

# Microbiome Premature

Jaewoong Lee

Ulsan National Institute of Science and Technology

*jwlee230@unist.ac.kr*

2020-10-06

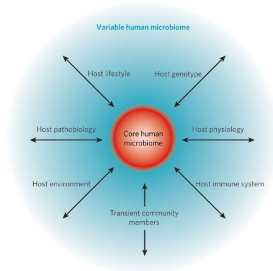
# Overview

- 1 Introduction
  - 2 Materials
  - 3 Literature Survey
  - 4 Methods
  - 5 Results
  - 6 Proceedings
- References

# Introduction

# Microbiome

- Microbiota: the microorganisms which live inside & on humans (Turnbaugh et al., 2007)
- Microbiome:  $10^{13}$  to  $10^{14}$  microorganisms whose 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:  $\geq 37$  weeks

# Materials

# 16S rRNA Sequencing

**16S rRNA sequencing** is the *reference method* for bacterial taxonomy & identification (Mignard & Flandrois, 2006)

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

- 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

Table: Metadata of Data

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

# Literature Survey

## BRIEF COMMUNICATIONS

nature.  
medicine

### Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer

Maria G Dominguez-Bello<sup>1,2</sup>, Kassandra M De Jesus-Laboy<sup>2</sup>, Nan Shen<sup>3</sup>, Laura M Cox<sup>1</sup>, Amnon Amir<sup>4</sup>, Antonio Gonzalez<sup>4</sup>, Nicholas A Bokulich<sup>1</sup>, Se Jin Song<sup>4,5</sup>, Marina Hoashi<sup>1,6</sup>, Juana I Rivera-Vinas<sup>7</sup>, Keimari Mendez<sup>7</sup>, Rob Knight<sup>4,8</sup> & Jose C Clemente<sup>3,9</sup>

estimated 15% of births that require C-section delivery to protect the health of the mother or baby<sup>11</sup>.

Here we exposed C-section-delivered infants to their maternal vaginal fluids at birth and longitudinally determined the composition of their microbiota to assess whether it developed more similarly to vaginally born babies than to unexposed C-section-delivered infants. We collected samples from 18 infants and their mothers, including 7 born vaginally and 11 delivered by scheduled C-section, of which four were exposed to the maternal vaginal fluids at birth (Supplementary Table 1). Briefly, the microbial restoration procedure, or vaginal microbial transfer, consists of incubating sterile gauze in the vagina of moth-

- Study Objectives
  - ① Compare Vaginally vs. Cesarean-section (C-section)
  - ② Restore the microbiota of C-section
- Microbial restoration procedure
  - ① Measure maternal vaginal pH
  - ② Put sterile gauze with saline solution in vagina for 1 hour
  - ③ Swab the infant with the gauze
- Sample collection procedure
  - ① Sample at right after birth, day 3 and weekly for the first month
  - ② Sample from oral, forehead, arm, foot and anal
- Notable Methods/Results
  - ① Using distance methods: e.g. UniFrac distance, Hamming distance

