

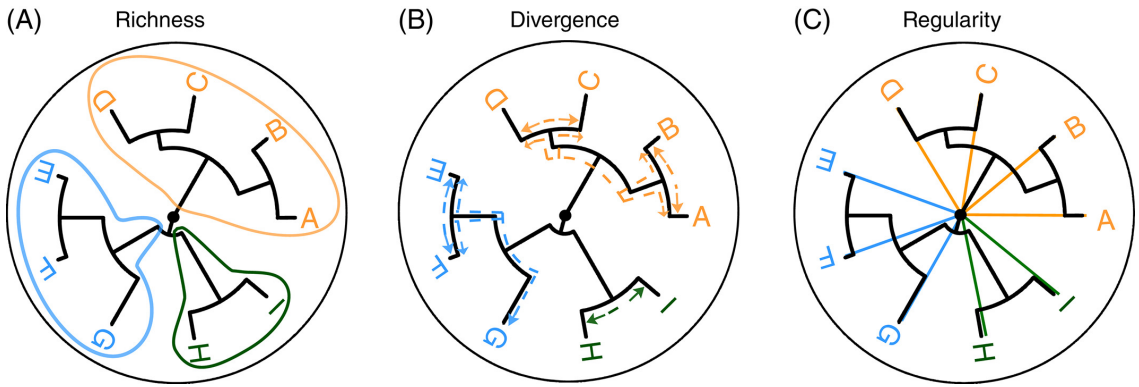


Figure: Three dimensions of phylogenic information (C. M. Tucker et al., 2017)

- A quantitative measure that shows richness, divergence, and regularity (C. M. Tucker et al., 2017)
- Alpha diversity indices: the richness of taxa **at a single community**
- Beta diversity indices: the taxonomic differentiation **between communities**

Alpha Diversity Indices

- Evenness index
- Faith's Phylogenetic Diversity (Faith PD) index
- Oberseved Features index
- Shannon's Diversity index

Beta Diversity Indices

- Bray-Curtis distance index
- Jaccard distnace index
- Unweighted UniFrac distance index
- Weighted UniFrac distance index

Methods

Miscellaneous

t-distributed Stochastic Neighbor Embedding (t-SNE)

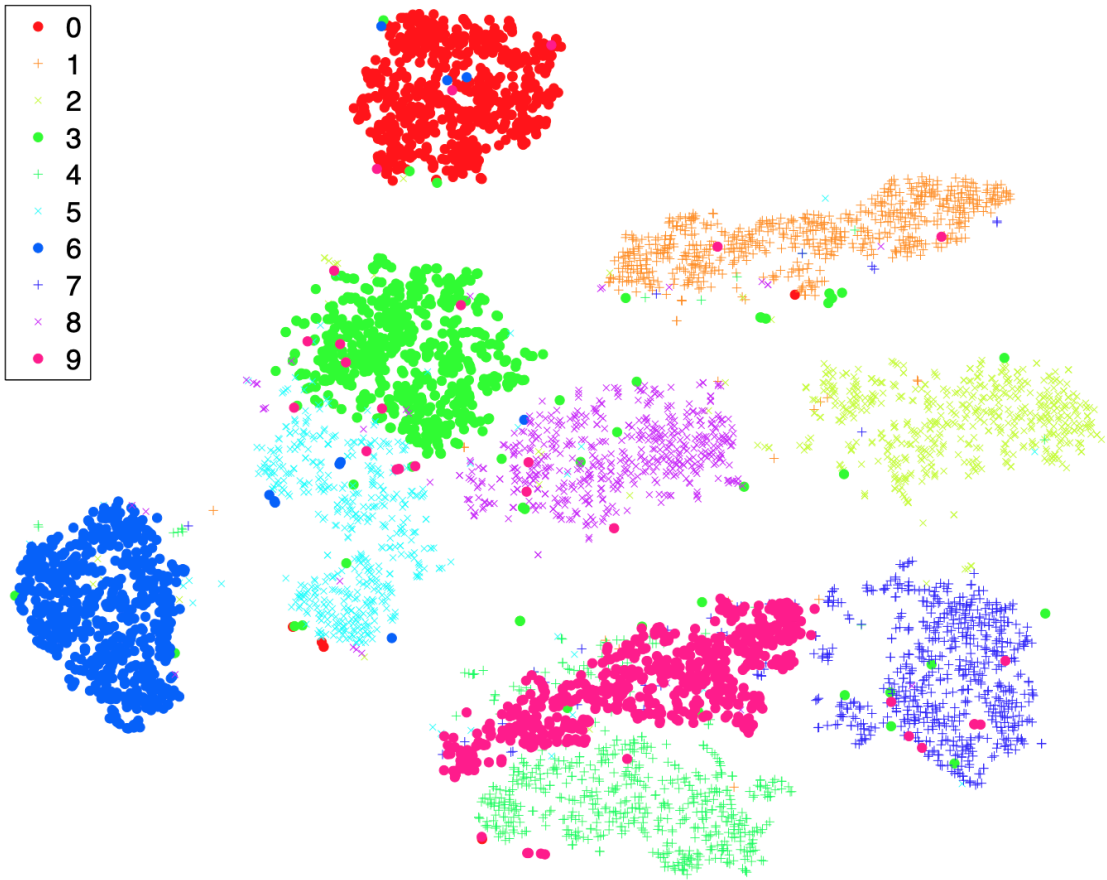


Figure: t-SNE with handwritten data (Maaten & Hinton, 2008)

Python Packages

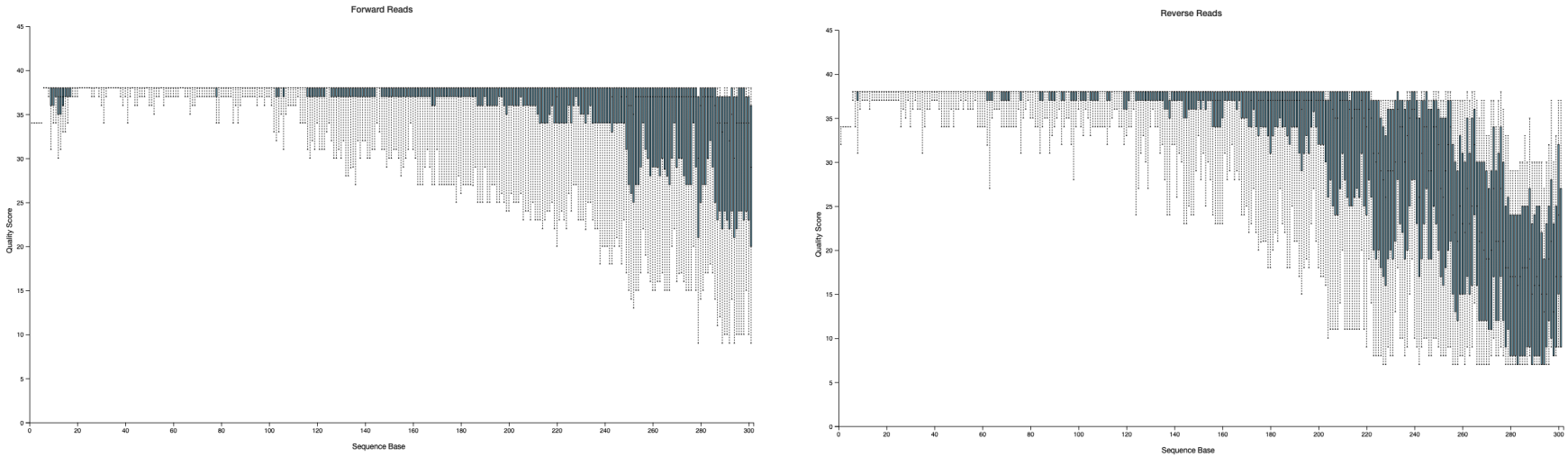
- Pandas (McKinney et al., 2011)
- Scikit-Learn (Pedregosa et al., 2011)
- SciPy (Virtanen et al., 2020)
- Matplotlib (Hunter, 2007)
- Seaborn (Waskom et al., 2020)
- Statannot

