Analysing 16S rRNA sequencing data

Short intro with practical

Vietnam School of Biology VSBO4

Dr. Hao Chung The

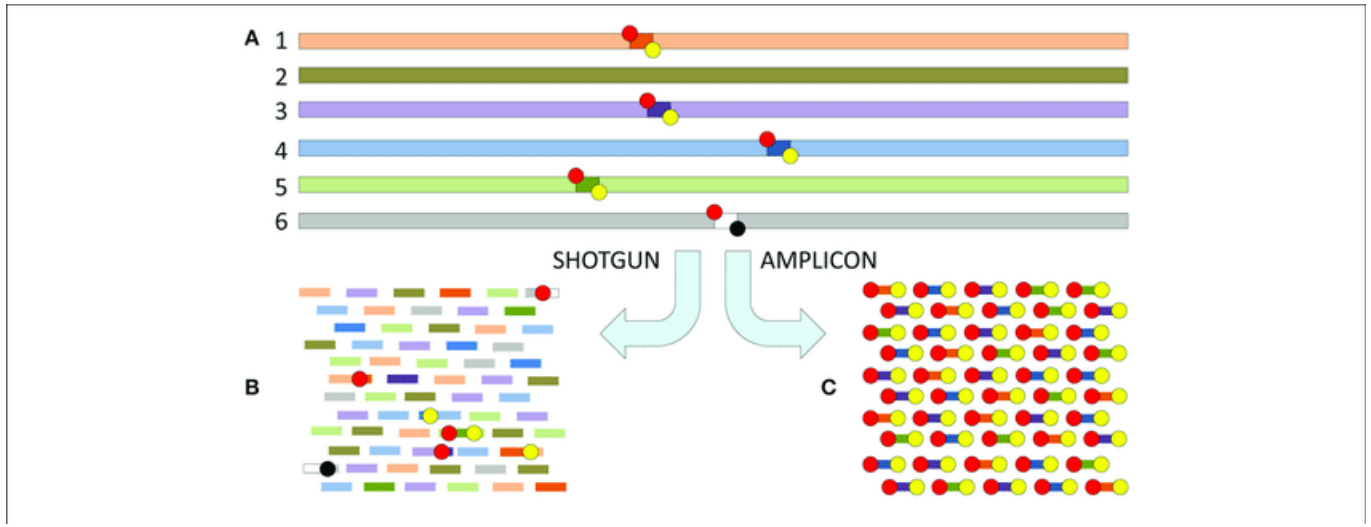
ICISE, Quy Nhon, Vietnam

5th September 2025

Outline

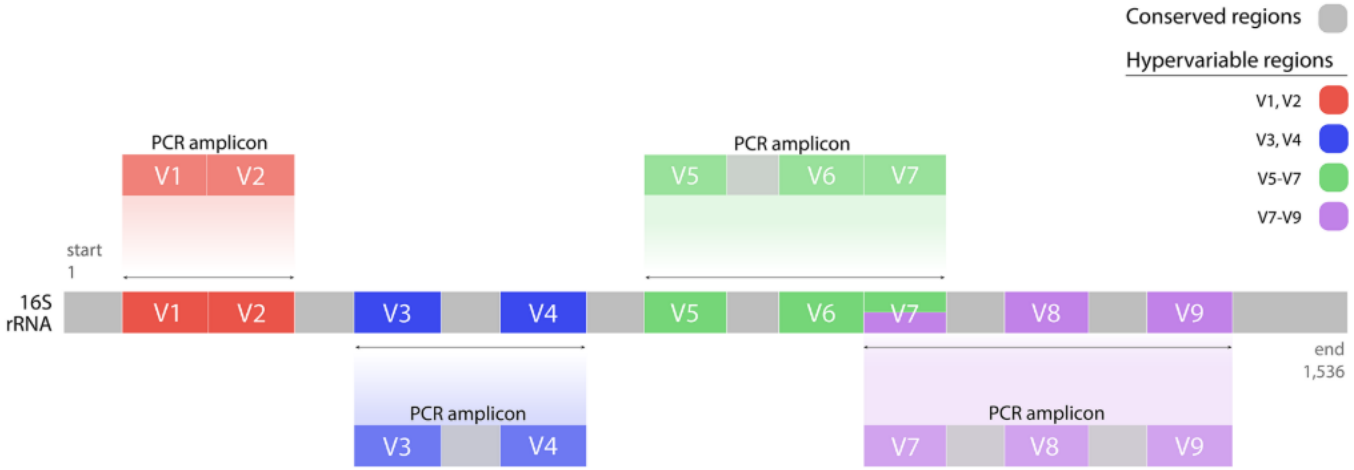
- 16S rRNA sequencing library preparation
- Note on wet-lab: low biomass samples
- Practical example: microbiome of tumour samples
- Rough look of sequencing data and QC check
- DADA2
- Phyloseq
- Exploratory analyses: alpha-div, beta-div, note on compositionality and sparsity
- Differential abundance analysis

