

Microbiome Premature

Jaewoong Lee

Semin Lee

2023-01-16

Contents

1	Introduction	3
1.1	Microbiome	3
1.2	rRNA	3
1.3	Premature	3
2	Materials	3
2.1	16S rRNA Sequencing	3
2.1.1	JBNU/Helixco Data	3
2.1.2	EBI Data	3
2.1.3	HMP Data	3
3	Methods	3
3.1	Docker	3
3.2	QIIME 2	3
3.3	Denoising Algorithms	3
3.3.1	DADA2	3
3.3.2	Deblur	5
3.4	Taxonomy Classification Algorithms	5
3.4.1	Greengenes	5
3.4.2	SILVA	5
3.5	Merging Taxonomy	5
3.6	t-SNE	5
3.7	Python Packages	5
3.7.1	Pandas	5
3.7.2	Scikit-Learn	5
3.7.3	Matplotlib	5
3.7.4	Seaborn	5
4	Results	5
4.1	Deciding Trimming Length	5
4.2	t-SNE for Comparing Databases	8
5	Discussion	8
5.1	t-SNE for Comparing Databases	8

List of Tables

1	Metadata of Data	4
2	Trimming Lengths of Databases	6

List of Figures

1	Concept of a core human microbiome. (?, ?)	4
2	Definition of Premature (?, ?)	4
3	Workflow of QIIME2	4
4	Denoising Algorithms	6
5	Taxonomy Classification Algorithms	6
6	Example Diagram for Merging Taxonomy	6
7	t-SNE Visualizations of handwritten digits from MNIST data (?, ?)	6
8	Sequence Quality Plot from First JBNU/Helixco Data	7
9	Sequence Quality Plot from Second JBNU/Helixco Data	7
10	Sequence Quality Plot from Stool JBNU/Helixco Data	7
11	Sequence Quality Plot from EBI Data	7
12	Sequence Quality Plot from HMP Data	9
13	Workflow of t-SNE for Brief Information	9
14	Count of Intersected Taxonomies	9
15	t-SNE for Comparing Databases	10

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 (?). Moreover, the microbiome, the collective genome from these microbiota (?), serve as the trait of individual have not to evolve on their own (?). 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* (?). 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 (?). 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 (?). Hence, 16S rRNA sequencing is the reference method for bacterial taxonomy classification and identification (?).

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 JBNU/Helixco Data

2.1.2 European Bioinformatics Institute Data

EBI data was collected by European Bioinformatics Institute (EBI) (?). 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) (?). 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 (?).

3.2 QIIME 2

QIIME 2 is a next-generation microbiome bioinformatics platform which is extensible, free, open-source, and community developed (?, ?, ?).

3.3 Denoising Algorithms

There are two denoising algorithms which are provided by QIIME as figure 4: DADA2 (?) and Deblur (?).

3.3.1 DADA2

DADA2 is an open-source software package for modeling and correcting Illumina-sequenced amplicon errors (?).

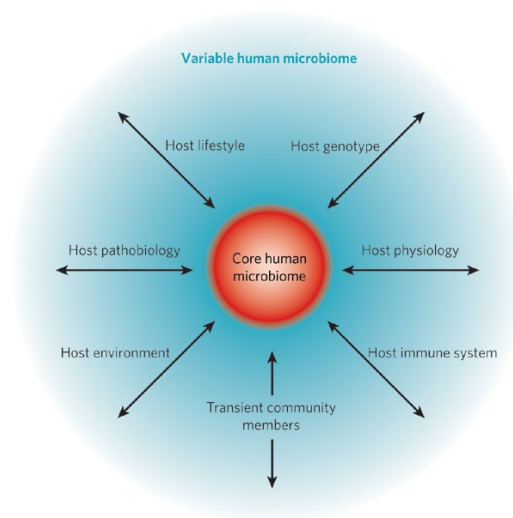


Figure 1: Concept of a core human microbiome. (?, ?)

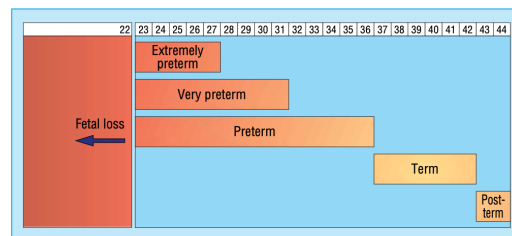


Figure 2: Definition of Premature (?, ?)

