Microbiome Premature

Jaewoong Lee

2020-10-13

Contents

List of Tables

List of Figures

1 Introduction

1.1 Microbiome

After the Human Genome Project was finished, the microorganisms which live along humans, as known as the microbiota, are considered overwhelmed human cells (Turnbaugh et al., 2007). Moreover, the microbiome, the collective genome from these microbiota (Gill et al., 2006), serve as the trait of individual have not to evolve on their own (Turnbaugh et al., 2007). Furthermore, human microbiome is effected by host's life style as figure 1.

1.2 Ribosomal RNA

Ribosomal RNA (rRNA) plays the main roles in a cell. This main roles include mRNA selection, tRNA binding, proof-reading, factor binding, and *et cetera* (Noller, 1991). Because of its momentous roles, rRNA could be preserved amongst whole bacteria throughout the evolution.

1.3 Premature

Premature (PTB; stands for Preterm Birth) is the birth of a baby earlier than 37 gestational weeks, as Figure 2 (Tucker & McGuire, 2004). Premature infants have more risk such as hearing problems and sight problems.

2 Materials

2.1 16S rRNA Sequencing

rRNA has been kept among bacteria; thus, 16S rRNA exists in almost bacteria, and its functions has not changed over time. Also, 16S rRNA is large enough for bioinformatics (Janda & Abbott, 2007). Hence, 16S rRNA sequencing is the reference method for bacterial taxonomy classification and identification (Mignard & Flandrois, 2006).

There are three databases which for machine learning: Helixco data, EBI data, and HMP data. Metadata of these databases is as table 1.

2.1.1 Helixco Data

2.1.2 European Bioinformatics Institute Data

EBI data was collected by European Bioinformatics Institute (EBI) (Dominguez-Bello et al., 2016). EBI data aimed to compare Cesarean section birth and vaginal birth; thus, every participants in EBI data is on term, not PTB.

2.1.3 Human Microbiome Project Data

HMP data was collected by Human Microbiome Project (HMP) (Fettweis et al., 2019). HMP data aimed to compare PTB and on-term birth; thus, every participants in HTMP data is PTB.

3 Methods

3.1 Docker

Docker is light-weight Linux containers for consistent development and deployment (Merkel, 2014).

3.2 QIIME 2

QIIME 2 is a next-generation microbiome bioinformatics platform which is extensible, free, open-source, and community developed (Bolyen et al., 2019; Mandal et al., 2015; McDonald et al., 2012).

3.3 Denoising Algorithms

There are two denoising algorithms which are provided by QIIME as figure 4: DADA2 (Callahan et al., 2016) and Deblur (Amir et al., 2017).

3.3.1 DADA2

DADA2 is an open-source software package for modeling and correcting Illumina-sequenced amplicon errors (Callahan et al., 2016).
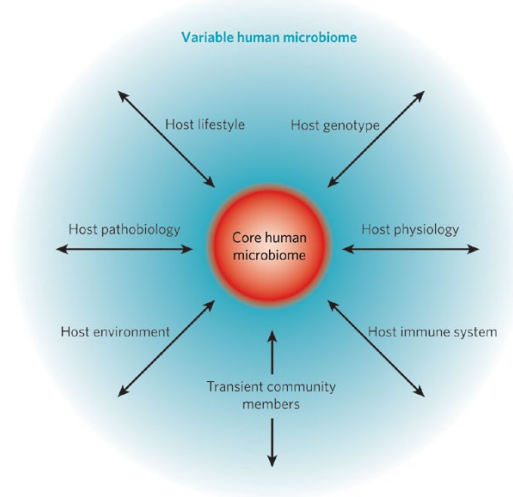
Figure 1: Concept of a core human microbiome. (Turnbaugh et al., 2007)

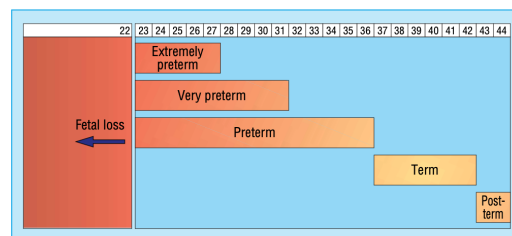
Figure 2: Definition of Premature (Tucker & McGuire, 2004)

