Microbiome Premature

Jaewoong Lee

Ulsan National Institute of Science and Technology

jwlee230@unist.ac.kr

2020-09-09

Overview

Microbiome

Ribosomal RNA

Premature

16s rRNA Sequencing

Qiime 2

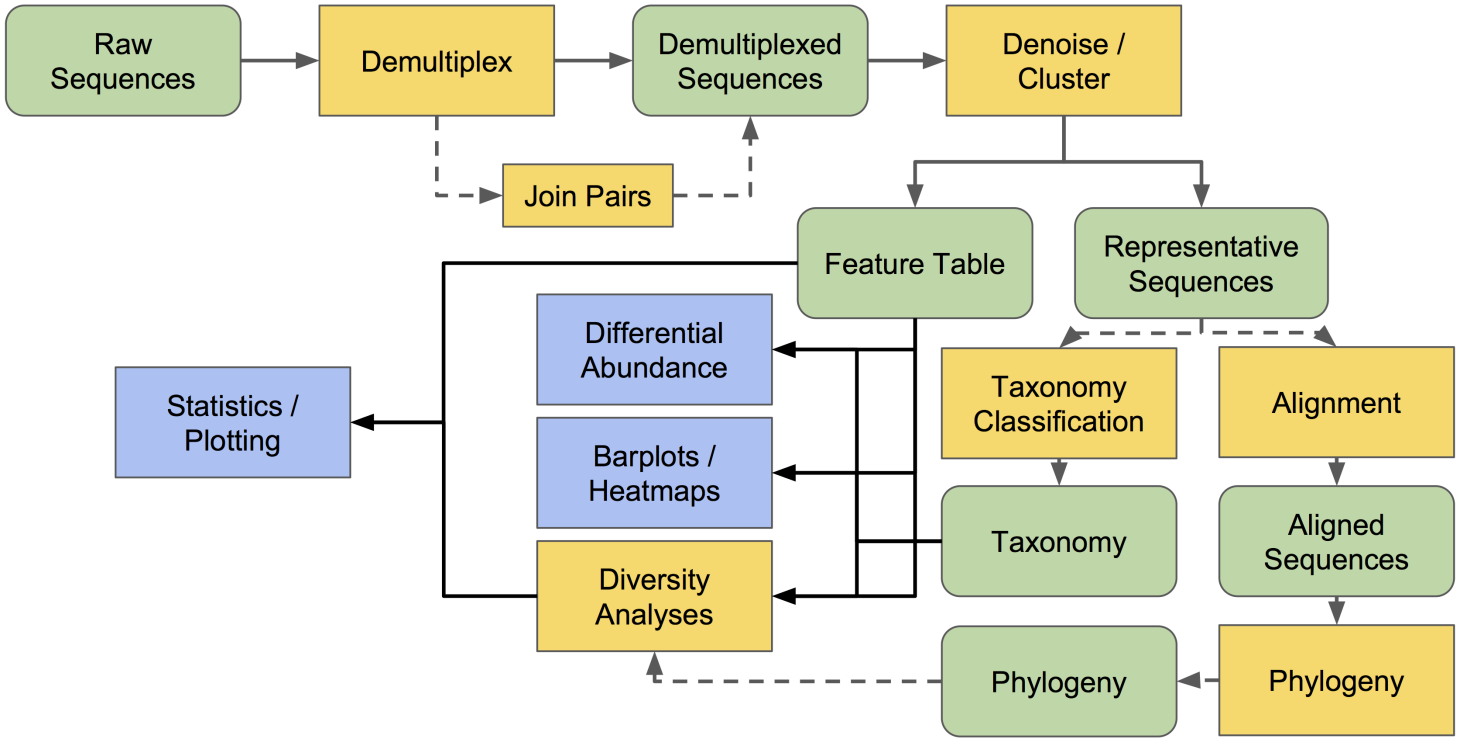


Figure: QIIME 2 workflow (Bolyen et al., 2019; Mandal et al., 2015; McDonald et al., 2012)

Denoising Techniques

- DADA2 (Callahan et al., 2016)
- Deblur (Amir et al., 2017)