Data	Participants	Samples	Remarks
First	24	107	-
Second	35	288	-
Third	10	106	-
Stool	63	126	Stool

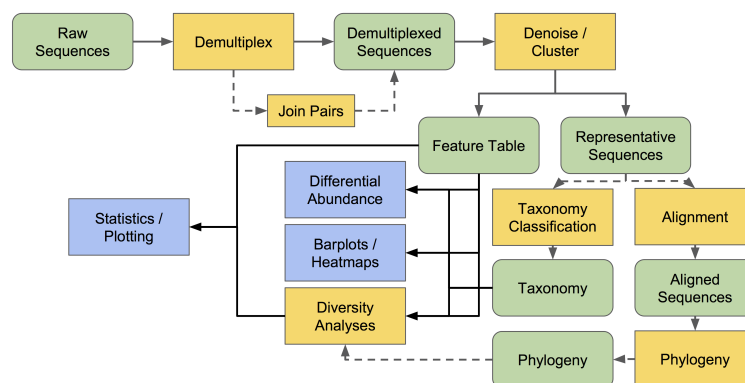


Figure 3: Workflow of QIIME2

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 (? , ?).

3.4 Taxonomy Classification Algorithms

There are two taxonomy classification algorithms which are provided by QIIME as figures 5: Greengenes (? , ?) and SILVA (? , ? , ?).

3.4.1 Greengenes

Greengenes (GG) is a chimera-checked 16S rRNA gene database (? , ?).

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 (? , ? , ?).

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 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 (? , ?).

3.7 Python Packages

3.7.1 Pandas

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

3.7.2 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 (? , ?).

3.7.3 Matplotlib

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

3.7.4 Seaborn

Seaborn is a Python data visualization library based on Matplotlib (? , ?).

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.

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

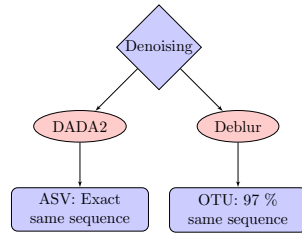


Figure 4: Denoising Algorithms

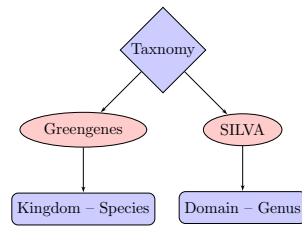


Figure 5: Taxonomy Classification Algorithms

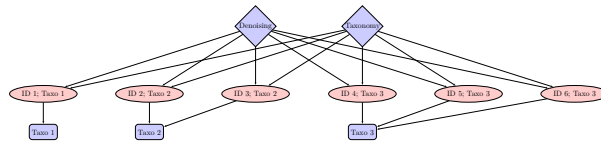


Figure 6: Example Diagram for Merging Taxonomy

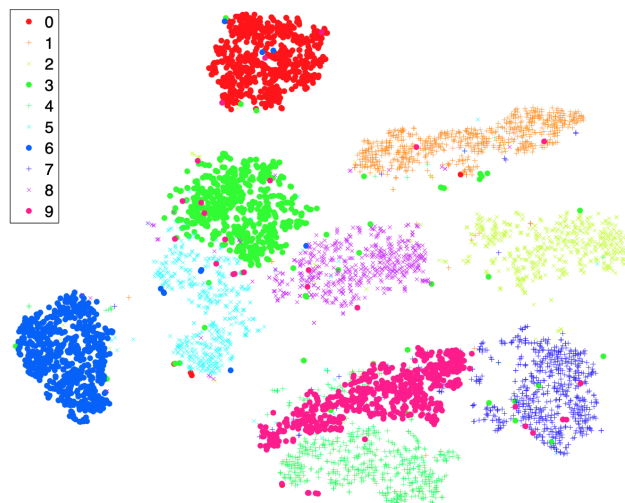


Figure 7: t-SNE Visualizations of handwritten digits from MNIST data (?, ?)