16S rRNA sequencing



Why is 16S rRNA sequencing still relevant?

- Samples with high level of host DNA or and/or low bacterial biomass
- Generally cheaper
- Less computational demanding
- Broad taxonomic information is concerned
- Paired with inter-kingdom analysis (ITS sequencing, virome RNA sequencing).



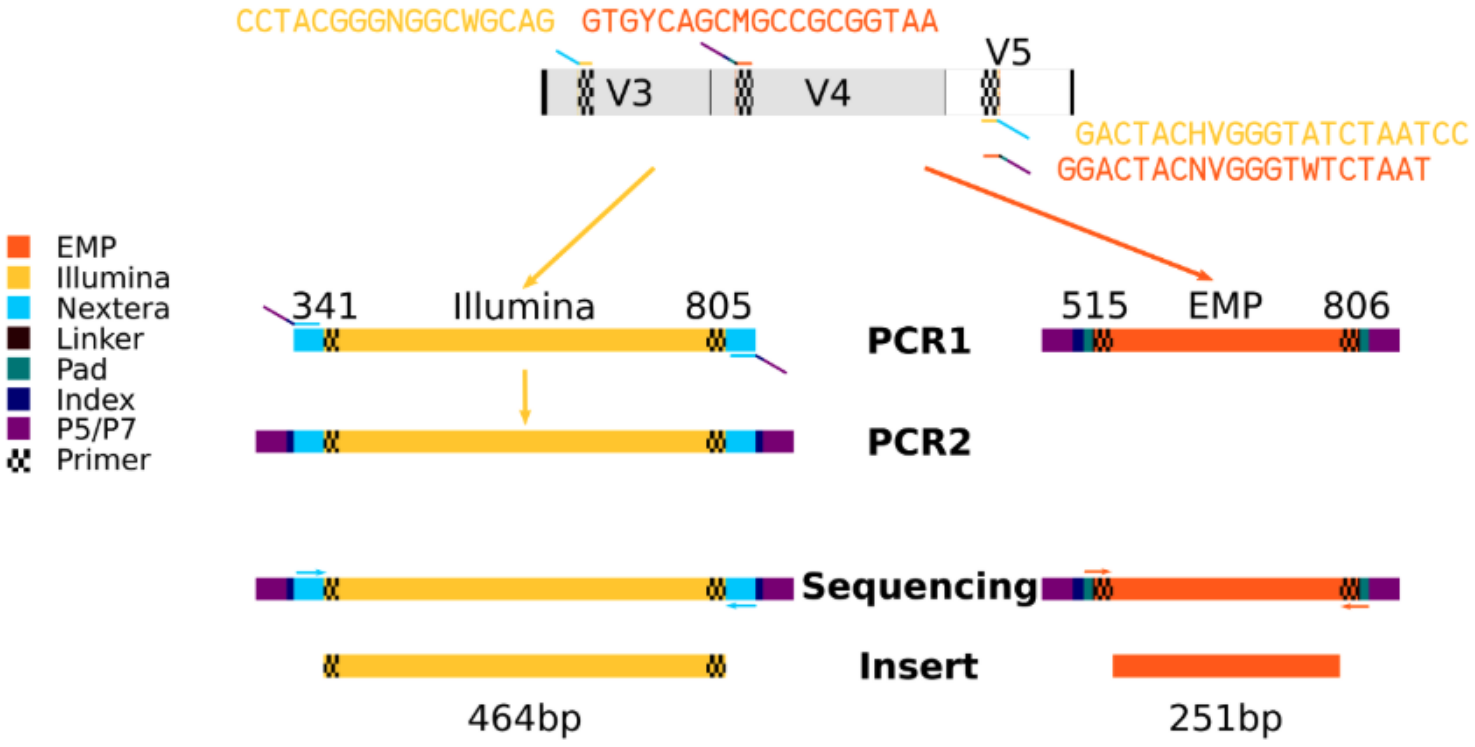
Factors affecting the choice of primers

- Nature of the sample's microbiome
- Harmonisation with previous studies
- Choice of sequencing platform

Lopez-Aladid et al., Sci Rep, 2023

Notes on wet lab

V3-V4 and V4 regions are the most commonly sequenced 16S rRNA gene region

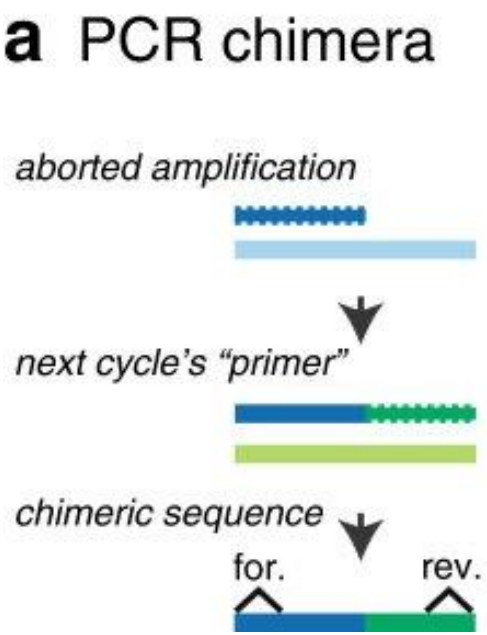


Credit: Oregon State University CQLS 16S sequencing protocol https://docs.hpc.oregonstate.edu/cqls/tips/16s-sequencing/

