Complete Microbiome Analysis

A Repository of Resources for Every Step of Microbiome Research

Alana Schick

2022-06-17

Contents

Introduction

Here is an overview of what this book contains...

Overview of sections

Part I

Part 1: Samples to Sequences

Chapter 1

Experimental Design Considerations

1.1 Amplicon versus Shotgun

Amplicon sequencing (often, but not necessarily the amplicon is the 16S rRNA gene) is a type of seuqencing where a specific region is targeted and amplified for sequencing. For 16S, this gene is found in all Bacteria and Archaea and will only identify these types of organisms.

Shotgun metagenomic sequencing This is an untargeted approach that aims to sequence all the genomic DNA present in a sample. Therefore this method is able to identify bacteria, viruses, and fungi. Analysing these reads requires more complex bioinformatics methods and can either be assembled to create partial or full microbial genomes, or aligned to databases of microbial marker genes.

Cost of shotgun depends on the depth of coverage required, which depends on microbial-to-host DNA ratio. Fecal samples have mostly microbial DNA, for example, and therefore require less depth. Skin swabs and cheek swabs may contain more human DNA, so 16S may be more suitable for the skin and oral microbiome. Some estimates: stool samples have less than 10% human DNA, other samples (saliva, throat, vaginal) may be more than 90% human DNA.

Paper investigating the effect of host DNA and sequencing depth on taxonomic resolution of shotgun metagenomic sequencing:

Pereira-Marques et al 2019. Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. Frontiers in Microbiology.

Summary: at a depth of ~30 million reads per sample, taxonomic resolution is similar for a mock community of bacteria only and a sample with 10% host DNA.

Resolution gets worse for samples with 90% or 99% host DNA. The sequencing depth seems to matter more for samples with 90% host DNA.

In addition, there are some methods to reduce the amount of host DNA present in samples prior to sequencing.

Resolution The level (ie. genus, species) of resolution achieved by amplicon sequencing is limited by two things: 1) the variation present across the targeted gene within a genus, species, strain, etc and 2) sequencing error

Composition versus Function Recent evidence suggests that functional metagenomic data may provide more power for identifying differences between healthy and diseased microbiomes. 16S cannot directly profile microbial genes/functions, but some tools (such as Picrust) attempt to predict microbiome function with 16S rRNA gene data. Shotgun can provide data on microbial gene content. Therefore, if the researcher is interested in microbiome functional profiles (such as antibiotic resistance genes or specific metabolic functions), shotgun is the better choice. The limitation is that current databases are still quite limited in identifying many functional genes.

Shallow shotgun sequencing is roughly defined as 1.5 million reads per sample (as few as 0.5 million).

Ultra-deep shotgun metagenomic sequencing would be 2.5 billion reads per sample (so roughly 1000 times more coverage).

1.2 16S Resources

The problem: different regions of the bacterial 16S rRNA gene evolve at different evolutionary rates.

Teng, F., Darveekaran Nair, S.S., Zhu, P. et al. Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. Sci Rep 8, 16321 (2018). https://doi.org/10.1038/s41598-018-34294-x

Tested the effect of DNA extraction method and 16S region on the accuracy of sequencing oral microbial communities. Used a mock community of only gram-positive species. Found that DNA extraction method had a much larger effect on variation in microbial community. Compared lysozyme method to bead-beating - found that bead-beating led to lower DNA yield but highter accuracy. Found V3-V4 and V4-V5 to be more reproducible than V1-V3.

Fouhy, F., Clooney, A.G., Stanton, C. et al. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiol 16, 123 (2016). https://doi.org/10.1186/s12866-016-0738-z

Compared V1-V2, V3-V4, and V4 regions on human fecal samples. Found V4 region samples had higher alpha diversity.

Rintala A, Pietilä S, Munukka E, et al. Gut Microbiota Analysis Results Are Highly Dependent on the 16S rRNA Gene Target Region, Whereas the Impact of DNA Extraction Is Minor. J Biomol Tech. 2017;28(1):19-30. doi:10.7171/jbt.17-2801-003

Compared V3-V4 and V4-V5 regions on human fecal samples. Found higher diversity in V3-V4 samples. The Firmicutes-to-Bacteroidetes ratio was significantly lower in V4–V5 sequencing. More specifically: "In the bacterial genus level, QIIME reported statistically significant differences in 21 genera between the V3–V4 and V4–V5 sequencing protocols. For example, the genus Parabacteroides was significantly more abundant in the samples analyzed with V4–V5 sequencing (FDR, $P < 0.05$), whereas Bifidobacterium, Coprococcus, and Blautia were more abundant in the V3–V4 samples (FDR, $P < 0.05$ for all). In addition, the genera Sphingomonas, Roseburia, and Bilophila were detectable only with V3–V4 sequencing, whereas Clostridium and Lactococcus could only be detected with V4–V5 sequencing."

Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. Primer and platform effects on 16S rRNA tag sequencing. Frontiers in microbiology. 2015 Aug 4;6:771.

Tested V4, V6-V8, and V7-V8 on mock communities. Found beta diversity metrics robust to region used. "V4 samples showed the highest similarity toward the expected taxonomic distribution. The largely bacteria-specific V7–V8 tags failed to amplify Halobacteria as expected, but also severely underrepresented Gammaproteobacteria and/or overrepresented Firmicutes."

