Introduction

Nate Olson

2017-11-02

Metagenomics, sequencing the DNA from a microibal community, has greatly advanced our understanding of the microbial world. Targeted sequencing of the 16S rRNA gene, 16S metagenomics, is a commonly used method for sequencing a microbial community, as the targeted approach allows for a more in-depth exploration of a microbial community taxonomic compositions compared to shotgun metagenomics where whole genomes are sequenced. 16S metagenomics is a complex measurement process comprised of multiple molecular laboratory and computation steps (Julia K Goodrich et al. 2014; Kim et al. 2017). There are numerous sources of error and bias in the measurement process, for both the molecular laboratory (e.g. PCR and sequencing) and computational steps (e.g. sequence clustering) (D'Amore et al. 2016; Julia K. Goodrich et al. 2014; Brooks et al. 2015). Appropriate methods are needed to evaluate the 16S measurement process inorder to characterize how these sources of bias and error impact the measurement results and determine where to focus communities efforts for improving the measurement process.

A key step in the measurement process is clustering, the grouping of sequences into biologically relevant units, or operational taxonomic units (OTUs). There are a number of different clustering methods.
The two most commonly used clustering methods are *de-novo* clustering and open-reference clustering. *de-novo* clustering algorithms group sequences based on a defined similarity threshold (Westcott and Schloss 2015). Open-reference clustering matches sequences to a set of previously clustered reference sequences (*de-novo*) then perform *de-novo* clustering on the sequences in the dataset that do not match to sequences in the reference dataset with the desired similarity threshold (He et al. 2015). A third methods for clustering, sequence inference, uses statistical models or algorithms to differentiate true biological sequences within a dataset from sequencing errors (Callahan et al. 2016; Amir et al. 2017; Eren et al. 2015).

Further challenging the measurement process is the compositional nature of the 16S data, that is the proporition of an organism within a sample is being measured and not the absolute abundance (Tsilimigras and Fodor 2016). Sequencing data only provide information regarding the relative abundance of organisms within a samples to other organisms within the same sample. When comparing the relative abundance of an organism across samples you are comparing organismal abundance relative to the rest of the organisms within the sample. As a result an organism can have the same absolute abundance in two samples but due to differences in either the microbial community composition or for targeted assays such as 16S metagenomics differences in the proportion of human DNA in the sample.

In order to characterize the accuracy of a measurement process you need a sample or dataset with an expected value to benchmark against. There have been a number of studies characterizing and evaluating different steps in the 16S rRNA metagenomics measurement process all of which use mock communities, simulated data, or environmental samples. Mock communities consisting of mixtures of cells or DNA from individual organisms and simulated data have been previously used to evaluate different aspects of the measurement process (Bokulich et al. 2016). Mock communities have an expected value but are not representative of the complexity of environmental samples in terms of the of number or abundance distributions of organisms. Similar to mock communities simulated data have an expected value that can be used for benchmarking. However, the sequencing error profile is not completely understood and therefore simulated sequencing data does not recapitulate the complexity of sequencing data generated from an environmental sample. While simulated data and mock communities are usefull in evaluating and benchmarking new methods one needs to consider that methods optimized for mock communities are not necessarily optimized to handle the additional biases, noise, and diversity present in real samples. Data generated from environmental samples, which include the biase, error, and diversity of real samples, are often used to benchmark new molecular laboratory and computational methods. However, without an expected value to compare to only measurement precision can be evaluated.

An alternative to these types of data is sequencing data generated from mixtures of environmental samples.

By mixing environmental samples at known proportions you can use information obtained from the unmixed samples and how they were mixed to obtain an expected value for use in assessing the measurement process. Mixtures of environmental samples have previously been used to evaluate gene expression measurements microarrays and RNAseq (Parsons et al. 2015; Pine, Rosenzweig, and Thompson 2011; Thompson et al. 2005)

- Application to 16S
  - We generated a data set using mixtures of extracted DNA from human stool samples for assessing the 16S metagenomic measurement process.
  - Processed the resulting dataset with three bioinformatic pipelines and performed a quantitative and qualitative assessment of the resulting count tables.
  - Results indicate that ....

Amir, Amnon, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, James T Morton, Zhenjiang Zech Xu, Eric P Kightley, et al. 2017. "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns." *mSystems* 2 (2). Am Soc Microbiol: e00191–16.