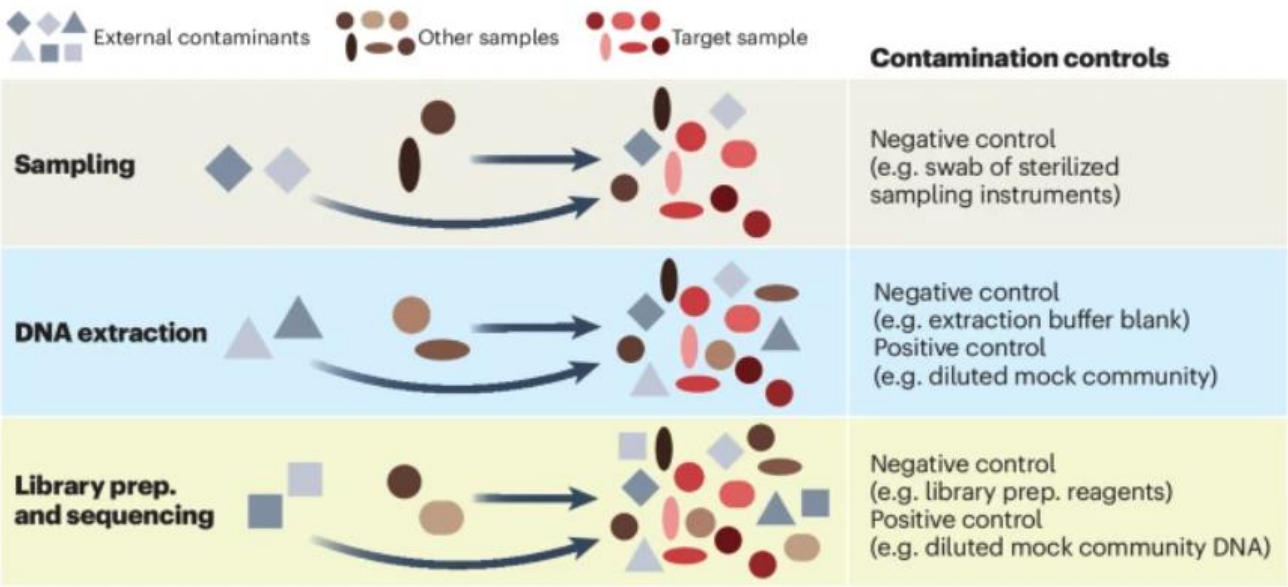
V3-V4: PE 300 bp
V4: PE 150bp – 250bp

Gohl et al., Nature Biotechnology, 2016

- Use of high-fidelity polymerases (i.e. KAPA)
- Optimise PCR conditions for primary amplification (15 – 25 cycles) to reduce chimeric amplifications
- Downstream cleaning, normalisation and pooling need close attention



Low biomass microbiome



Fierer et al., Nat Microbiology, 2025

Stage of experiment:	Recommended practice:
Experimental design	• Randomize sample types and treatment groups to prevent influences of batch/day-level variation of contaminant DNA or other confounding variables • Record batch numbers for reagents
Sampling	• Wear cleansuit; clean gloves; face mask • Use sampling blank controls
DNA extraction	• Wear cleansuit; clean gloves; face mask • Decontaminate working area before use • Perform lab work in a controlled environment (cabinet), physically isolated from post-PCR facilities • Use DNA extraction blank controls