Table 2: Trimming Lengths of Databases		
	Sequence Quality Plot	Trimming Length
Helixco Data	265	Figure 8
EBI Data	150	Figure 11
HMP Data	226	Figure 12

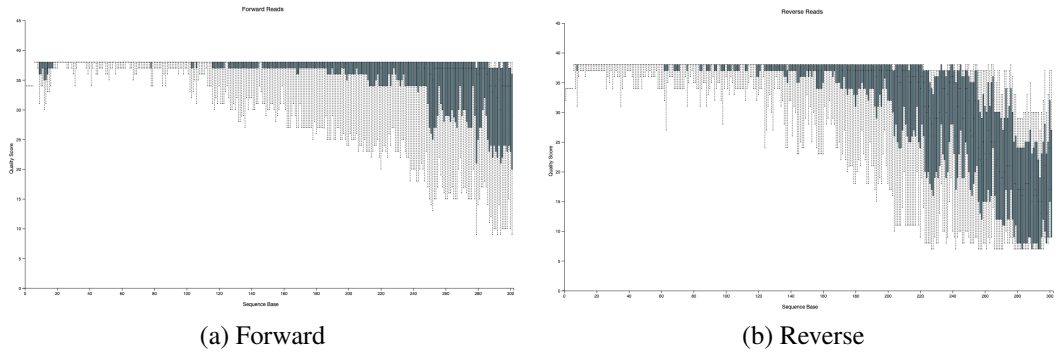


Figure 8: Sequence Quality Plot from First JBNU/Helixco Data

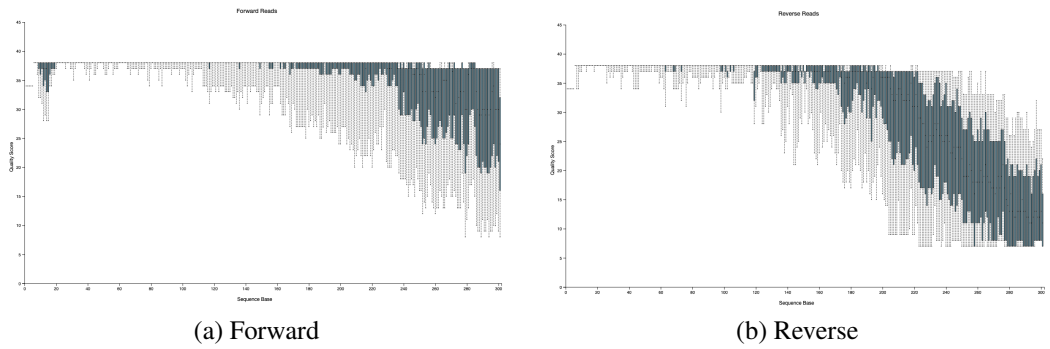


Figure 9: Sequence Quality Plot from Second JBNU/Helixco Data

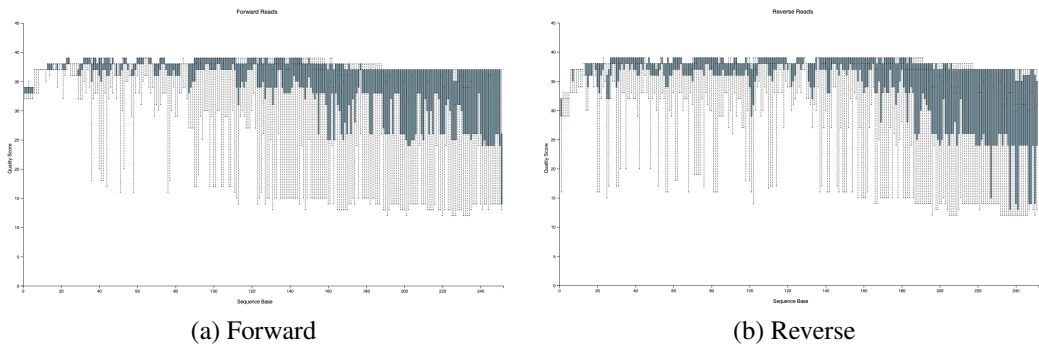


Figure 10: Sequence Quality Plot from Stool JBNU/Helixco Data

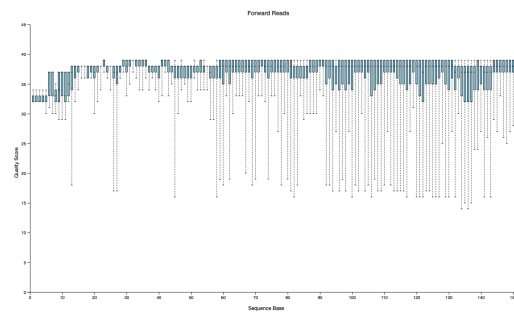


Figure 11: Sequence Quality Plot from EBI Data

4.2 t-SNE for Comparing Databases

To compare three databases, workflow, which as figure 13, 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 14. There are around 200 - 400 intersected taxa. With these intersected taxa, t-SNE plot was derived as figure 15.

