Amplicon Analysis Workshop

Introduction & DADA2 Pipeline

Crandall Lab, Penn State

Amplicon Analysis Workshop
Introduction & DADA2 Pipeline

Penicillium (Latin: Painter's brush)

Workshop Agenda

1. Unix / R introduction (optional)
2. Microbiome Analysis Basics - From Planning to Sequencing
3. Amplicon Data Analysis in R
4. Amplicon Data Analysis and QIIME 2
5. Shotgun Analysis Binning & Assembly

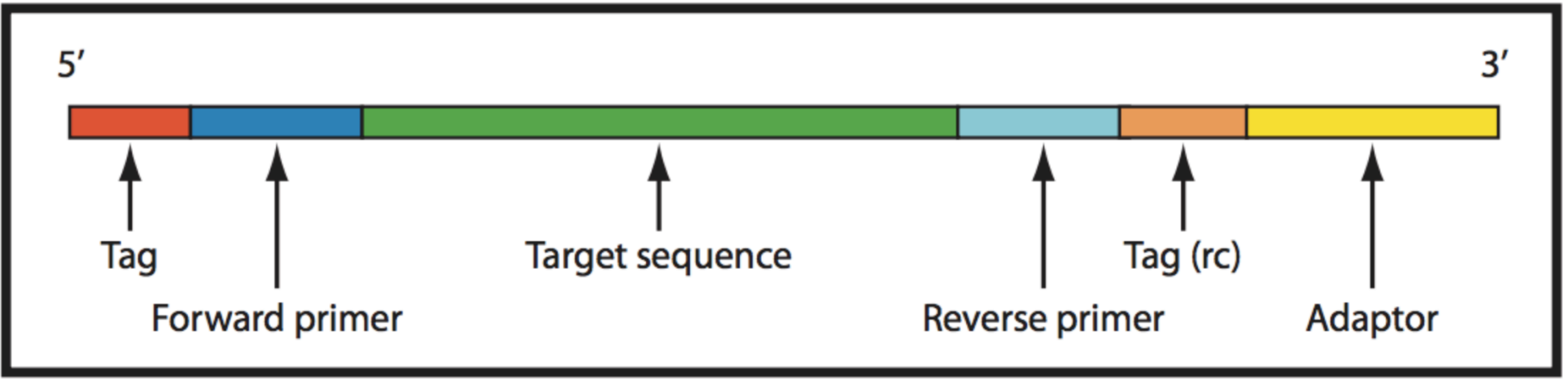
Today's Learning Objectives

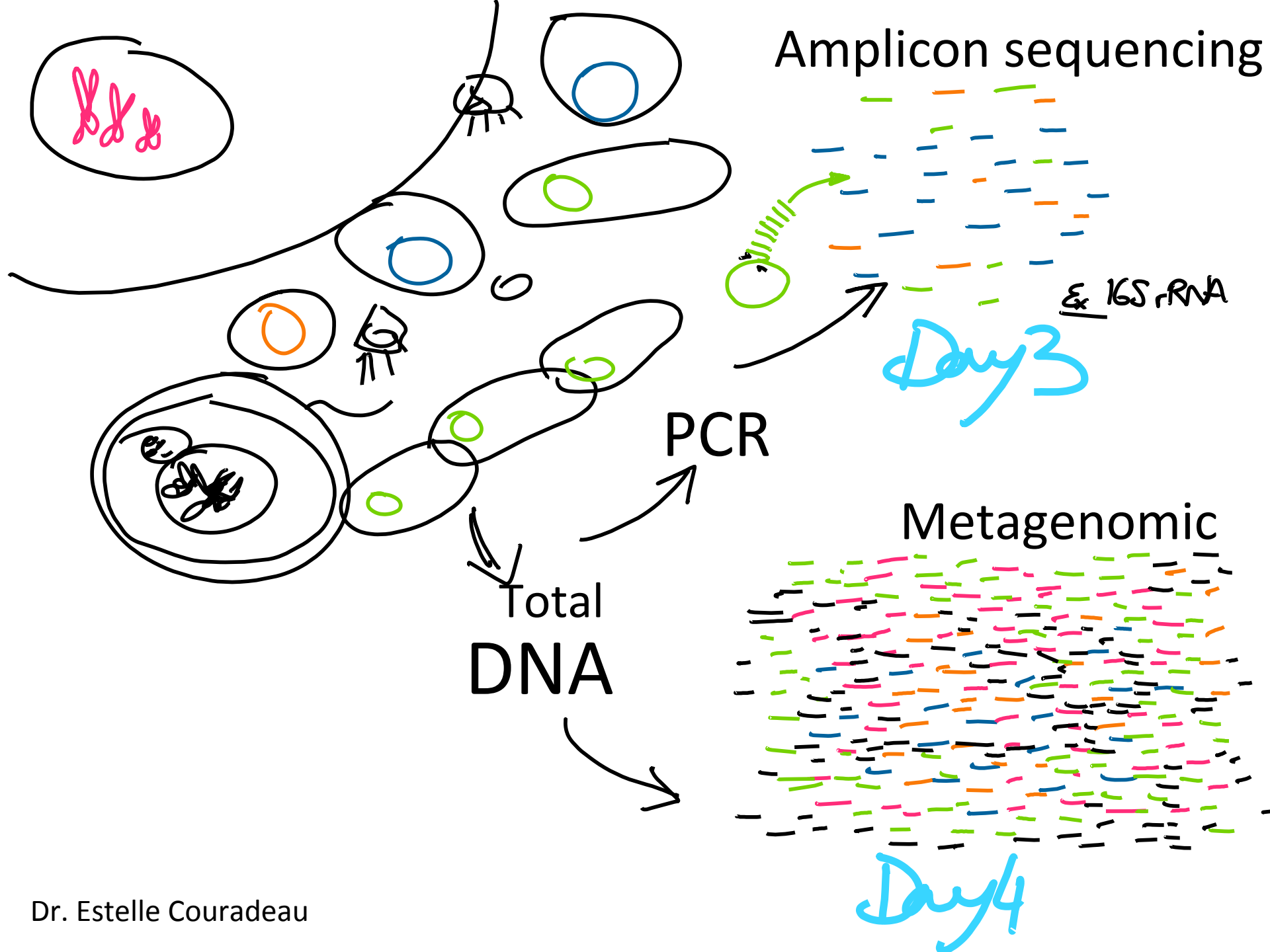
- ✓ Understand the basics of amplicon-based sequencing for microbiome research
- ✓ Run the DADA2 workflow in R – from raw to processed Illumina 16S sequences
- ✓ Explore amplicon data through visual analyses in R

What's an amplicon?

A piece of DNA or RNA that is the source and/or product of amplification or replication events.

It can be formed artificially, using various methods including polymerase chain reactions (PCR) or naturally through gene duplication.



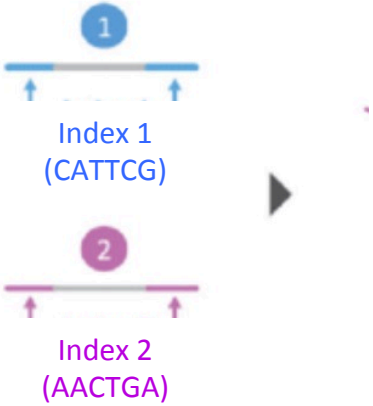


Dr. Estelle Couradeau

DNA Preparation & Sequencing

1

Library Preparation



Next-Gen DNA Sequencing & Analyses

- Illumina Miseq
- ITS1,2 gene region
- DADA2, QIIME2 pipelines
- R

QIAGEN

DADA2
Amplicon Sequencing. Exactly. Version 1.14

Fast and accurate sample inference from amplicon data with single-nucleotide resolution