Library preparation	• Wear cleansuit; clean gloves; face mask • Decontaminate working area before use • Perform lab work in a controlled environment (cabinet) • Use no-template amplification controls
DNA sequencing	• Ensure use of unique redundant barcodes
Data analysis	• Report taxa found in negative controls • Comparison of taxa found in controls to those identified in biological samples • Use of subtractive filtering to remove contaminants from biological samples [68–71]

Opinion

Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations

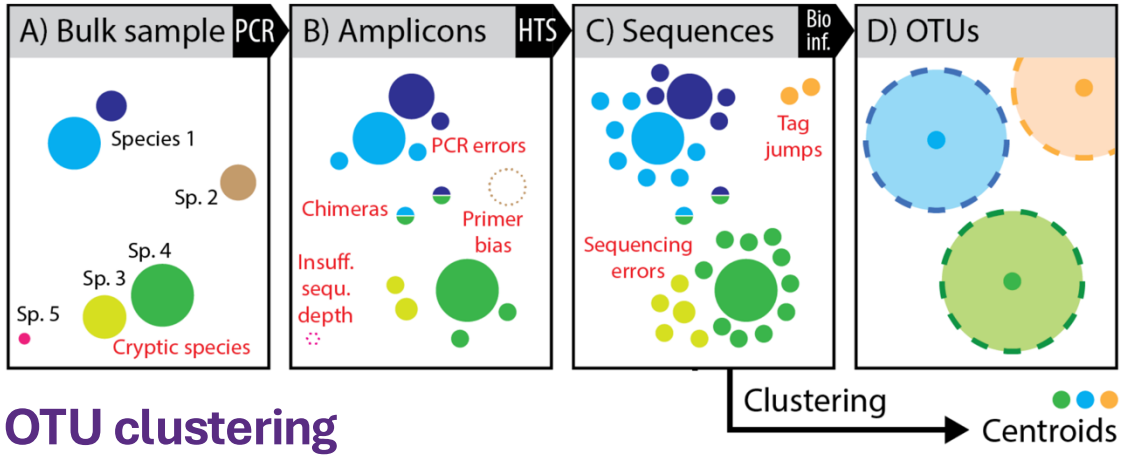
Eisenhofer et al., Trends in Microbiology, 2019

Common laboratory contamination genera *Ralstonia, Pseudomonas, Acinetobacter, Enhydrobacter*, etc.