Table 1: Metadata of Data

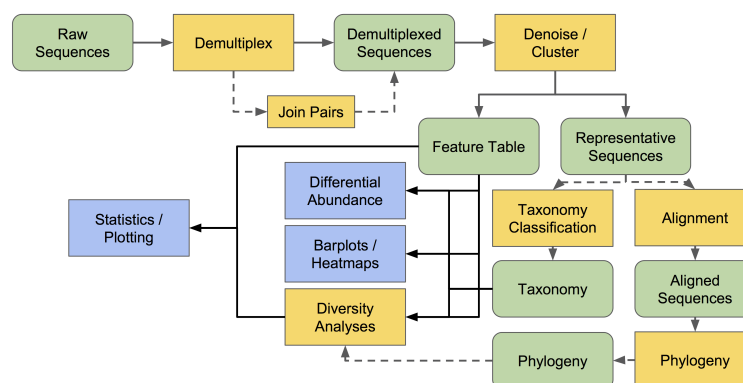| Data	Participants	Samples	Remarks
Helixco	24	107	-
EBI	18	1016	Only Normal
HMP	1572	9205	Only Premature

Figure 3: Workflow of QIIME2

3.3.2 Deblur

Deblur is a software packages which uses error profiles to obtain putative error-free sequences from Illumina MiSeq and HiSeq sequencing platforms (Amir et al., 2017).

3.4 Taxonomy Classification Algorithms

There are two taxonomy classification algorithms which are provided by QIIME as figures 5: Greengenes (DeSantis et al., 2006) ans SILVA (Pruesse et al., 2007; Quast et al., 2012).

3.4.1 Greengenes

Greengenes (GG) is a chimera-checked 16S rRNA gene database (DeSantis et al., 2006).

3.4.2 SILVA

SILVA is a comprehensive web resource for up-to-date, quality-controlled databases of aligned rRNA gene sequences from the Bacteria domains (Pruesse et al., 2007; Quast et al., 2012).

3.5 Merging Taxonomy

After applying denoising algorithms and taxonomy classification algorithms, some reads have different IDs (ASV OTU), but are identified as same taxonomy. In that cases, these reads are merged into same taxonomy as figure 6.

3.6 Mothur

Mothur is an open-source software package for bioinformatics data processing, especially for the analysis of DNA from microbes (Schloss et al., 2009).

3.7 t-distributed Stochastic Neighbor Embedding

T-distributed stochastic neighbor embedding (t-SNE) visualizes high-dimensional data by giving each data-point a location in a two-dimensional map (Maaten & Hinton, 2008).

3.8 Python Packages

3.8.1 Pandas

Pandas is a Python library of rich data structure and tools for working with structured data sets (McKinney et al., 2011).

3.8.2 Scikit-Learn

Scikit-learn is a Python module which integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems (Pedregosa et al., 2011).

3.8.3 Matplotlib

Matplotlib is a two-dimensional graphics package used for Python for image generation (Hunter, 2007).

3.8.4 Seaborn

Seaborn is a Python data visualization library based on Matplotlib (Waskom et al., 2020).

4 Results

4.1 Deciding Trimming Length

Deblur require filtering step; though, DADA2 contains filtering step. For filtering step, trimming length should be decided. In other words, the sequence which is longer than specific length should be removed, because of the quality of sequences. There is no canonically admitted methods for deciding trimming length; thus, trimming length n would be decided as equation 1. Hence, trimming lengths are decided as shown as table 2.
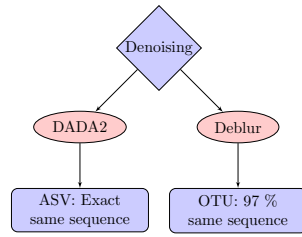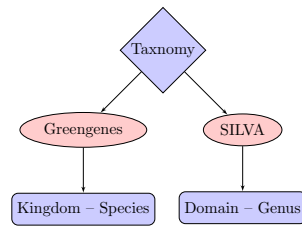
Figure 4: Denoising Algorithms

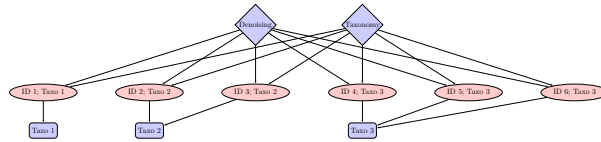Figure 5: Taxonomy Classification Algorithms

