

Microbiome Premature

Jaewoong Lee

Ulsan National Institute of Science and Technology

jwlee230@unist.ac.kr

2020-09-22

Overview

1 Introduction

2 Materials

3 Methods

4 Results

5 Proceedings

References

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)

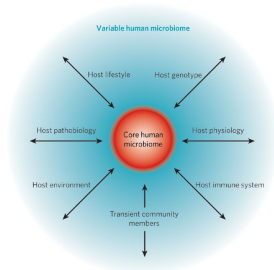


Figure: Concept of a core human microbiome (Turnbaugh et al., 2007)

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

Premature (Preterm Birth)

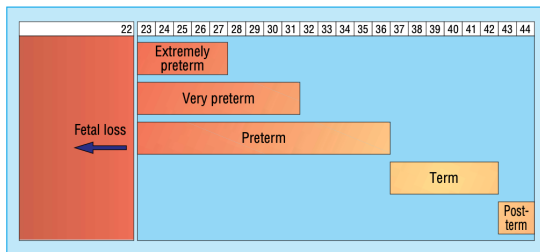


Figure: Definitions of Premature (Tucker & McGuire, 2004)

∴ Hence, in this study,

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

Materials

16s rRNA Sequencing

Train Data vs. Validate Data

Train/Test data

- Helixco: Data collected by Helixco

Validate data

- EBI (European Bioinformatics Institute): Data collected by Dominguez-Bello et al., 2016
- HMP (Human Microbiome Project): Data collected by Fettweis et al., 2019

Methods

Qiime 2

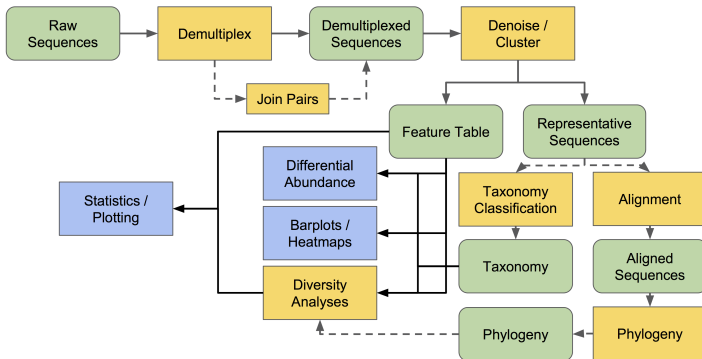


Figure: QIIME 2 workflow (Bolyen et al., 2019; Mandal et al., 2015; McDonald et al., 2012)

Filtering with Quality Score I

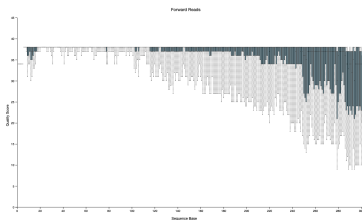
Drawback between:

- Longer sequence read
- Higher quality value

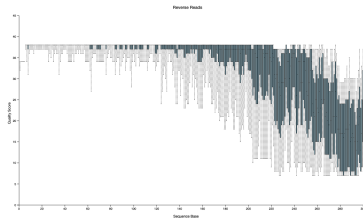
∴ I select the length n where:

$$\begin{aligned} \forall n_i \in \{n_k | \text{MedianQualityScore} \geq 30\} \\ \exists ! n \in \{n_i\} : n \geq n_i \end{aligned} \quad (1)$$

