Microbiome Premature

Jaewoong Lee

2020-09-07

Contents

List of Tables

List of Figures

1 Introduction

1.1 Microbiome

1.2 Ribosomal RNA

1.3 Premature

2 Materials

2.1 16S rRNA Sequencing

3 Methods

3.1 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).

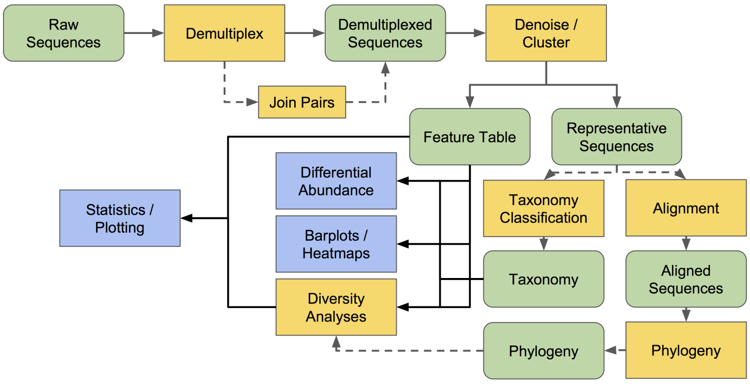

Figure 1: Workflow of QIIME2

3.2 Denoising Techniques

3.2.1 DADA2

DADA2 is an open-source software package for modeling and correcting Illumina-sequenced amplicon erros (Callahan et al., 2016).

3.2.2 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).

3.3 Taxonomy Classification

3.3.1 Greengenes

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

3.3.2 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).

3.4 Mothur

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

3.5 t-distributed Stochastic Neighbor Embedding

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

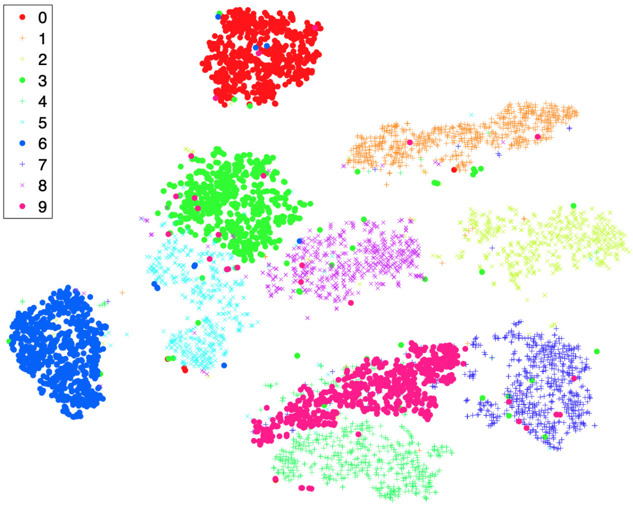

Figure 2: t-SNE Visualizations of handwritten digits from MNIST data (Maaten & Hinton, 2008)

3.6 Python Packages

3.6.1 Pandas

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

3.6.2 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).

3.6.3 Matplotlib

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

3.6.4 Seaborn

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

4 Results

5 Discussion

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.

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).

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.

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

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