## ARTICLES

<https://doi.org/10.1038/s41591-019-0450-2>

nature  
medicine

OPEN

## The vaginal microbiome and preterm birth

Jennifer M. Fettweis<sup>1,2,3</sup>, Myrna G. Serrano<sup>1,3</sup>, J. Paul Brooks<sup>3,4</sup>, David J. Edwards<sup>3,5</sup>,  
Philippe H. Girerd<sup>2,3</sup>, Hardik I. Parikh<sup>1</sup>, Bernice Huang<sup>1</sup>, Tom J. Arodz<sup>3,6</sup>, Laahirie Edupuganti<sup>1,3</sup>,  
Abigail L. Glascock<sup>7</sup>, Jie Xu<sup>3,8,9</sup>, Nicole R. Jimenez<sup>1,3</sup>, Stephany C. Vivadellj<sup>1,3</sup>, Stephen S. Fong<sup>3,10</sup>,  
Nihar U. Sheth<sup>11</sup>, Sophonie Jean<sup>1</sup>, Vladimir Lee<sup>1,3</sup>, Yahya A. Bokhari<sup>6</sup>, Ana M. Lara<sup>1</sup>, Shreni D. Mistry<sup>1</sup>,  
Robert A. Duckworth III<sup>1</sup>, Steven P. Bradley<sup>1</sup>, Vishal N. Koparde<sup>11</sup>, X. Valentine Orendo<sup>11</sup>,  
Sarah H. Milton<sup>2</sup>, Sarah K. Rozycki<sup>12</sup>, Andrey V. Matveyev<sup>1</sup>, Michelle L. Wright<sup>13,14,15</sup>,  
Snehalata V. Huzurbazar<sup>16</sup>, Eugenie M. Jackson<sup>16</sup>, Ekaterina Smirnova<sup>17,18</sup>, Jonas Korlach<sup>19</sup>,  
Yu-Chih Tsai<sup>19</sup>, Molly R. Dickinson<sup>1</sup>, Jamie L. Brooks<sup>1</sup>, Jennifer I. Drake<sup>1</sup>, Donald O. Chaffin<sup>20</sup>,  
Amber L. Sexton<sup>20</sup>, Michael G. Gravett<sup>20,21</sup>, Craig E. Rubens<sup>20</sup>, N. Romesh Wijesooriya<sup>9</sup>,  
Karen D. Hendricks-Muñoz<sup>3,8,9</sup>, Kimberly K. Jefferson<sup>1,3</sup>, Jerome F. Strauss III<sup>2,3</sup> and Gregory A. Buck<sup>1,3,6\*</sup>

# HMP Data (Fettweis et al., 2019) II

- Study Objectives

- ① Predicting & Preventing premature
- ② Report community resources
- ③ Provide an analysis of the longitudinal, comprehensive, multi-omic profiling of vaginal samples

- Sample collection Procedure

- ① Premature birth vs. Matched normal birth
- ② Ethnically diverse cohort

- Notable Methods/Results

- ① Imitate figures

# HMP Data (Fettweis et al., 2019) III

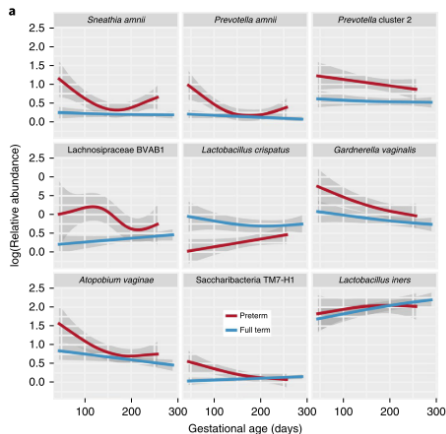
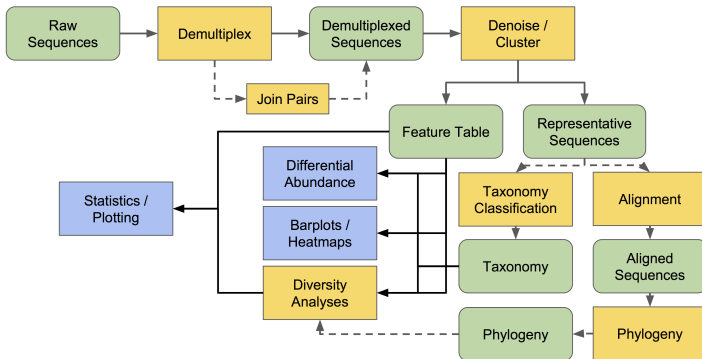


Figure: Microbiome Composition during Pregnancy

# Methods



# Qiime 2 Workflow



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

# Filtering with Quality Score I

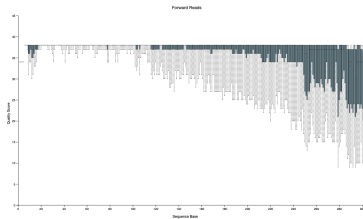
Drawback between:

- Longer sequence read
- Higher quality value

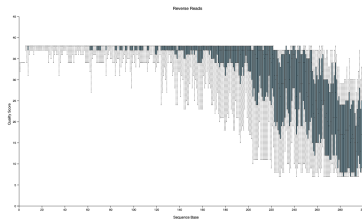
∴ Select the maximum length  $n$  where:

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

# Filtering with Quality Score II



(a) Forward



(b) Reverse

Figure: Sequence Quality Plot from Helixco Data

Maximum Length: 265

# Filtering with Quality Score III

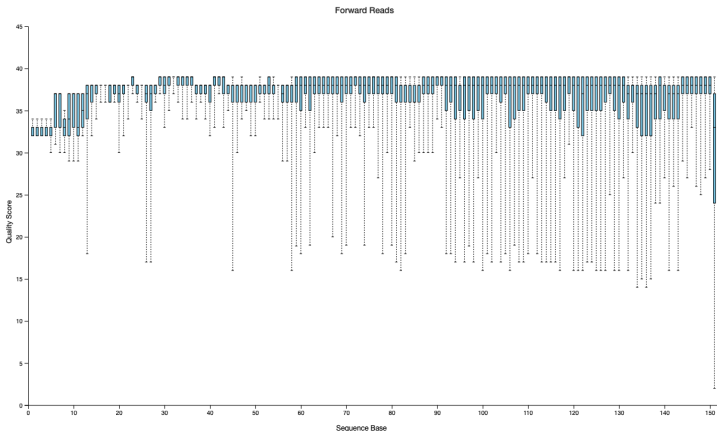
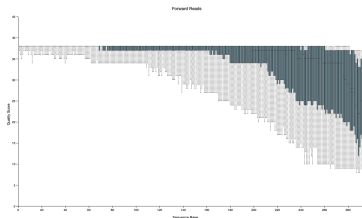


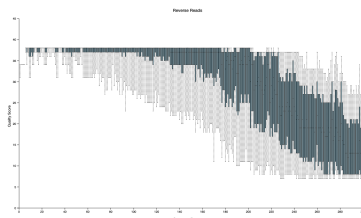
Figure: Sequence Quality Plot from EBI

Maximum Length: 150

# Filtering with Quality Score IV



(a) Forward



(b) Reverse

Figure: Sequence Quality Plot from HMP Data

Maximum Length: 226

# Denoising Techniques

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

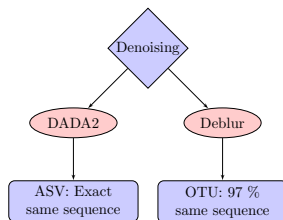


Figure: Denoising Algorithms

# Taxonomy Classification

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

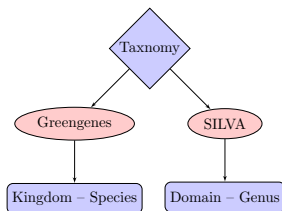


Figure: Taxonomy Classifications

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



Figure: Mothur

Note: Still in progress



# t-distributed Stochastic Neighbor Embedding (t-SNE)

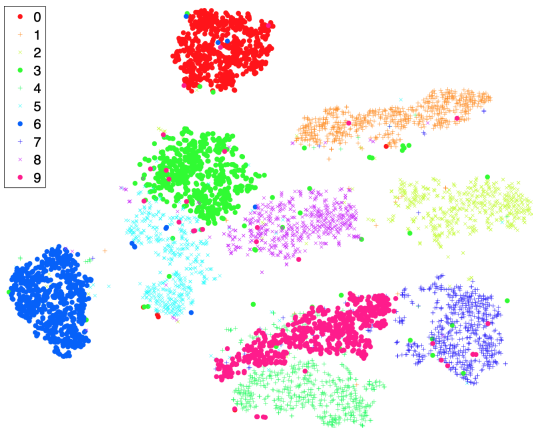


Figure: t-SNE with handwritten data (Maaten & Hinton, 2008)