Bokulich, Nicholas A, Jai Ram Rideout, William G Mercurio, Arron Shiffer, Benjamin Wolfe, Corinne F Maurice, Rachel J Dutton, Peter J Turnbaugh, Rob Knight, and J Gregory Caporaso. 2016. "Mockrobiota: A Public Resource for Microbiome Bioinformatics Benchmarking." *mSystems* 1 (5). Am Soc Microbiol: e00062–16.

Brooks, J Paul, David J Edwards, Michael D Harwich, Maria C Rivera, Jennifer M Fettweis, Myrna G Serrano, Robert A Reris, et al. 2015. "The Truth About Metagenomics: Quantifying and Counteracting Bias in 16S rRNA Studies." *BMC Microbiology* 15 (1). BioMed Central: 66.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods*. Nature Publishing Group.

D'Amore, Rosalinda, Umer Zeeshan Ijaz, Melanie Schirmer, John G Kenny, Richard Gregory, Alistair C Darby, Christopher Quince, and Neil Hall. 2016. "A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling." *BMC Genomics* 17. BMC Genomics: 1–40. doi:10.1186/s12864-015-2194-9.

Eren, A Murat, Hilary G Morrison, Pamela J Lescault, Julie Reveillaud, Joseph H Vineis, and Mitchell L Sogin. 2015. "Minimum Entropy Decomposition: Unsupervised Oligotyping for Sensitive Partitioning of High-Throughput Marker Gene Sequences." *The ISME Journal* 9 (4). Nature Publishing Group: 968–79.

Goodrich, Julia K, Sara C Di Rienzi, Angela C Poole, Omry Koren, William A Walters, J Gregory Caporaso, Rob Knight, and Ruth E Ley. 2014. "Conducting a Microbiome Study." *Cell* 158 (2). Elsevier: 250–62.

Goodrich, Julia K., Sara C. Di Rienzi, Angela C. Poole, Omry Koren, William A. Walters, J. Gregory Caporaso, Rob Knight, and Ruth E Ley. 2014. "Conducting a Microbiome Study." *Cell* 158 (2). Elsevier Inc.: 250–62. doi:10.1016/j.cell.2014.06.037.

He, Yan, J Gregory Caporaso, Xiao-Tao Jiang, Hua-Fang Sheng, Susan M Huse, Jai Ram Rideout, Robert C Edgar, et al. 2015. "Stability of Operational Taxonomic Units: An Important but Neglected Property for Analyzing Microbial Diversity." *Microbiome* 3 (1). BioMed Central: 20.

Kim, Dorothy, Casey E Hofstaedter, Chunyu Zhao, Lisa Mattei, Ceylan Tanes, Erik Clarke, Abigail Lauder, et al. 2017. "Optimizing Methods and Dodging Pitfalls in Microbiome Research." *Microbiome* 5 (1). BioMed Central: 52.

Parsons, Jerod, Sarah Munro, P Scott Pine, Jennifer McDaniel, Michele Mehaffey, and Marc Salit. 2015. "Using Mixtures of Biological Samples as Process Controls for Rna-Sequencing Experiments." *BMC Genomics* 16 (1). BioMed Central: 708.

Pine, P Scott, Barry A Rosenzweig, and Karol L Thompson. 2011. "An Adaptable Method Using Human Mixed Tissue Ratiometric Controls for Benchmarking Performance on Gene Expression Microarrays in Clinical Laboratories." *BMC Biotechnology* 11 (1). BioMed Central: 38.

Thompson, Karol L, Barry A Rosenzweig, P Scott Pine, Jacques Retief, Yaron Turpaz, Cynthia A Afshari, Hisham K Hamadeh, et al. 2005. "Use of a Mixed Tissue Rna Design for Performance Assessments on Multiple Microarray Formats." *Nucleic Acids Research* 33 (22). Oxford University Press: e187–e187.

Tsilimigras, Matthew CB, and Anthony A Fodor. 2016. "Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges." *Annals of Epidemiology* 26 (5). Elsevier: 330–35.

Westcott, Sarah L, and Patrick D Schloss. 2015. "De Novo Clustering Methods Outperform Reference-Based Methods for Assigning 16S rRNA Gene Sequences to Operational Taxonomic Units." *PeerJ* 3. PeerJ Inc.: e1487.