5 Discussion

5.1 t-SNE for Comparing Databases

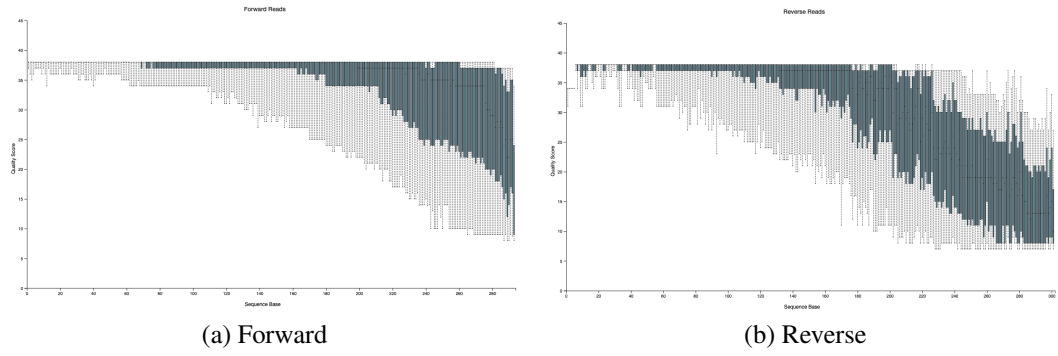


Figure 12: Sequence Quality Plot from HMP Data

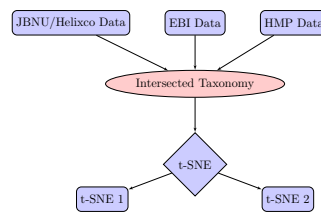


Figure 13: Workflow of t-SNE for Brief Information

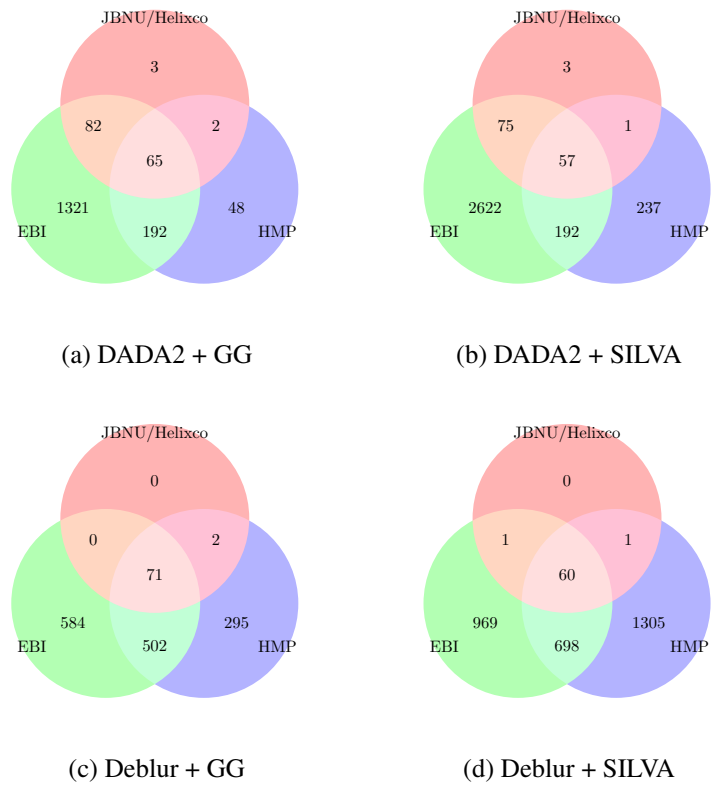


Figure 14: Count of Intersected Taxonomies



Figure 15: t-SNE for Comparing Databases