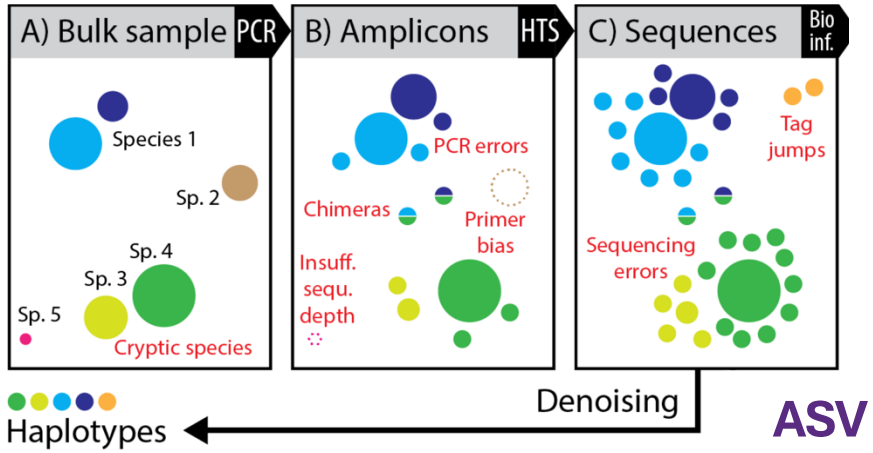
Understand what you would expect from your samples

What to expect from 16S sequencing?

Raw sequences → merge → classify → taxonomy → **sample x taxa table** → downstream analyses



OTU clustering



ASV approach

Credit: NCI Bioinformatic training
https://bioinformatics.ccr.cancer.gov/docs/qiime2/Lesson3/

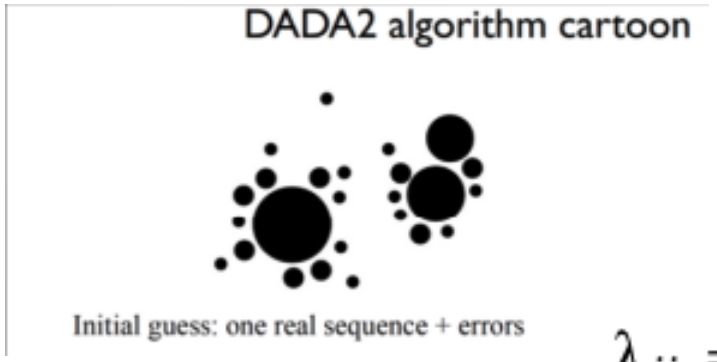
Feature	OTU (Operational Taxonomic Unit)	ASV (Amplicon Sequence Variant)
Definition	Sequences with similarity (commonly ≥97%) grouped into an OTU	Infers exact biological sequences by denoising and error correction
Resolution	May group several closely related species together	Higher chance for species classification
Reproducibility	Poor – OTU clustering is study specific	High – ASV represents consistent sequences reproducible

Revised from ChatGPT output

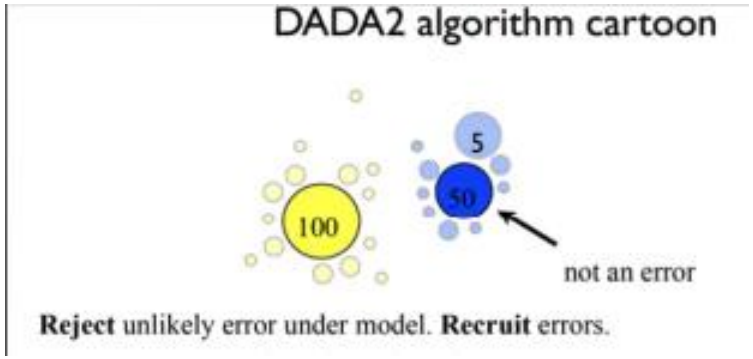
Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