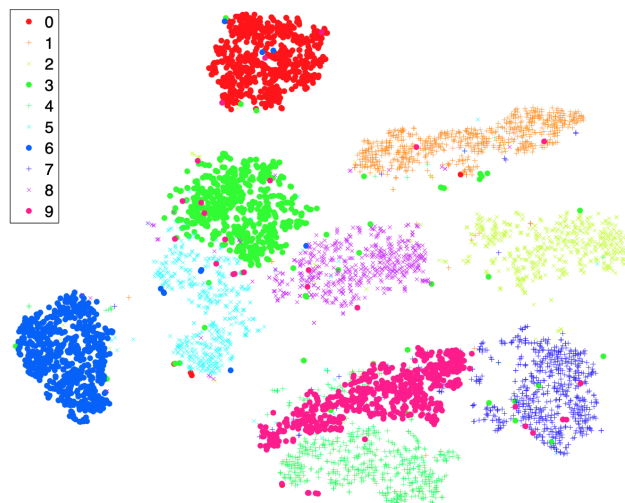Figure 6: Example Diagram for Merging Taxonomy

Figure 7: t-SNE Visualizations of handwritten digits from MNIST data (Maaten & Hinton, 2008)

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

4.2 t-SNE for Comparing Databases

To compare three databases, workflow, which as figure 11, was executed:

1. Select intersected taxonomies that means the taxonomy which can be founded in every database.

2. Draw t-SNE plot in 2-dimension with intersected taxonomies.

Thereupon, intersected taxonomies are as figure 12. There are around 200 - 400 intersected taxonomies. With these intersected taxonomies, t-SNE plot was derived as figure 13.

5 Discussion

5.1 t-SNE for Comparing Databases

6 References

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.

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.

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).

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, *2014*(239), 2.

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.

Noller, H. F. (1991). Ribosomal rna and translation. *Annual review of biochemistry*, *60*(1), 191–227.

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.
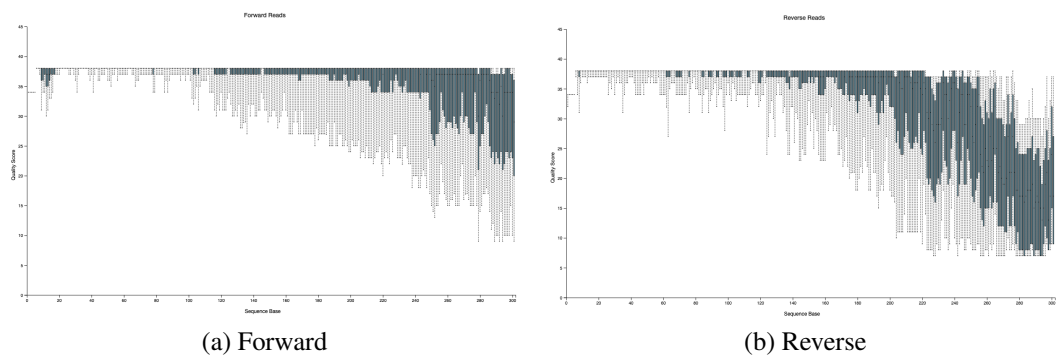
(a) Forward   (b) Reverse

Figure 8: Sequence Quality Plot from Helixco Data
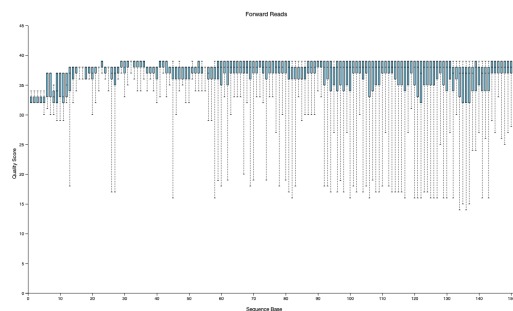
Figure 9: Sequence Quality Plot from EBI Data
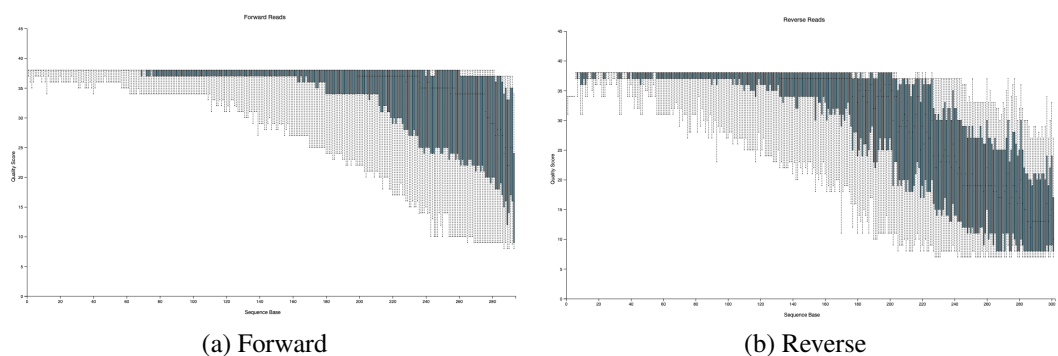