Results

Results

Filtering Results

Quality Score from First Data

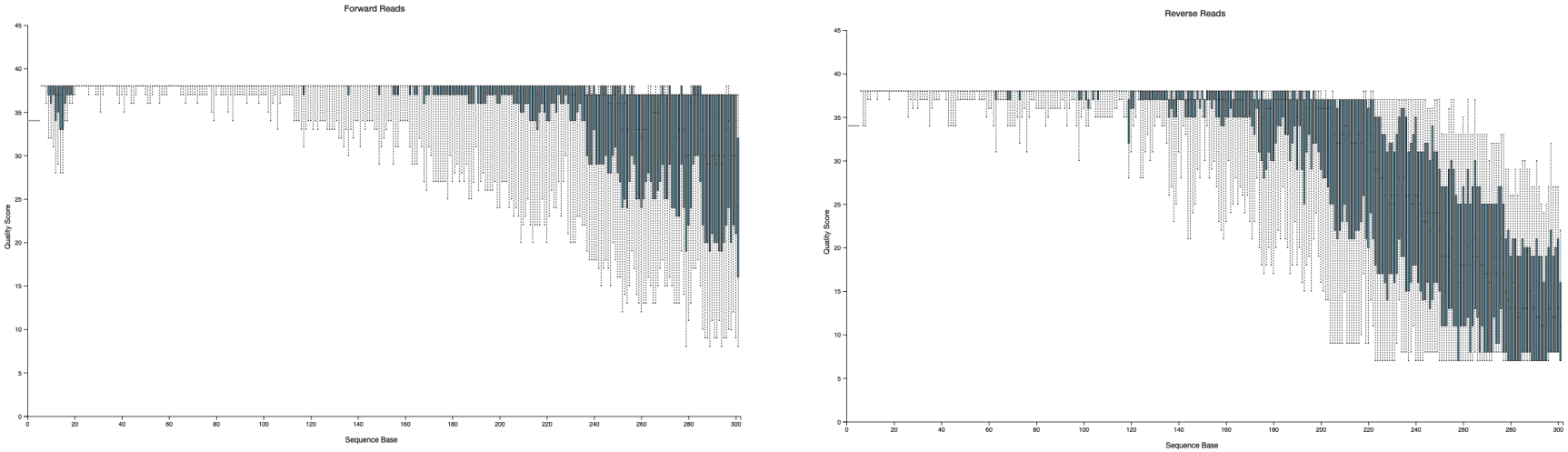


(a) Forward (b) Reverse

Figure: Sequence Quality Plot from Helixco Data

Maximum Length: $n_{Forward} = 300$, $n_{Reverse} = 265$

Quality Score from Second Data

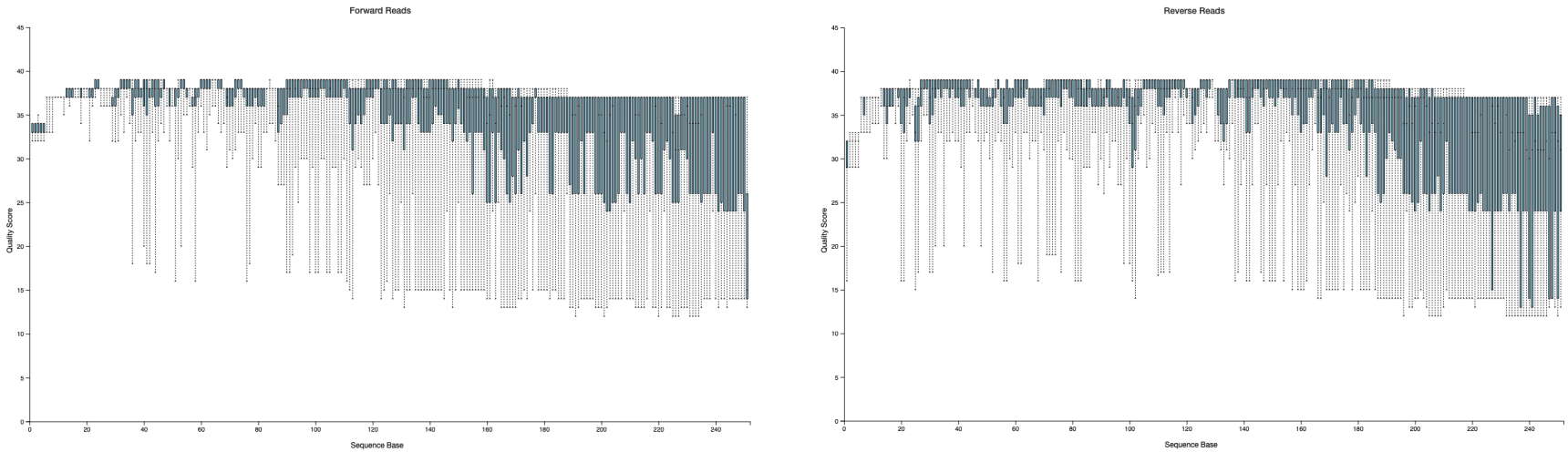


(a) Forward (b) Reverse

Figure: Sequence Quality Plot from Helixco Data