Quantitative Insights Into Microbial Ecology
qiime

Further reading on available pipelines

PLoS One. 2020; 15(1): e0227434.
Published online 2020 Jan 16. doi: 10.1371/journal.pone.0227434

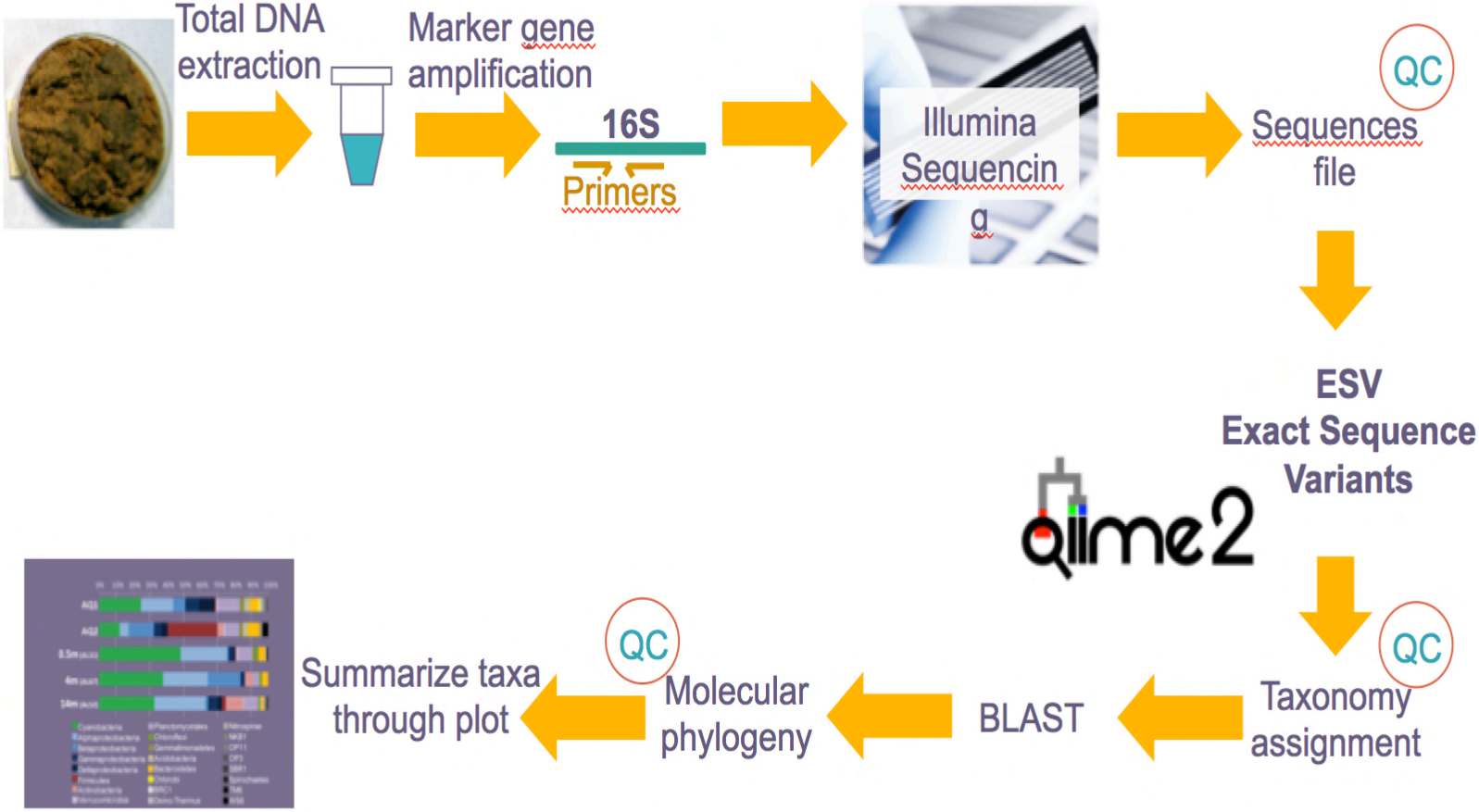
PMCID: PMC6964864
PMID: 31945086

Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing

Andrei Prodan, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing,[1,*] Valentina Tremaroli, Investigation, Resources, Writing – review & editing,[2] Harald Brolin, Writing – review & editing,[2] Aeilko H. Zwinderman, Conceptualization, Writing – review & editing,[3] Max Nieuwdorp, Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing,[1] and Evgeni Levin, Conceptualization, Supervision, Writing – review & editing[1,4]

Jeong-Sun Seo, Editor

Culture Independent description of microbial diversity
16S rDNA libraries Illumina sequencing



Dr. Estelle Couradeau

What is an ASV (ESV)?

- **A**mplicon **S**equence **V**ariant (or **E**xact) differences in genetic sequences

vs.

- **O**perational **T**axonomic **U**nit (OTU) clusters of sequencing reads that differ by less than a fixed dissimilarity threshold

Callahan et al. 2017 ISME

The OTU approach clusters similar reads into one representative group, potentially containing more than one organism from the sample.

'The ASV approach identifies single, exact sequences that are statistically supported as being present in the sample.