Taxonomy Classification

- Greengenes (GG): Kingdom - Species (DeSantis et al., 2006)
- SILVA: Domain - Genus (Pruesse et al., 2007; Quast et al., 2012)

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

Mothur

Figure: Mothur

Note: Still in progress

t-distributed Stochastic Neighbor Embedding (t-SNE)

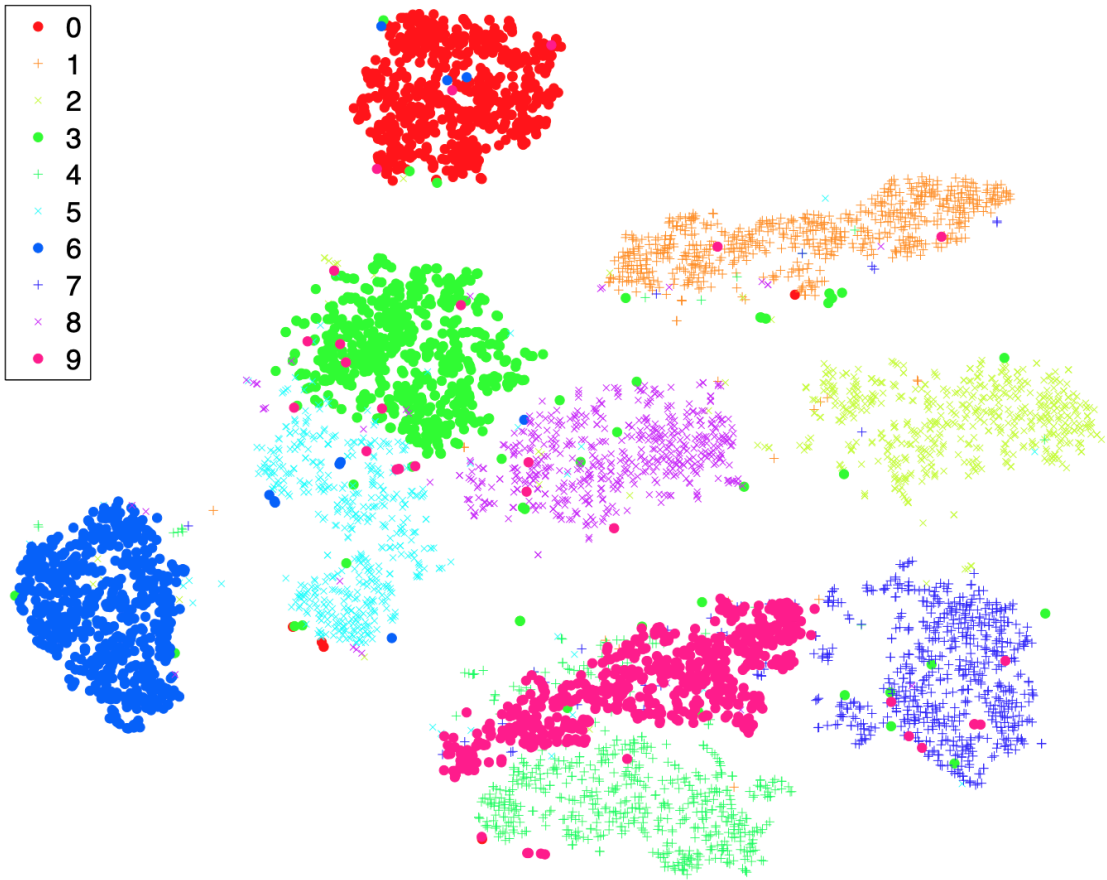
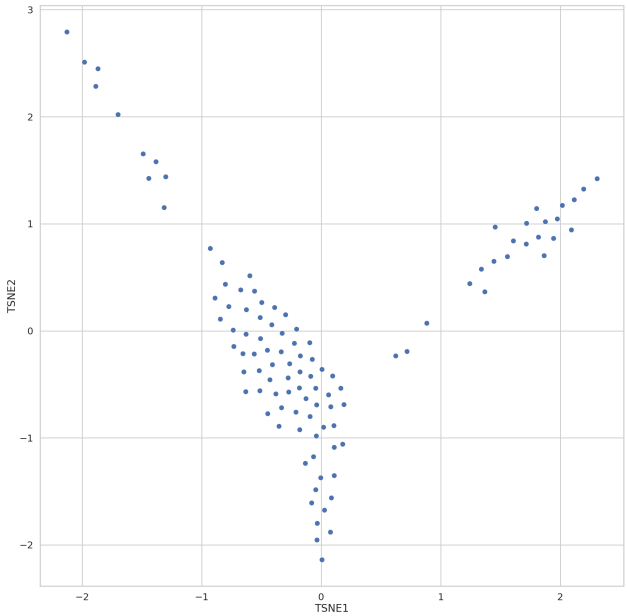