Maximum Length: $n_{Forward} = 300$, $n_{Reverse} = 222$

Quality Score from Stool Data



(a) Forward (b) Reverse

Figure: Sequence Quality Plot from Stool Data

Maximum Length: $n_{Forward} = 250$, $n_{Reverse} = 251$

Results

t-SNE with Site/Premature Information

Workflow for t-SNE with Site/Premature Information

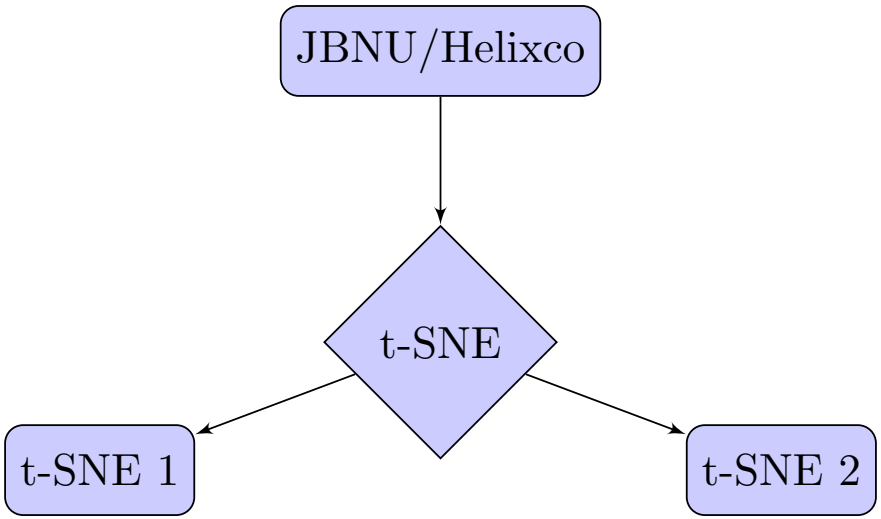
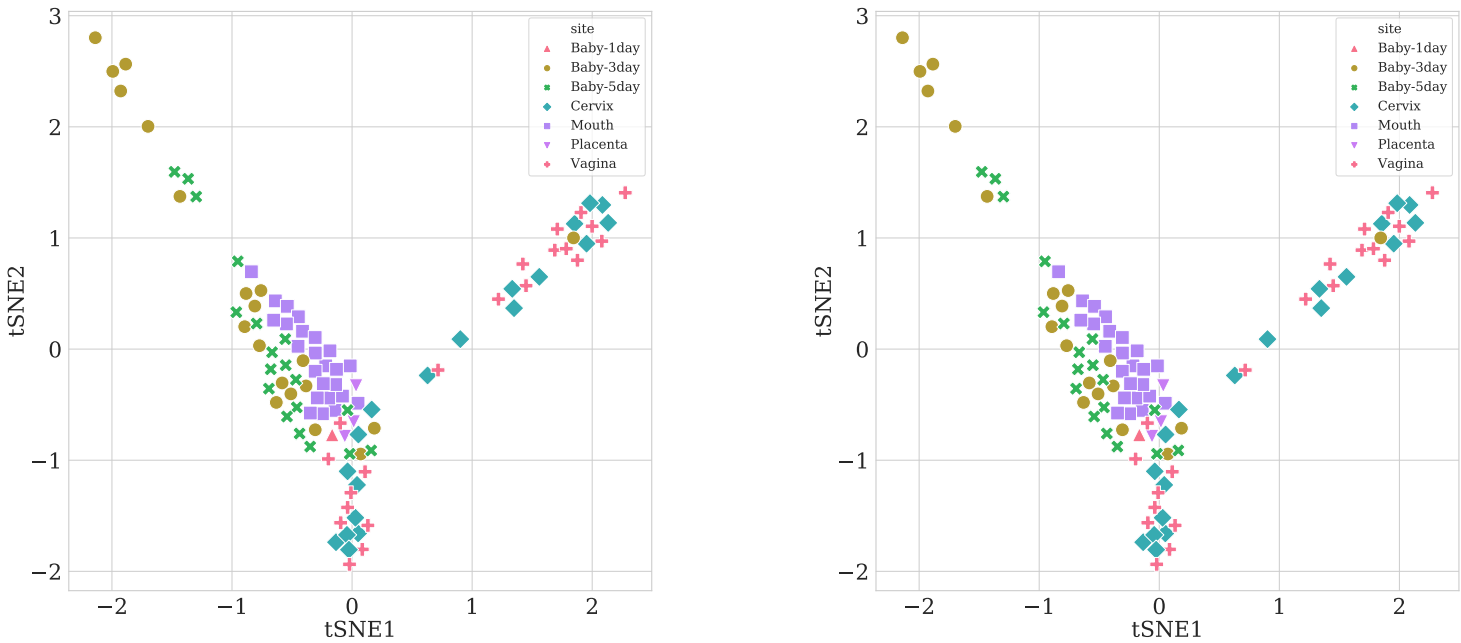