Ghyselinck J, Pfeiffer S, Heylen K, Sessitsch A, De Vos P. The effect of primer choice and short read sequences on the outcome of 16S rRNA gene based diversity studies. PloS one. 2013 Aug 19;8(8):e71360.

Tested ten well established universal primers. Found V4 primers to be best and V6 primers least reliable.

Chapter 2

Power Analyses

Resources:

https://medium.com/brown-compbiocore/power-analyses-for-microbiome-studies-with-micropower-8ff28b36dfe3

https://gist.github.com/brendankelly/6673e8596d3cde3fac7d493d7747aa80

http://joey711.github.io/waste-not-supplemental/simulation-cluster-accuracy/simulation-cluster-accuracy-server.html

Chapter 3

Sequencing samples

Part II

Working with Sequence Data

Chapter 4

Filtering and cleaning

Chapter 5

Generating Abundance Tables

5.1 Dada2 notes

Species assignment. The function 'addSpecies' in dada2 has an option to allow multiple species assignments. This can be set to a number (ie. 3) or `TRUE`, the default is `FALSE`). If set to a number, if there are more than that number, NA will be returned.

From dada2 documentation:

allowMultiple

Default FALSE. Defines the behavior when multiple exact matches against different species are returned. By default only unambiguous identifications are return. If TRUE, a concatenated string of all exactly matched species is returned. If an integer is provided, multiple identifications up to that many are returned as a concatenated string.

In general, best practice it to set this to `TRUE`.

Chapter 6

Remove Contaminants

Chapter 7

Functional Annotation

7.1 Amplicon sequences

7.1.1 Picrust2 notes

How it works:

Picrust2 has a reference database of annotated genomes placed on a phylogenetic tree, including a model of how well conserved genes are across the tree. In other words, given how close two species are on the tre, how likely is it that a given gene will be shared, for all genes. The amplicon sequences are placed on the tree, then Picrust2 estimates what genes that species is likely to have based on the annotated genomes of the species near it on the tree and how well conserved those genes are. Doing this for all amplicon sequences will estimate the functional content of a sample.

The problems arise when either A) a species has conserved amplicon sequence but substantially variable gene content, so even if you know where it goes on the tree you cant make a good guess about the presence or absence of those variable genes, or B) your observed amplicon sequence comes from a place on the tree without a lot of information - it's nearest neighbors aren't very near or you don't have enough information to estimate gene conservation well - so you end up with highly speculative, weak estimates, which are probably largely incorrect. If either of those problem cases applies to highly abundant species/16s sequences in your sample, your overall picture of the samples functional content can be way off.

On the other hand, if your sample is mainly comprised of very well characterized species which are well represented in the reference data and which don't have a lot of intraspecies functional variation, you can infer a remarkably accurate functional profile of your sample using picrust. (In terms of functional capacity/genes present at least; you don't know anything about what's actually being

expressed… But shotgun metagenomics doesn't help you there, either.)

7.2 Shotgun sequences

Part III

Working with Abundance Tables

Chapter 8

Alpha diversity

And now for some alpha diversity methods

Chapter 9

Beta Diversity

Chapter 10

Differential Abundance

Comparing methods:

https://www.nicholas-ollberding.com/post/identifying-differentially-abundant-features-in-microbiome-data/

10.1 Corncob notes

Beta binomial regression.

How to interpret coefficients:

By default, corncob uses the logit link between the expected relative abundance and the covariates. This can be interpreted on the log-odds scale, similar to with logistic regression. Typically, I recommend users interpret parameters directly on the logit scale, for example "The expected difference in the logit-transformed relative abundance between two samples that differ by one unit change in [my covariate] controlling for [all controls] is [coefficient value]". If you want to interpret more directly, I'd recommend looking up the interpretation of log-odds, or logistic regression parameters. Unfortunately, the definition is not intuitive, and it is necessary to use a link function in order to appropriately model something on the [0,1] scale, such as relative abundance.

https://github.com/bryandmartin/corncob/issues/68 (Note: I have ideas (but not time) about how to improve this feature)

https://github.com/bryandmartin/corncob/blob/master/R/print_differentialTest.R (This is the code for the function that performs this differential test)

10.2 DESeq2 notes

DESeq2: normalizes by estimating the negative binomial distribution for each taxon in each sample https://bioconductor.org/packages/devel/bioc/vignettes/phyloseq/inst/doc/phyloseq-mixture-models.html

DESEQ analysis on RNA-seq data: https://compbiocore.github.io/deseq-workshop-1/assets/deseq_workshop_1.html

Analyzing RNA-seq data with DESeq2: http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#contrasts

10.3 metagenomeSeq

MetagenomeSeq: uses sample quantiles to normalize accounting for undersampling.

Chapter 11

Normalization Methods

Variance-stabilizing transformation: https://github.com/joey711/phyloseq/issues/283 https://github.com/joey711/phyloseq/issues/492

Chapter 12

Multivariate Analysis

Lots of variables? Use multivariate methods!

Part IV

Advanced Topics

Chapter 13

Reproducible Workflows

13.1 Snakemake

Snakemake is...

Resources: https://vincebuffalo.com/blog/2020/03/04/understanding-snakemake.html

Chapter 14

Multiomic Methods

14.1 Correlation Analysis

14.2 Network Analysis