Figure: t-SNE with handwritten data (Maaten & Hinton, 2008)

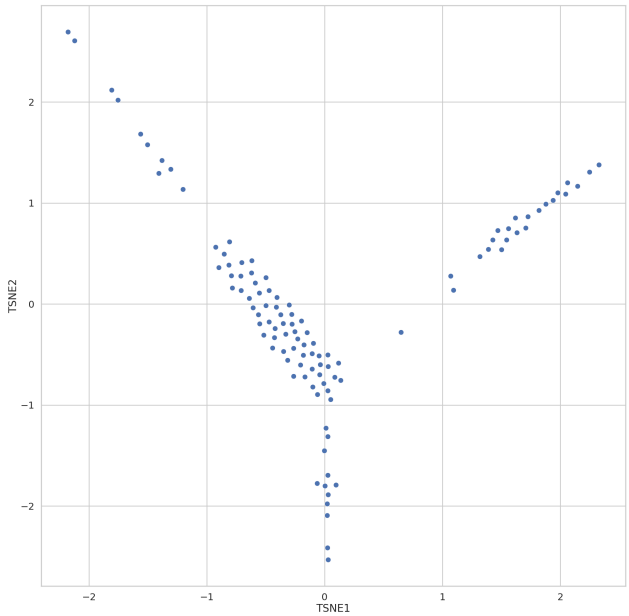
Python Packages

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

t-SNE for Brief Information



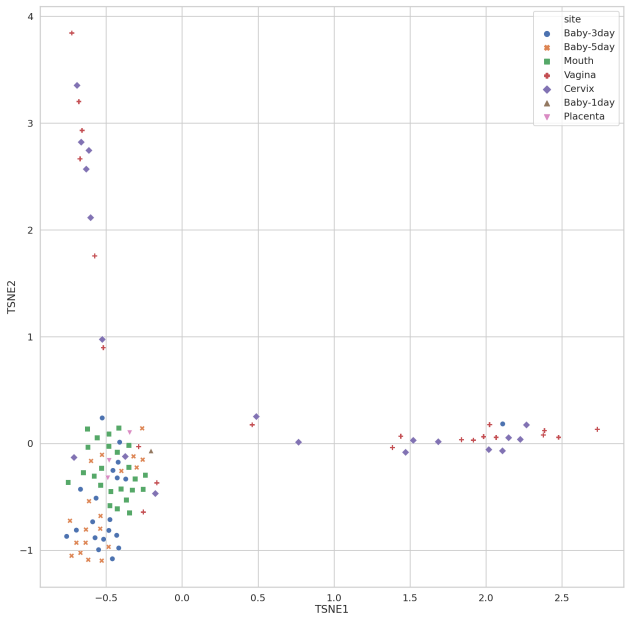
(a) DADA2



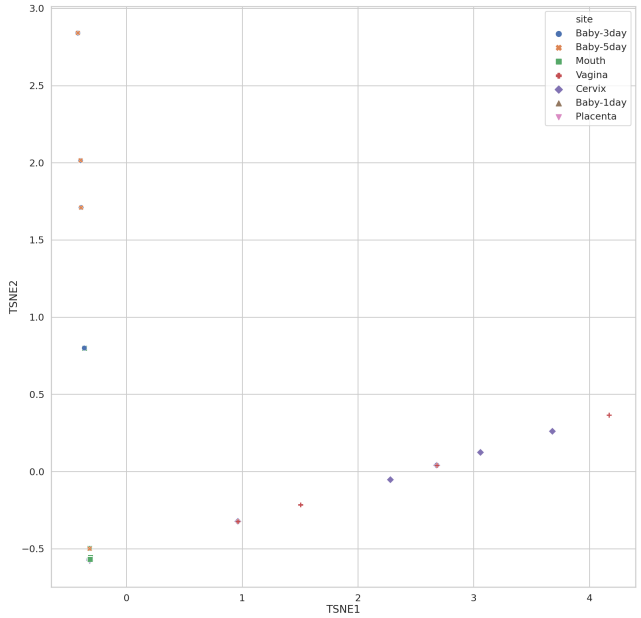
(b) Deblur

Figure: t-SNE for Brief Information

t-SNE with Site I

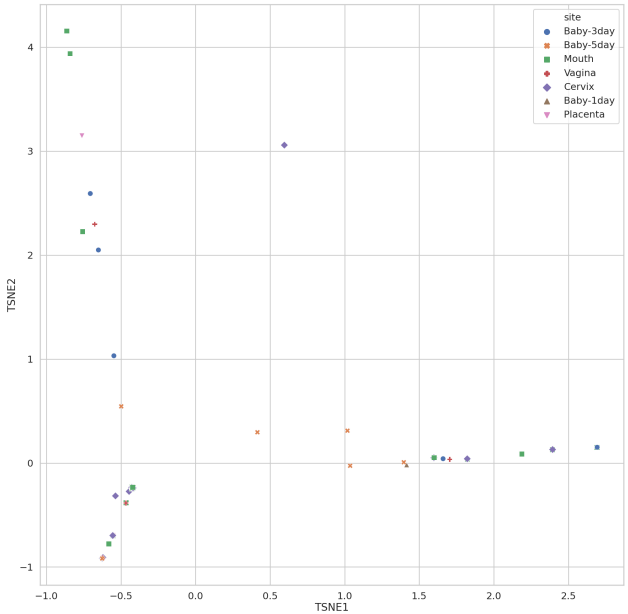


(a) DADA2 + GG

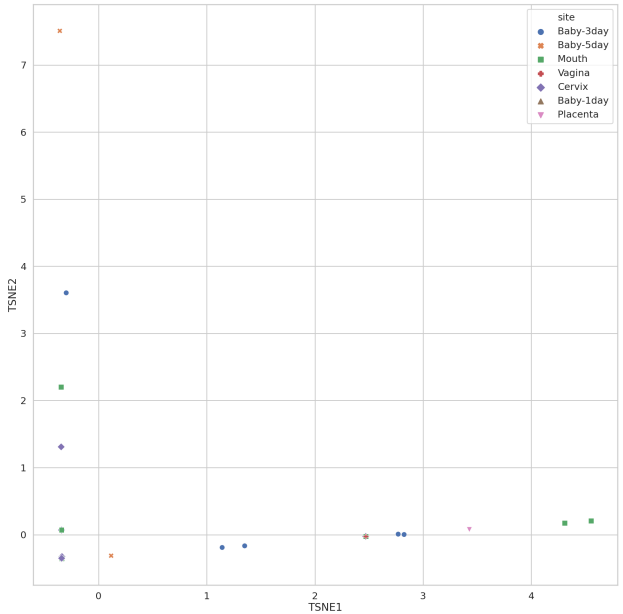


(b) DADA2 + SILVA