Figure: Workflow of t-SNE for Site/Premature Information

t-SNE with Site Information I

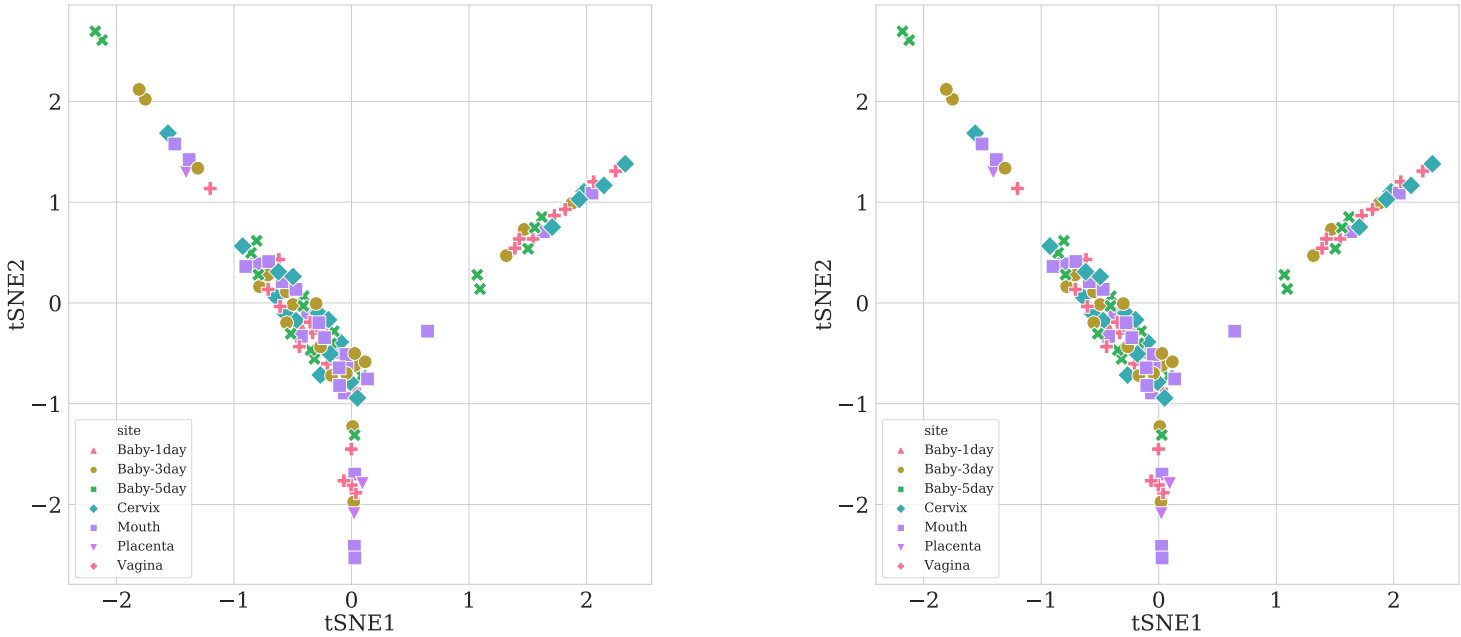


(a) DADA2 + GG

(b) DADA2 + SILVA

Figure: t-SNE with Site by DADA2

t-SNE with Site Information II

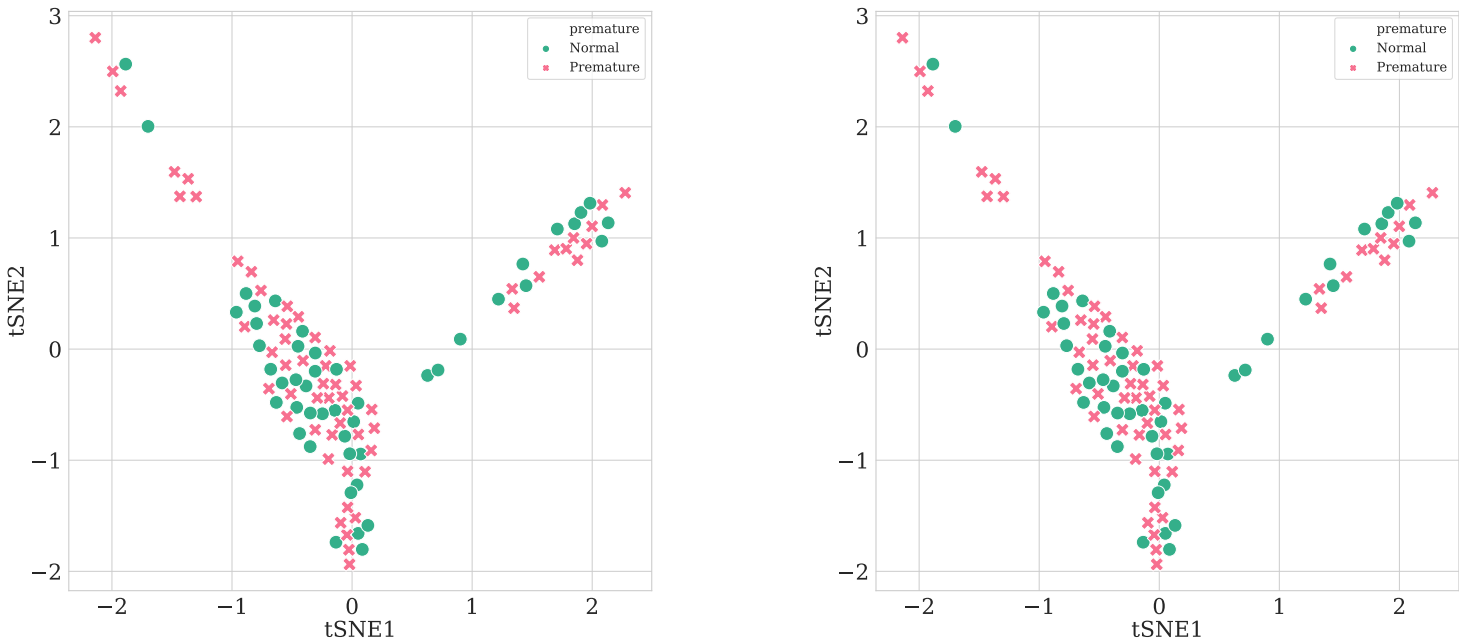


(c) Deblur + GG

(d) Deblur + SILVA

Figure: t-SNE with Site by Deblur

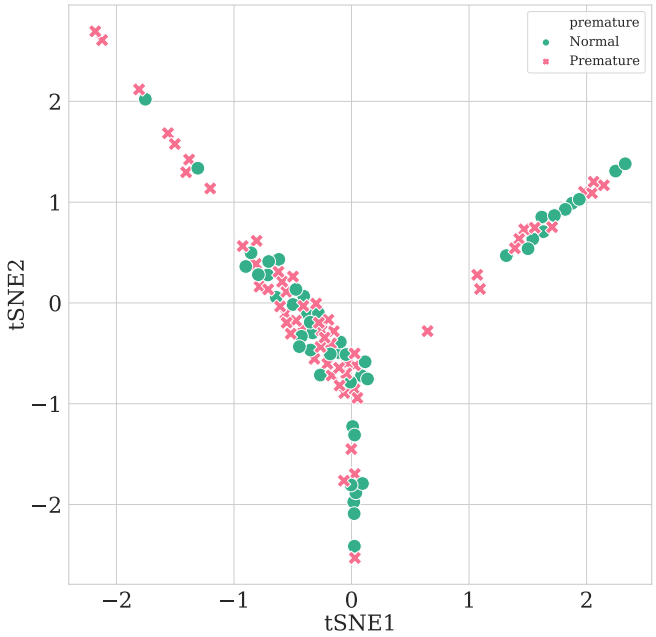
t-SNE with Premature Information I



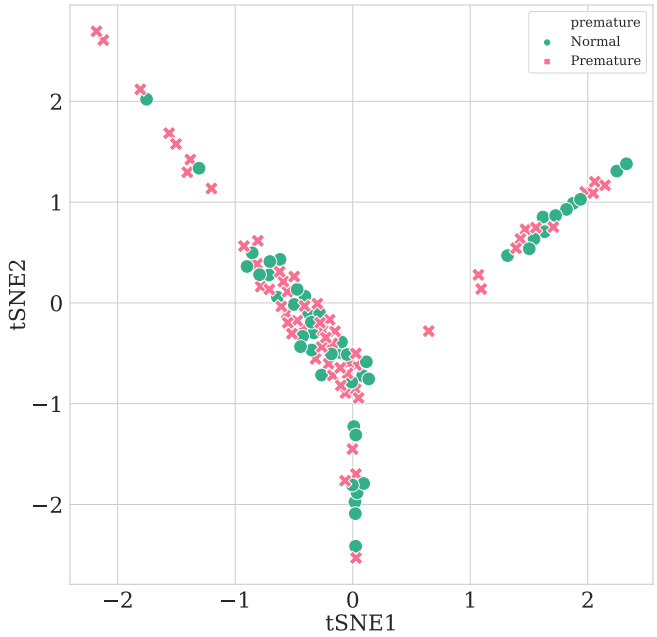
(a) DADA2 + GG (b) DADA2 + SILVA

Figure: t-SNE with Premature by DADA2

t-SNE with Premature Information II



(c) Deblur + GG



(d) Deblur + SILVA

Figure: t-SNE with Premature by Deblur

Results

Bacterial Abundance Test with ANCOM

ANCOM?