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

## Results

# t-SNE for Brief Information I

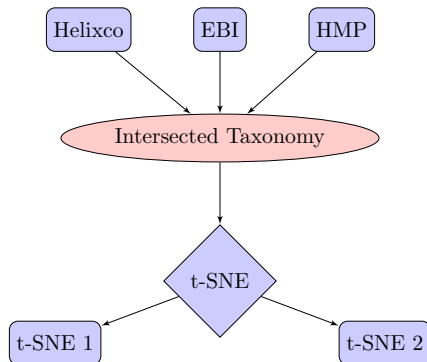


Figure: Workflow of t-SNE for Brief Information

# t-SNE for Brief Information II



(a) DADA2 + GG



(b) DADA2 + SILVA



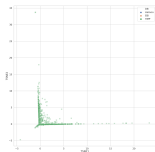
(c) Deblur + GG



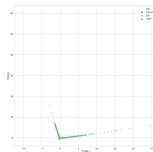
(d) Deblur + SILVA

Figure: Intersected Taxa Information

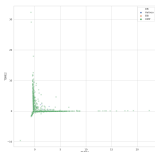
# t-SNE for Brief Information III



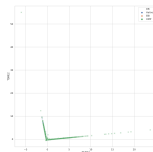
(a) DADA2 + GG



(b) DADA2 + SILVA



(c) Deblur + GG



(d) Deblur + SILVA

Figure: t-SNE for Brief Information

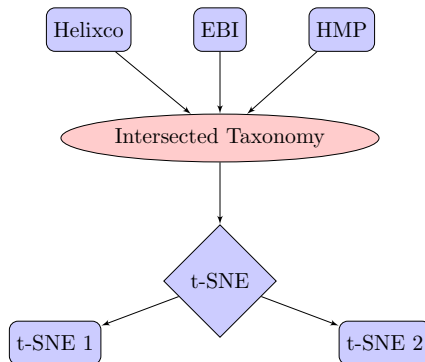
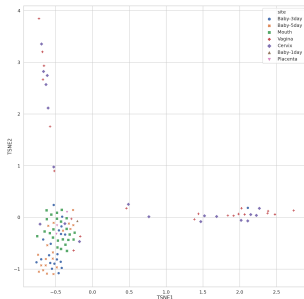
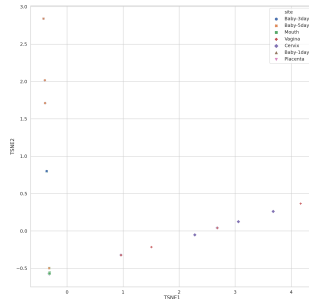