Fast and accurate sample inference from amplicon data with single-nucleotide resolution

Divisive Amplicon Denoising Algorithm

- Pipeline control errors sufficiently such (ASVs) can be resolved exactly, down to the level of single-nucleotide differences over the sequenced gene region.
- Finer resolution = = better taxonomic resolution

Callahan et al. 2017 ISME

Paradigm shift from OTU to ASV?

> ISME J. 2017 Dec;11(12):2639-2643. doi: 10.1038/ismej.2017.119. Epub 2017 Jul 21.

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

Benjamin J Callahan [1], Paul J McMurdie [2], Susan P Holmes [3]

Affiliations + expand

PMID: 28731476 PMCID: PMC5702726 DOI: 10.1038/ismej.2017.119

Free PMC article

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan ✉, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson & Susan P Holmes

Nature Methods **13**, 581–583(2016) | Cite this article

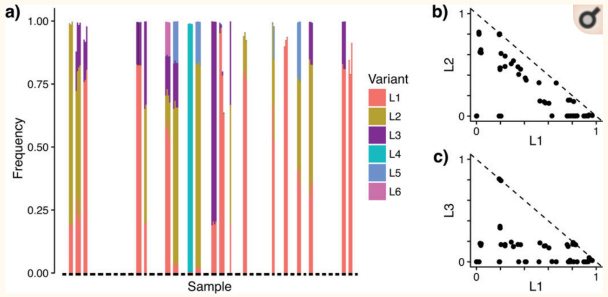


Figure 2

Lactobacillus crispatus sequence variants in the human vaginal community during pregnancy

DADA2 identified six *Lactobacillus crispatus* 16S rRNA sequence variants present in multiple samples and a significant fraction of all reads (L1: 19.7%, L2: 11.1%, L3: 6.5%, L4: 3.1%, L5: 1.3%, L6: 0.4%). (a) The frequency of L1–L6 in each sample. Black bars at the bottom link samples from the same subject. The frequency of (b) L1 vs. L2, and (c) L1 vs. L3, by sample. The dashed line indicates a total frequency of 1.