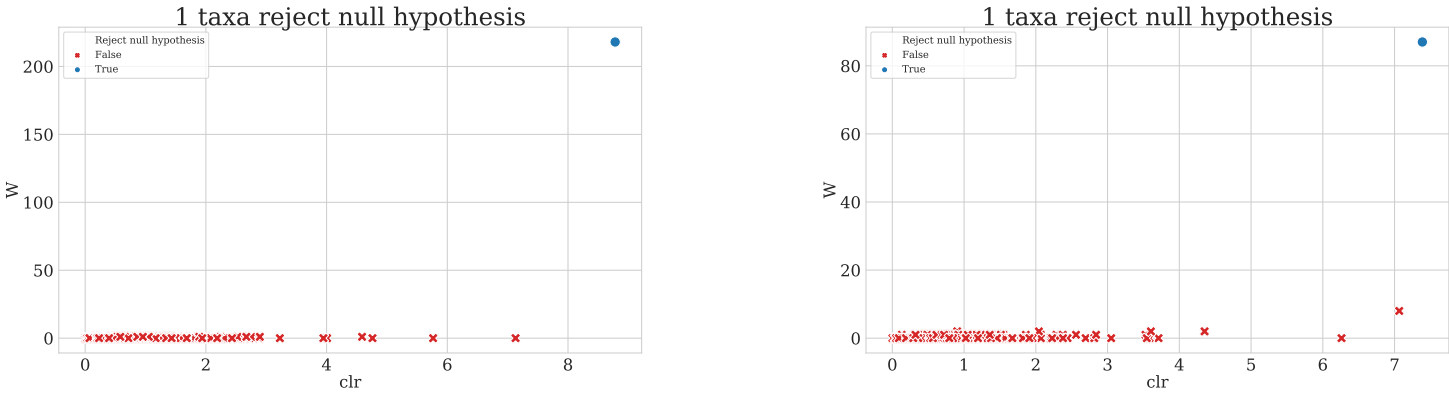
- Analysis of composition of microbiomes
- ANCOM can be used for analyzing the composition of microbiomes in multiple populations (Mandal, Van Treuren, White, Eggesbø, et al., 2015)
- Differential abundance testing

1. clr: Centered log($Ratio$)
2. W: a count of the number of sub-hypothesis which have passed for given species

ANCOM with ...

- Site where get the microbiome
- Premature – Before 37 weeks and After 37 weeks
- Detailed Premature – Before 34 weeks, After 37 weeks, and the other
- C-section
- PROM – Premature rupture of membranes
- Using Steroid?
- Using anti-biotic?

ANCOM with Detailed Premature

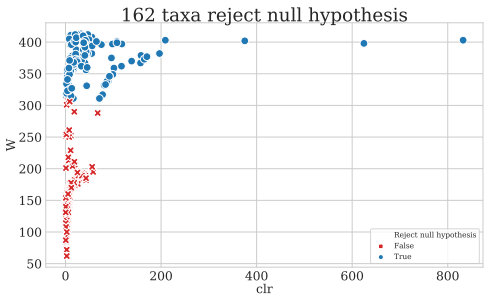


(a) DADA2 + GG (b) Deblur + GG

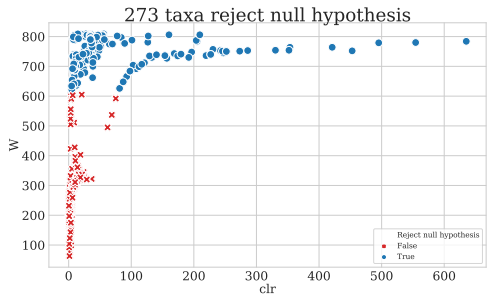
Figure: ANCOM results with Detailed Premature

- *Ureaplasma* genus
- *Aerococcus* genus

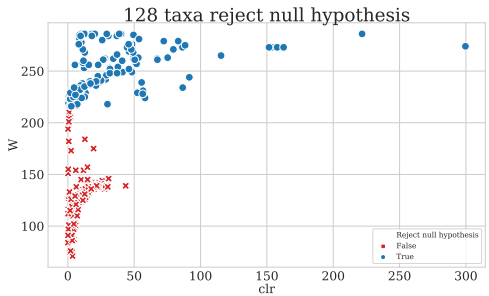
ANCOM with Site



(a) DADA2 + GG



(b) DADA2 + Silva



(c) Deblur + GG

Figure: ANCOM results with Site

ANCOM with PROM

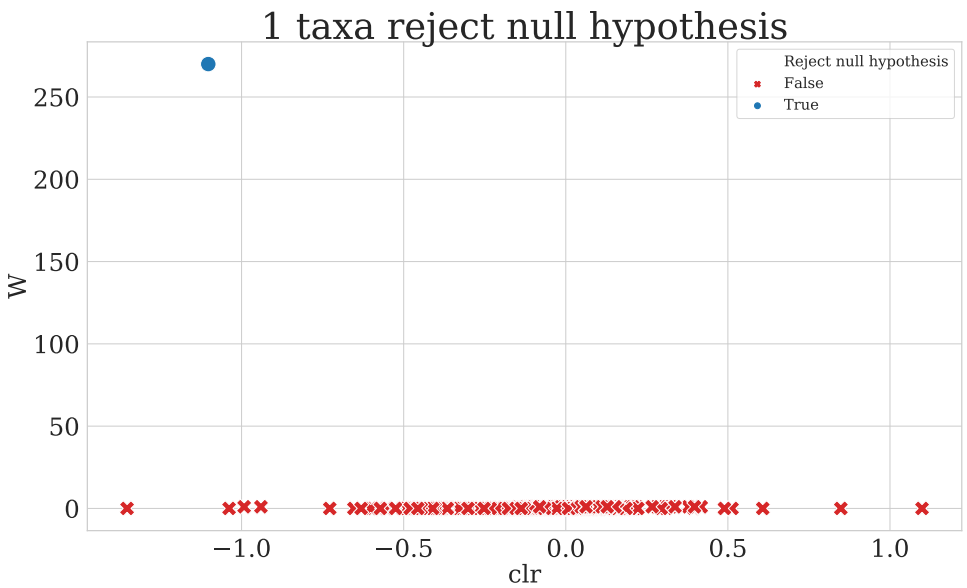
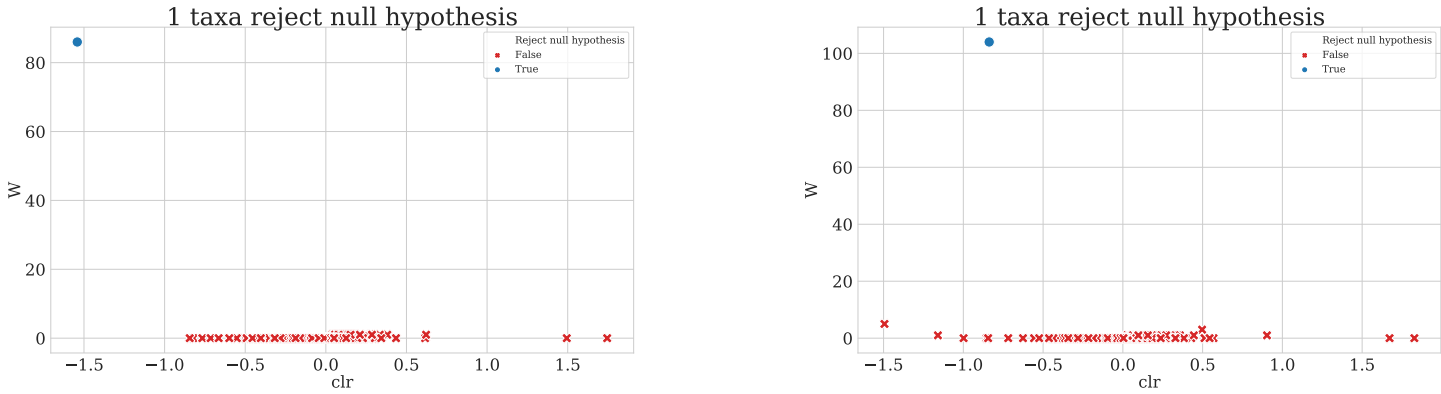