Filtering with Quality Score II



(a) Forward



(b) Reverse

Figure: Sequence Quality Plot from Helixco Data

Filtering with Quality Score III

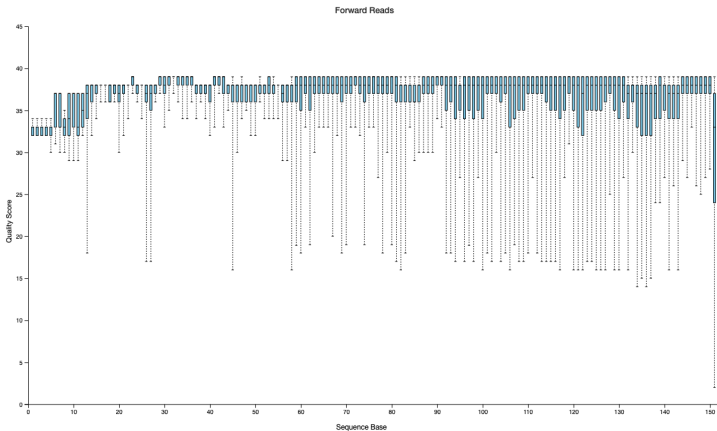
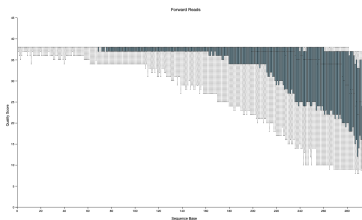
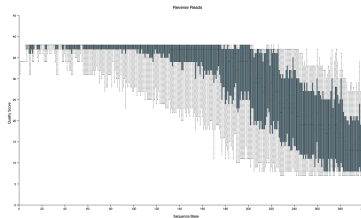


Figure: Sequence Quality Plot from EBI

Filtering with Quality Score IV



(a) Forward



(b) Reverse

Figure: Sequence Quality Plot from HMP Data

Denoising Techniques

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

Taxonomy Classification

- Greengenes (GG): Kingdom \leftrightarrow Species (DeSantis et al., 2006)
- SILVA: Domain \leftrightarrow 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)

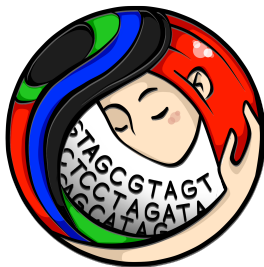


Figure: Mothur

Note: Still in progress

t-distributed Stochastic Neighbor Embedding (t-SNE)

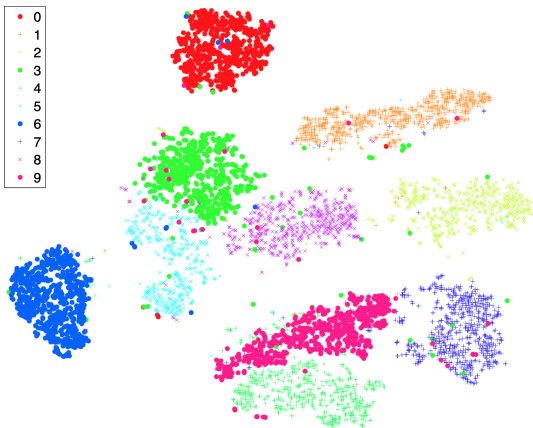
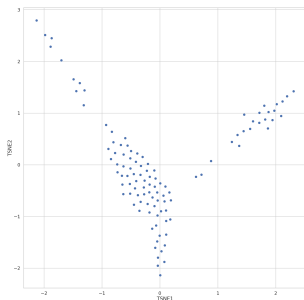


Figure: t-SNE with handwritten data (Maaten & Hinton, 2008)

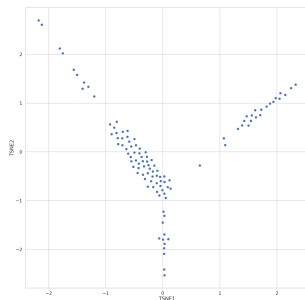
- Pandas (McKinney et al., 2011)
- Scikit-Learn (Pedregosa et al., 2011)
- Matplotlib (Hunter, 2007)
- Seaborn (Waskom et al., 2020)