Figure: Workflow of t-SNE for Site Information

# t-SNE with Site II



(a) DADA2 + GG

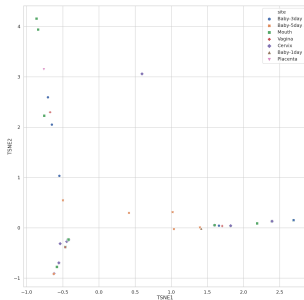


(b) DADA2 + SILVA

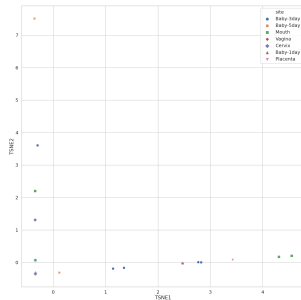
Figure: t-SNE with Site



# t-SNE with Site III



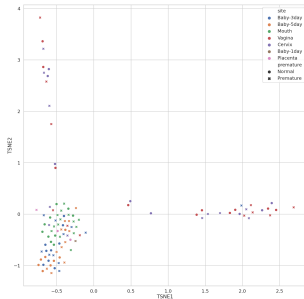
(c) Deblur + GG



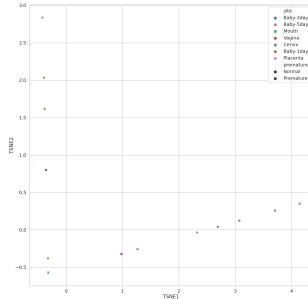
(d) Deblur + SILVA

Figure: t-SNE with Site

# t-SNE with Site + Premature



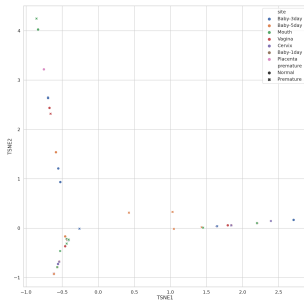
(a) DADA2 + GG



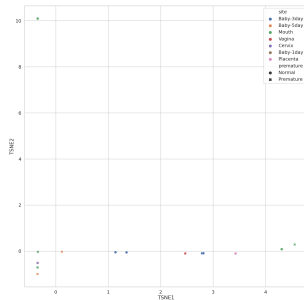
(b) DADA2 + SILVA

Figure: t-SNE with Site + Premature

# t-SNE with Site + Premature II



(c) Deblur + GG



(d) Deblur + SILVA

Figure: t-SNE with Site + Premature

# Random Forest Classifier I

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

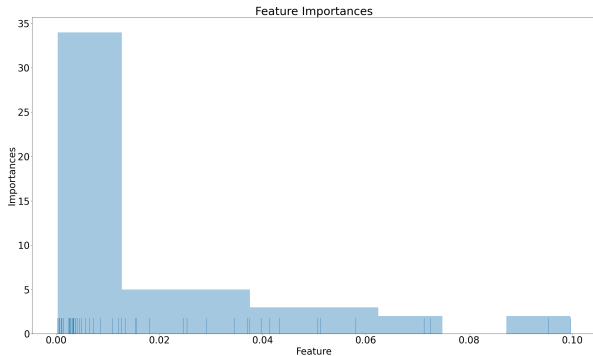


Figure: Feature Importance derived by Random Forest Classifier

# Random Forest Classifier II

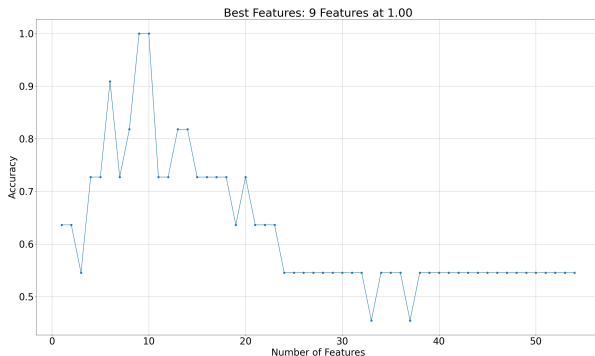


Figure: Number of Features vs. Accuracy

# Random Forest Classifier III

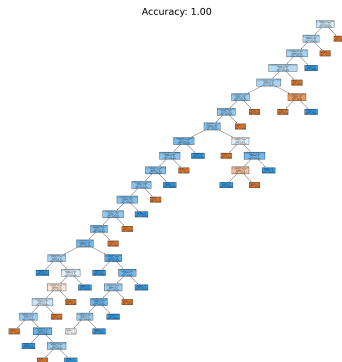


Figure: Random Forest Classifier

# Random Forest Classifier IV

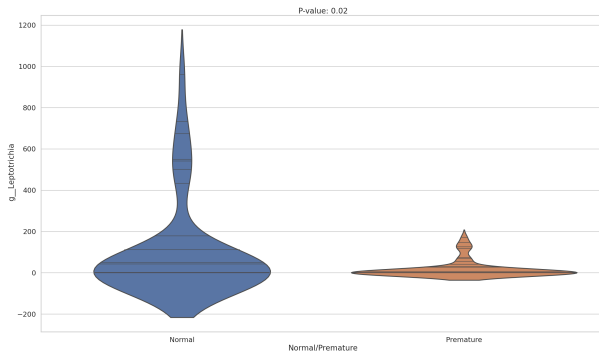


Figure: Violin Plot of *Leptotrichia*

*Fusobacteriota Fusobacteriia Fusobacteriales Leptotrichiaceae Leptotrichia*

# Proceedings



- t-SNE with databases
- Random Forest Classifier

# Requirements I

- More data
- Metadata for databases
- Mothur pipeline

- 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.
- 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. 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).
- 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.
- 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.
- 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