Figure: ANCOM result with PROM (DADA2 + GG)

- *Campylobacteraceae* genus *Campylobacter* species

ANCOM with Using Steroid



(a) DADA2 + GG (b) Deblur + GG

Figure: ANCOM results with Using Steroid

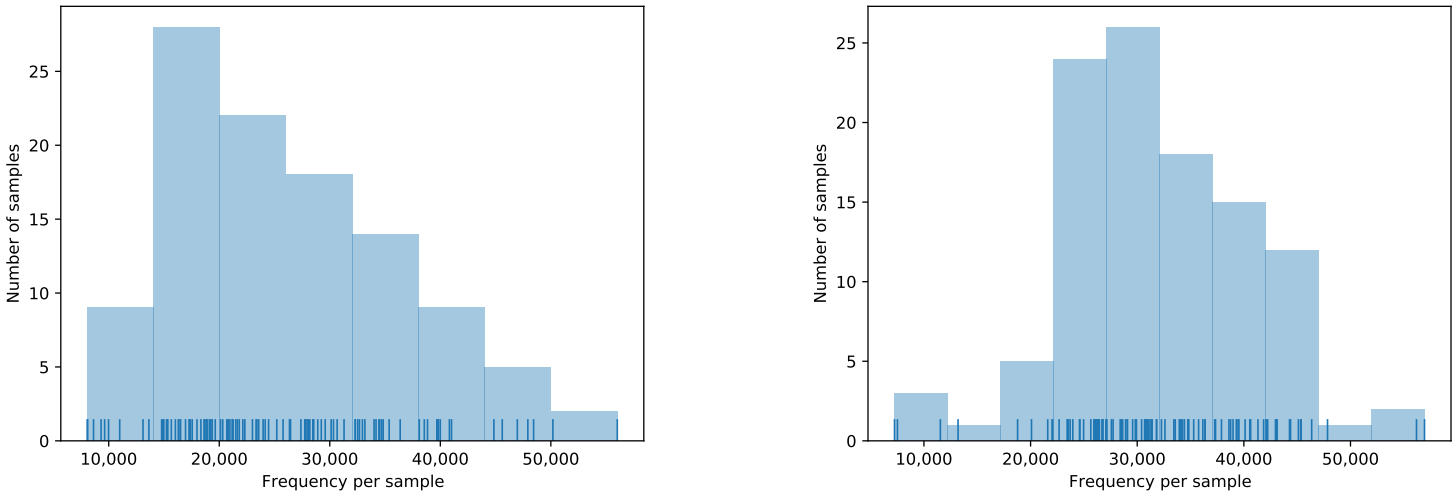
- *Ureaplasma* genus
- *Aerococcus* genus

Results

Rarefaction

Rarefaction?

Rarefaction from First Data



(a) DADA2

(b) Deblur

Figure: Rarefaction from the First Data

- DADA2: 8062
- Deblur: 7239

Results

Alpha-Diversity

Alpha-Diversity

Results

Beta-Diversity

Beta-Diversity

Results

Classification

Workflow for Classification

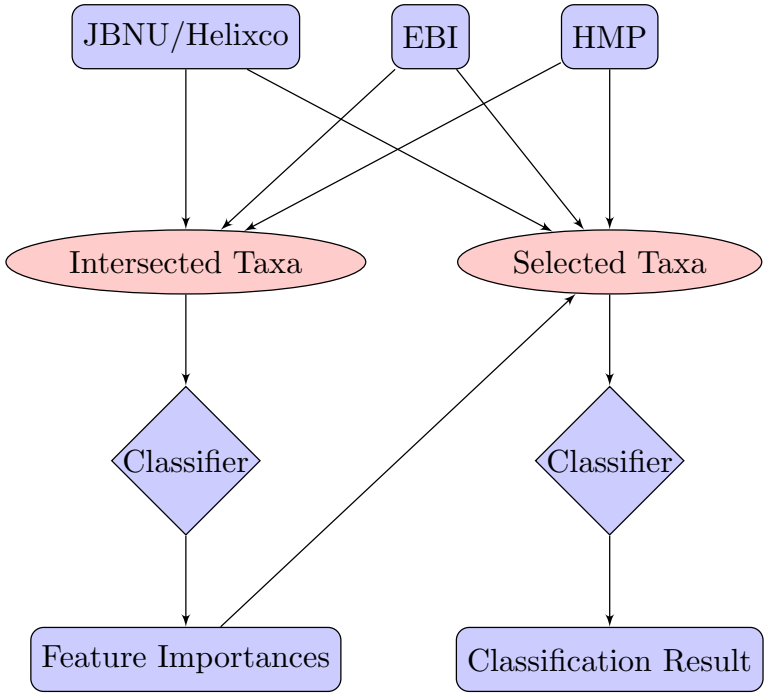


Figure: Workflow with Classification

Random Forest Classifier I

Input Data was treated with **Deblur** and **SILVA**.