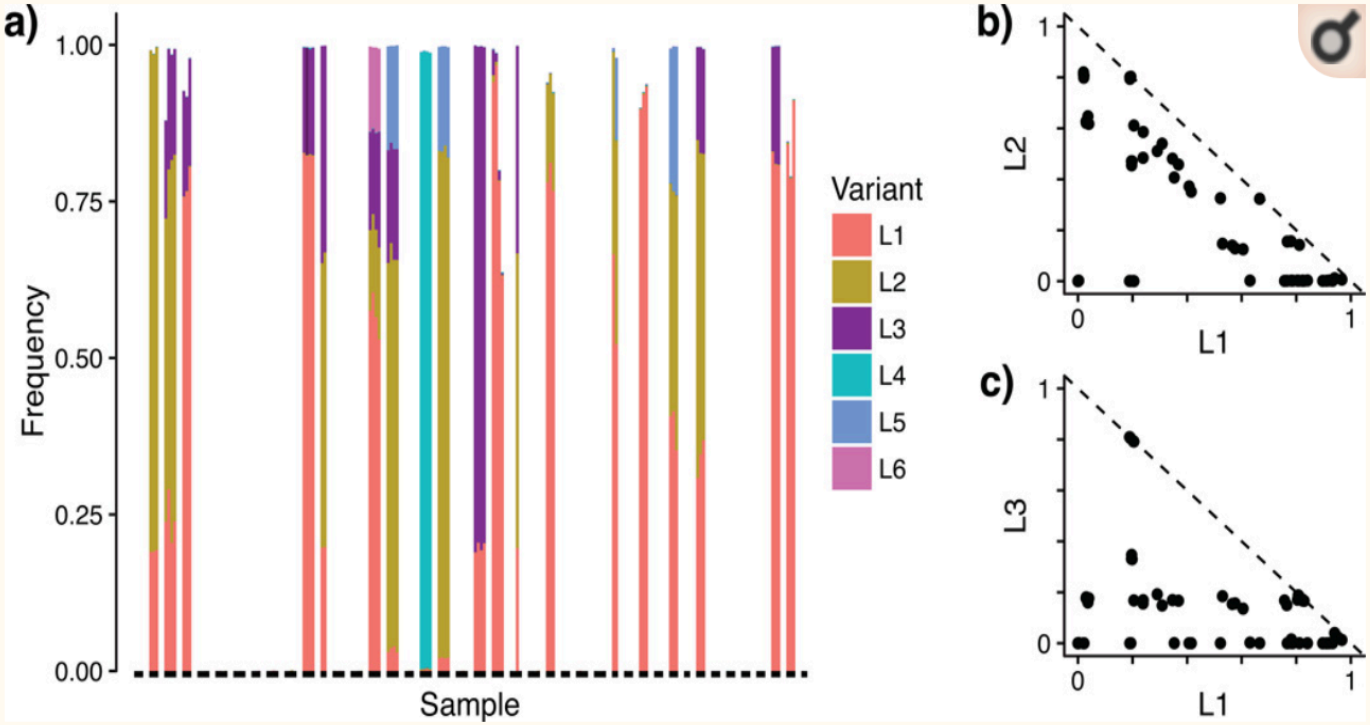


Figure 2

Lactobacillus crispatus sequence variants in the human vaginal community during pregnancy

DADA2 identified six *Lactobacillus crispatus* 16S rRNA sequence variants present in multiple samples and a significant fraction of all reads (L1: 19.7%, L2: 11.1%, L3: 6.5%, L4: 3.1%, L5: 1.3%, L6: 0.4%). (a) The frequency of L1–L6 in each sample. Black bars at the bottom link samples from the same subject. The frequency of (b) L1 vs. L2, and (c) L1 vs. L3, by sample. The dashed line indicates a total frequency of 1.

In several mock communities DADA2 identified more real variants and output fewer spurious sequences than other methods.

Vaginal samples from a cohort of pregnant women, revealing a diversity of previously undetected *Lactobacillus crispatus* variants.

General Pipeline

- Filter and trim: filterAndTrim()
- Dereplicate: derepFastq()
- Learn error rates: learnErrors()
- Infer sample composition: dada()
- Merge paired reads: mergePairs()
- Make sequence table: makeSequenceTable()
- Remove chimeras: removeBimeraDenovo()

raw data

data
processing
/ pipeline

results & data visualization

Let's get coding!