Results

t-SNE for Brief Information



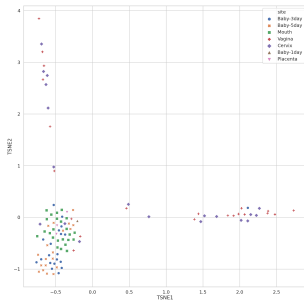
(a) DADA2



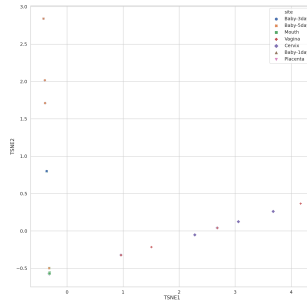
(b) Deblur

Figure: t-SNE for Brief Information

t-SNE with Site I

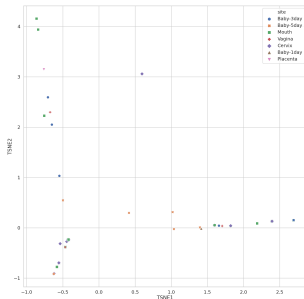


(a) DADA2 + GG

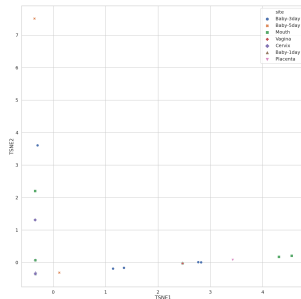


(b) DADA2 + SILVA

t-SNE with Site II



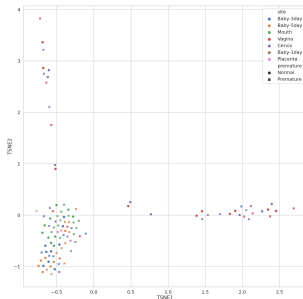
(c) Deblur + GG



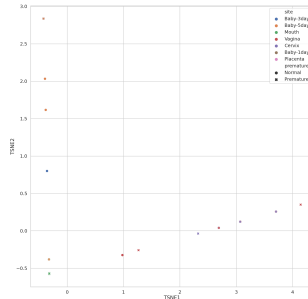
(d) Deblur + SILVA

Figure: t-SNE with Site

t-SNE with Site + Premature



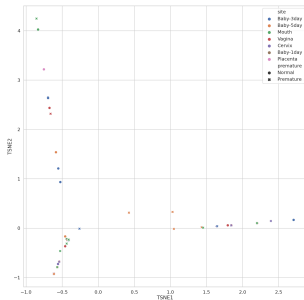
(a) DADA2 + GG



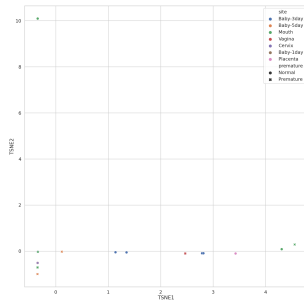
(b) DADA2 + SILVA

Figure: t-SNE with Site

t-SNE with Site + Premature II



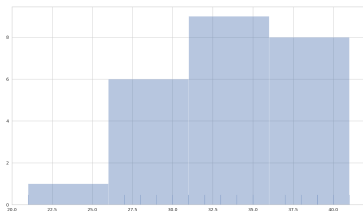
(c) Deblur + GG



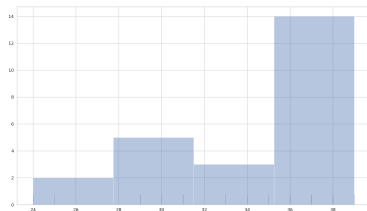
(d) Deblur + SILVA

Figure: t-SNE with Site

Histogram with Clinical Information



(a) Age



(b) Weeks

Figure: Histogram with Clinical Information

Proceedings

- 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

- Classifier result (Statistical values)

- 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.

- 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.
- Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov), 2579–2605.
- 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).
- Olsen, G. J., & Woese, C. R. (1993). Ribosomal rna: a key to phylogeny. *The FASEB journal*, 7(1), 113–123.
- 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.
- 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.

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