t-SNE with Site II



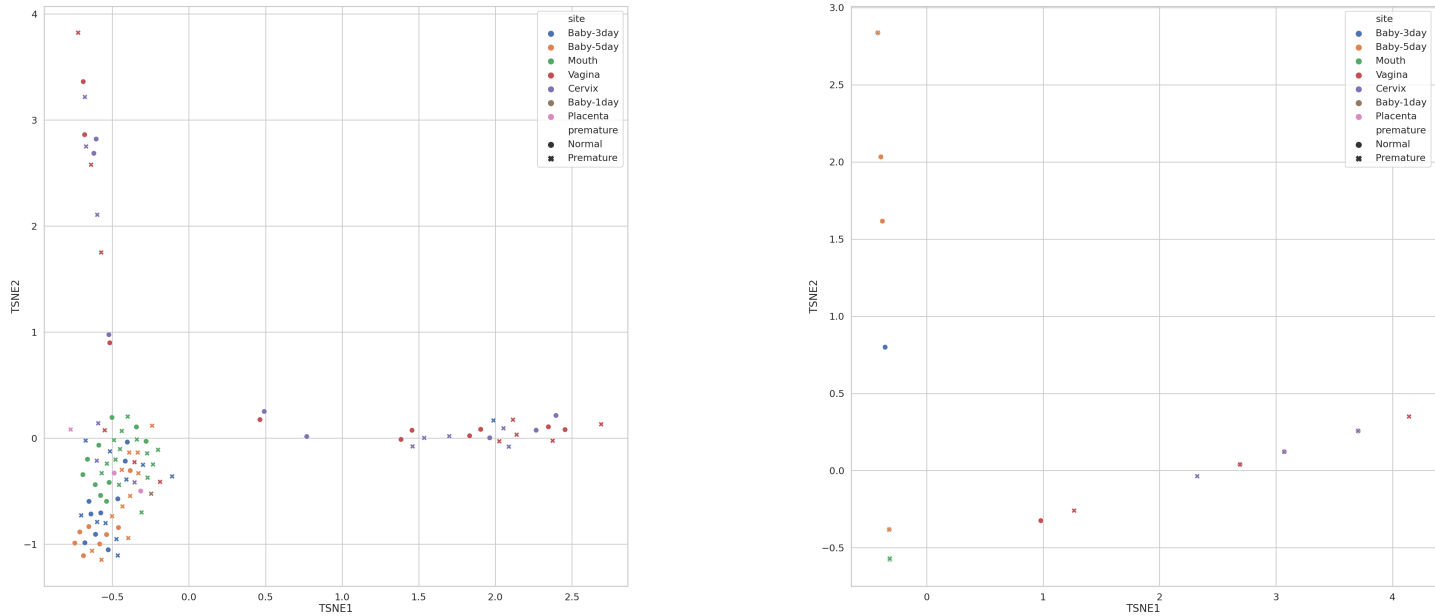
(c) Deblur + GG



(d) Deblur + SILVA

Figure: t-SNE with Site

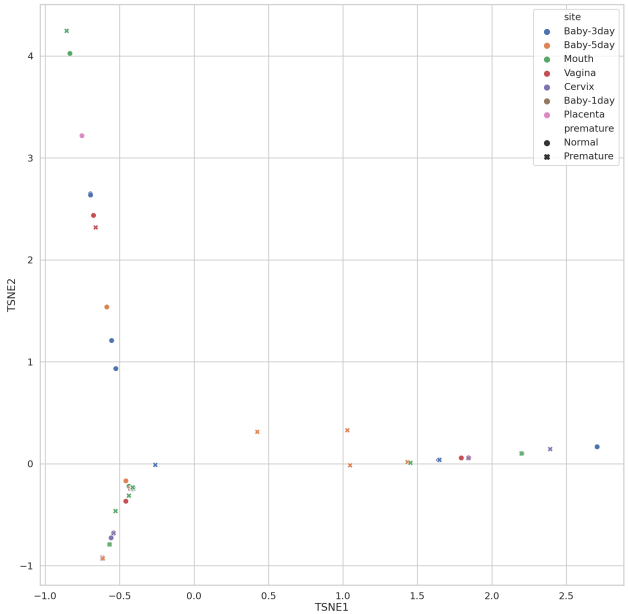
t-SNE with Site + Premature I



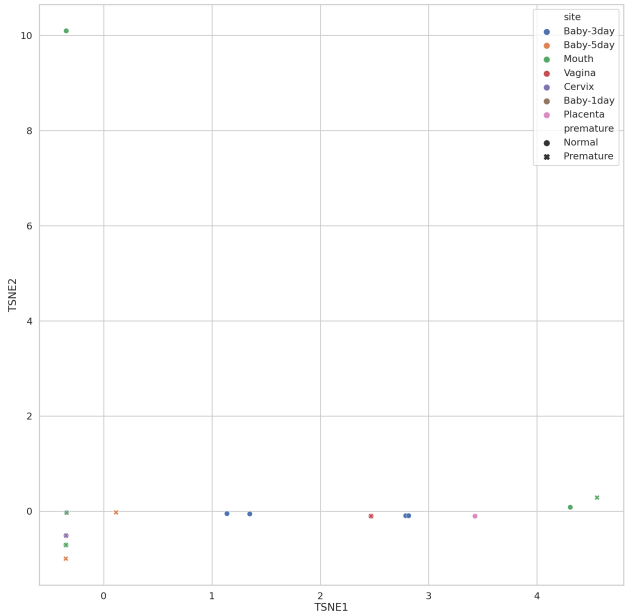
(a) DADA2 + GG (b) DADA2 + SILVA

Figure: t-SNE with Site

t-SNE with Site + Premature II



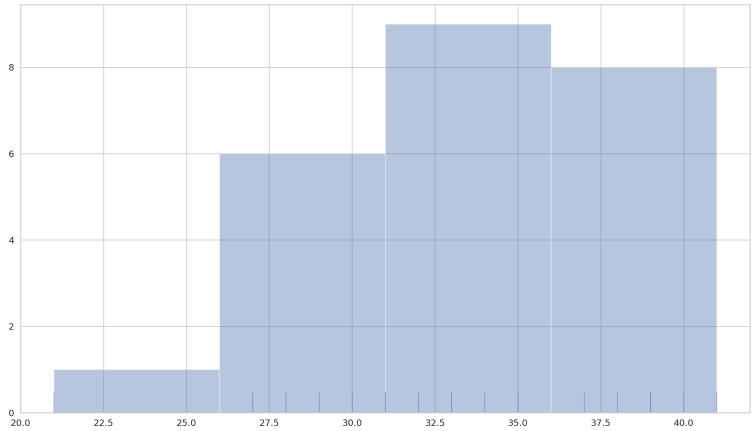
(c) Deblur + GG



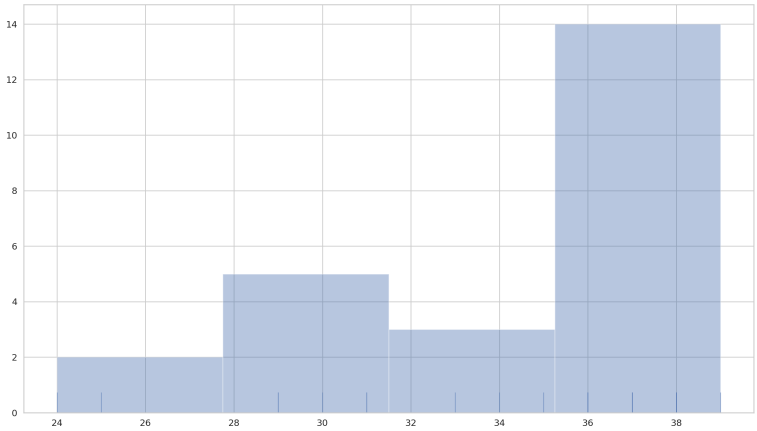
(d) Deblur + SILVA

Figure: t-SNE with Site

Histogram with Clinical Information



(a) Age



(b) Weeks

Figure: Histogram with Clinical Information

Yields I

- t-SNE plots
  - in Brief
  - by Site
  - by Site + Premature
- Histogram
  - by Age
  - by Weeks
- Have tried (but in vain)
  - ANCOM with premature/normal
  - Classification with raw TSV

Requirements I

- More data
- Mothur pipeline
- Classification

Expectations I

- Classifier result (Statistical values)

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.

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.

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

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

References III

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

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).

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.

References IV

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.

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