(a) Forward   (b) Reverse

Figure 10: Sequence Quality Plot from HMP Data

Table 2: Trimming Lengths of Databases

	Sequence Quality Plot	Trimming Length
Helixco Data	265	Figure 8
EBI Data	150	Figure 9
HMP Data	226	Figure 10
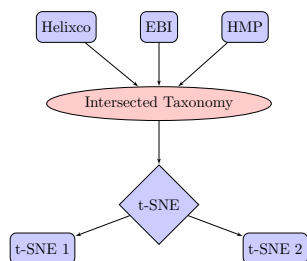
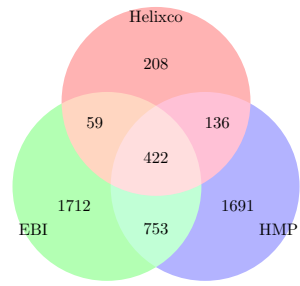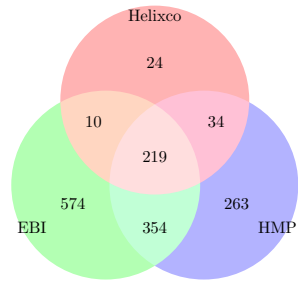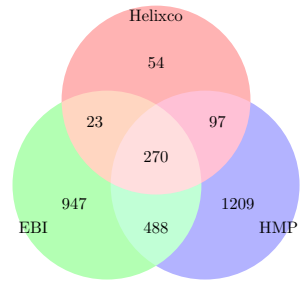Figure 11: Workflow of t-SNE for Brief Information
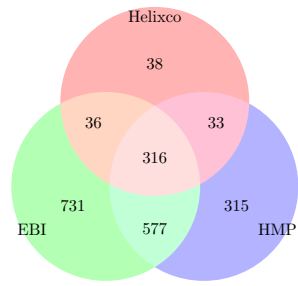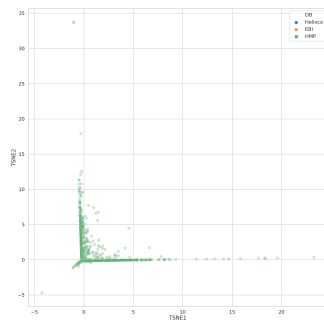
(a) DADA2 + GG
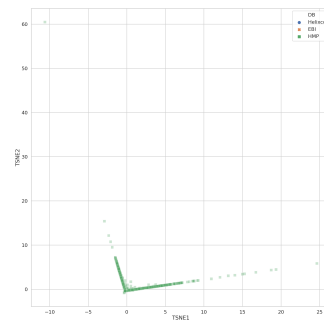
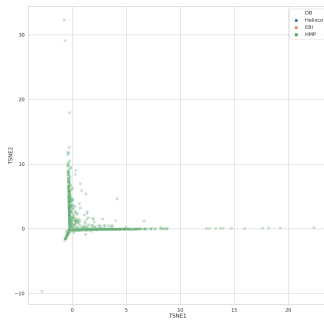(b) DADA2 + SILVA

(c) Deblur + GG

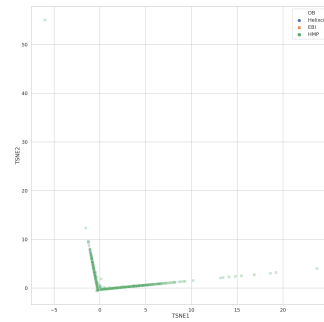(d) Deblur + SILVA

Figure 12: Count of Intersected Taxonomies

(a) DADA2 + GG

(b) DADA2 + SILVA

(c) Deblur + GG

(d) Deblur + SILVA

Figure 13: t-SNE for Comparing Databases

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.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. Retrieved from `https://aem.asm.org/content/75/23/7537` doi: 10.1128/AEM.01541-09

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