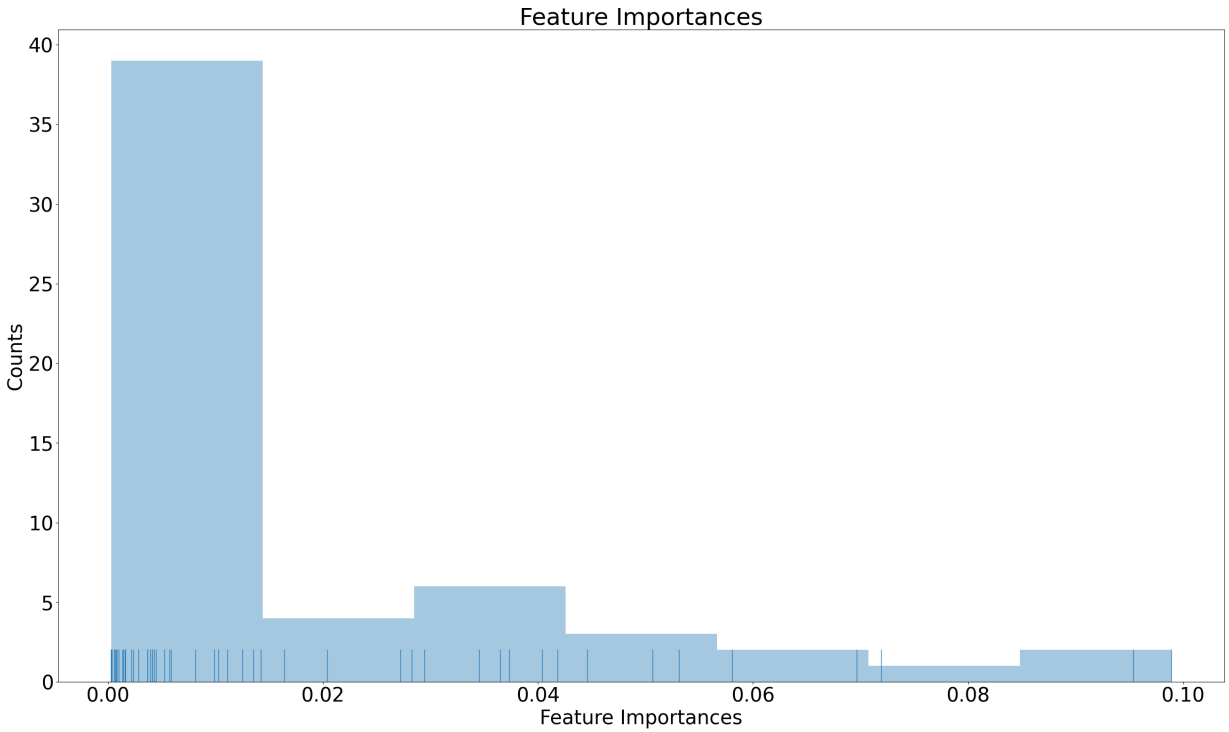


Figure: Feature Importance derived by Random Forest Classifier

Random Forest Classifier II

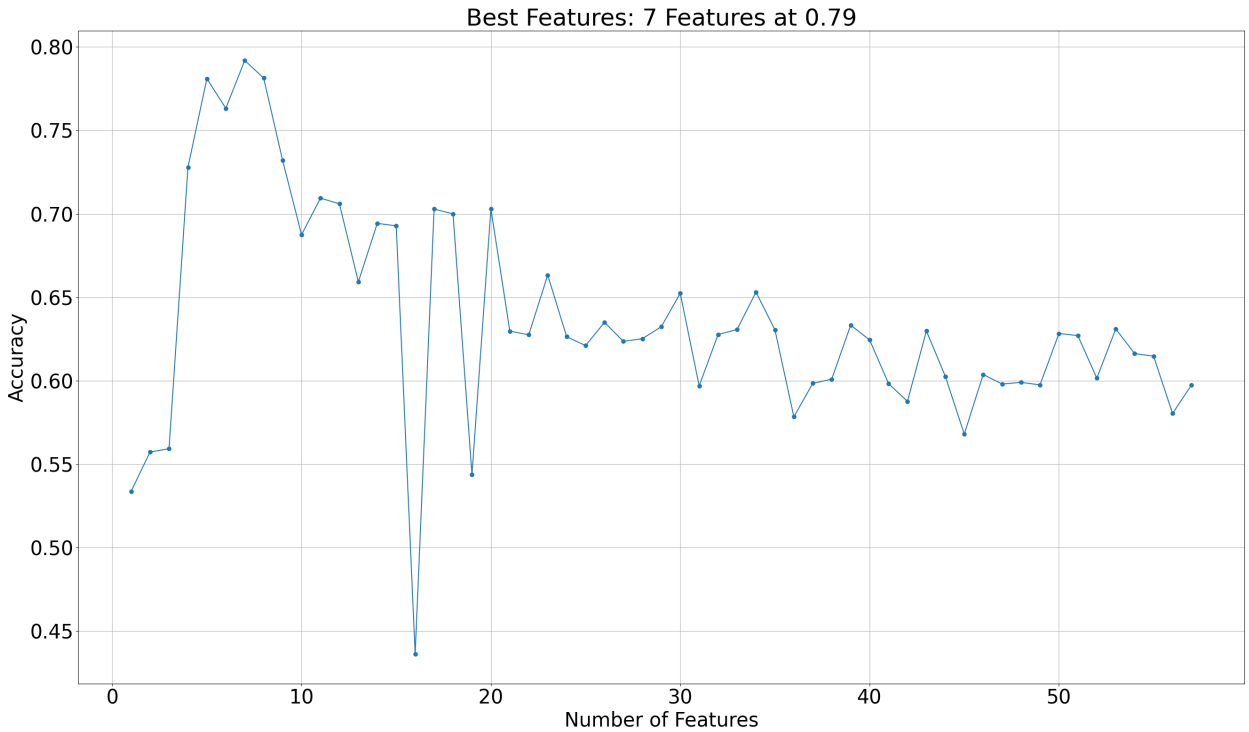


Figure: Number of Features vs. Accuracy

Random Forest Classifier III

1. *Bacteria Firmicutes Bacilli Lactobacillales Lactobacillaceae Lactobacillus Lactobacillus iners*
2. *Bacteria Fusobacteriota Fusobacteriia Fusobacteriales Leptotrichiaceae Leptotrichia*
3. *Bacteria Actinobacteriota Actinobacteria*
4. *Bacteria Firmicutes Bacilli Lactobacillales Lactobacillaceae Lactobacillus*
5. *Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Peptostreptococcaceae Romboutsia*
6. *Bacteria Firmicutes Bacilli Mycoplasmatales Mycoplasmataceae Ureaplasma*
7. *Bacteria Actinobacteriota Actinobacteria Corynebacteriales Corynebacteriaceae Corynebacterium Corynebacterium matruchotii*

Random Forest Classifier IV

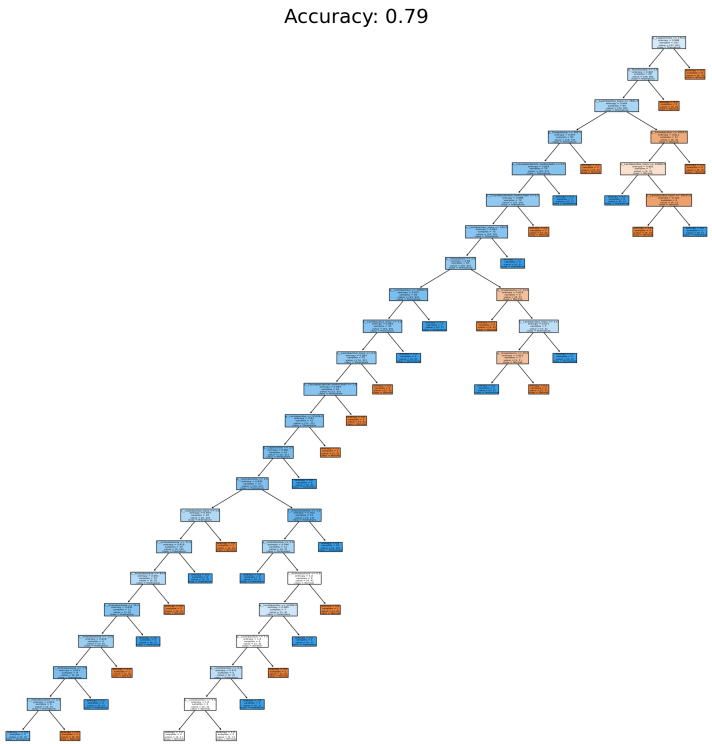
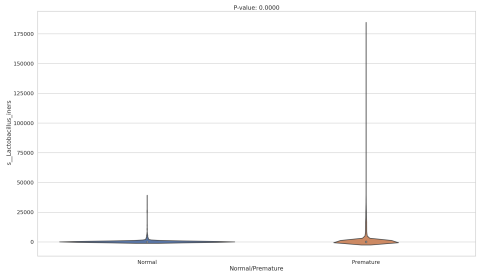
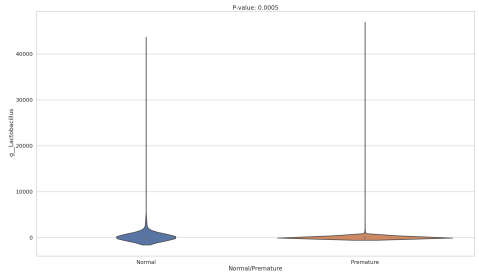


Figure: Random Forest Classifier

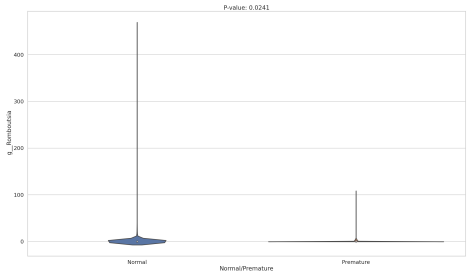
Random Forest Classifier V



(a) *Lactobacillus iners*



(b) *Lactobacillus*



(c) *Romboutsia*

Figure: Violin Plot of Taxonomy

- a *Bacteria Firmicutes Bacilli Lactobacillales Lactobacillaceae Lactobacillus Lactobacillus iners*
- b *Bacteria Firmicutes Bacilli Lactobacillales Lactobacillaceae Lactobacillus*
- c *Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Peptostreptococcaceae Romboutsia*

Lactobacillus (Lb.)

- Vaginal *Lb.* may be clinically useful tools at PTB under 33 weeks. (Usui et al., 2002)
- Presence of *Lb.* sp (odds ratio 0.2) was negatively associated. (MARTIUS et al., 1988)
- *Lb. crispatus/gasseri* could decrease the risk of PTB. (Stafford et al., 2017)
- *Lb.* were associated with decreased risk of PTB. (Tabatabaei et al., 2019)

Romboutsia

References I

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., ... others (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, *2*(2).

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature Biotechnology*, *37*(8), 852-857. Retrieved from `https://doi.org/10.1038/s41587-019-0209-9` doi: 10.1038/s41587-019-0209-9

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods*, *13*(7), 581–583.

References II

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, *72*(7), 5069–5072.

Dominguez-Bello, M. G., De Jesus-Laboy, K. M., Shen, N., Cox, L. M., Amir, A., Gonzalez, A., ... others (2016). Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nature medicine*, *22*(3), 250.

Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., ... others (2019). The vaginal microbiome and preterm birth. *Nature medicine*, *25*(6), 1012–1021.

References III

Gihawi, A., Rallapalli, G., Hurst, R., Cooper, C. S., Leggett, R. M., & Brewer, D. S. (2019). Sepath: benchmarking the search for pathogens in human tissue whole genome sequence data leads to template pipelines. *Genome biology*, *20*(1), 1–15.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., ... Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *science*, *312*(5778), 1355–1359.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(3), 90–95.

Janda, J. M., & Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, *45*(9), 2761–2764.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

References IV

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, *26*(1), 27663.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, *26*(1), 27663. doi: 10.3402/mehd.v26.27663

MARTIUS, J., KROHN, M. A., HILLIER, S. L., STAMM, W. E., HOLMES, K. K., & ESCHENBACH, D. A. (1988). Relationships of vaginal lactobacillus species, cervical chlamydia trachomatis, and bacterial vaginosis to preterm birth. *Obstetrics & Gynecology*, *71*(1), 89–95.

References V

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., ... Caporaso, J. G. (2012). The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. *GigaScience*, *1*(1), 7. doi: 10.1186/2047-217X-1-7

McKinney, W., et al. (2011). pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*(9).

Mignard, S., & Flandrois, J.-P. (2006). 16s rrna sequencing in routine bacterial identification: a 30-month experiment. *Journal of microbiological methods*, *67*(3), 574–581.

Olsen, G. J., & Woese, C. R. (1993). Ribosomal rna: a key to phylogeny. *The FASEB journal*, *7*(1), 113–123.

References VI

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids research*, *35*(21), 7188–7196.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2012). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, *41*(D1), D590–D596.

Stafford, G. P., Parker, J. L., Amabebe, E., Kistler, J., Reynolds, S., Stern, V., ... Anumba, D. O. (2017). Spontaneous preterm birth is associated with differential expression of vaginal metabolites by lactobacilli-dominated microflora. *Frontiers in physiology*, *8*, 615.

References VII

Tabatabaei, N., Eren, A., Barreiro, L., Yotova, V., Dumaine, A., Allard, C., & Fraser, W. (2019). Vaginal microbiome in early pregnancy and subsequent risk of spontaneous preterm birth: a case–control study. *BJOG: An International Journal of Obstetrics & Gynaecology*, *126*(3), 349–358.

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., ... others (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, *92*(2), 698–715.

Tucker, J., & McGuire, W. (2004). Epidemiology of preterm birth. *Bmj*, *329*(7467), 675–678.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*(7164), 804–810.

References VIII

Usui, R., Ohkuchi, A., Matsubara, S., Izumi, A., Watanabe, T., Suzuki, M., & Minakami, H. (2002). Vaginal lactobacilli and preterm birth. *Journal of perinatal medicine*, *30*(6), 458–466.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, *17*(3), 261–272.

Waskom, M., Botvinnik, O., Ostblom, J., Gelbart, M., Lukauskas, S., Hobson, P., ... Brian (2020, April). *mwaskom/seaborn: v0.10.1 (april 2020).* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.3767070` doi: 10.5281/zenodo.3767070