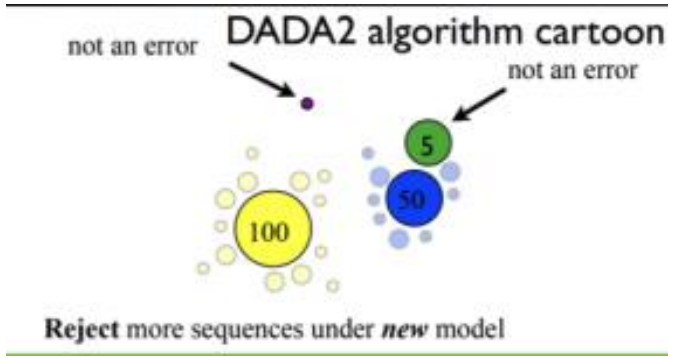
Benjamin J Callahan[1], Paul J McMurdie[2] and Susan P Holmes[3]

ASV approach



$$\lambda_{ji} = \prod_{l=0}^{L} p(j(l) \rightarrow i(l), q_i(l))$$





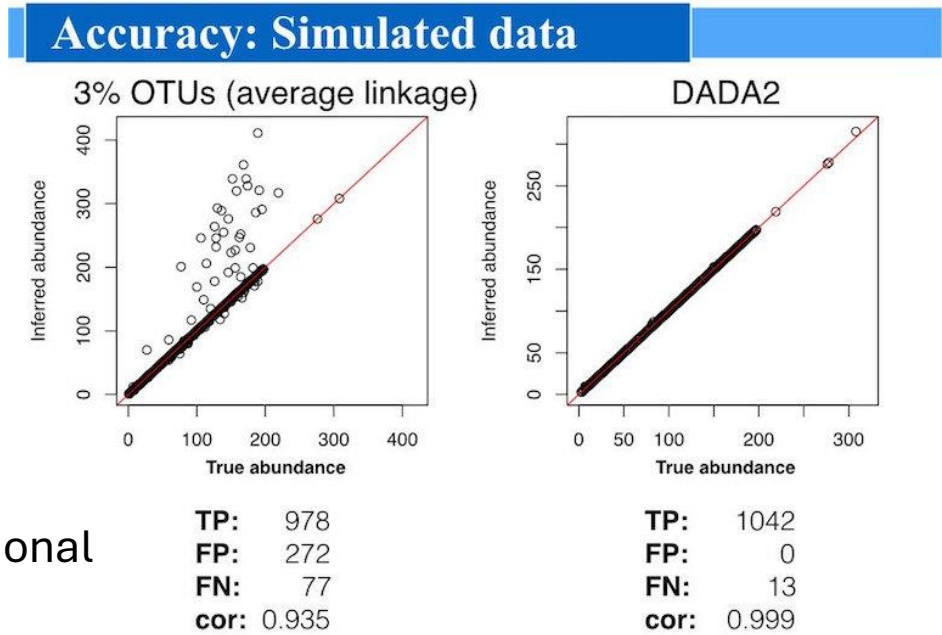
DADA2 algorithm includes three main steps:

- Divisive partitioning by sequence comparison
- Error model construction
- Alternating consistency until consistency

DADA2: High-resolution sample inference from Illumina amplicon data

Callahan et al., Brief Comms, 2016 - >30,000 citations

DADA2 recovers additional sequence variation compared to UPARSE



Data: Kopylova, et al. mSystems, 2016.

Data for tutorial

Rationale and study design

- Colorectal cancer microbiome is well-studied in Western developed settings, but under-explored in developing settings
- Enrolled 43 CRC patients (cases) and 25 patients with colorectal polyps (controls) in Ho Chi Minh City
- Collected saliva, tissues from gut (excised tumour and non-tumour sites for cases; polyp and non-polyp biopsy from controls)

ARTICLE OPEN

Tran et al., npj Biofilms and Microbiomes, 2022

Tumour microbiomes and *Fusobacterium* genomics in Vietnamese colorectal cancer patients

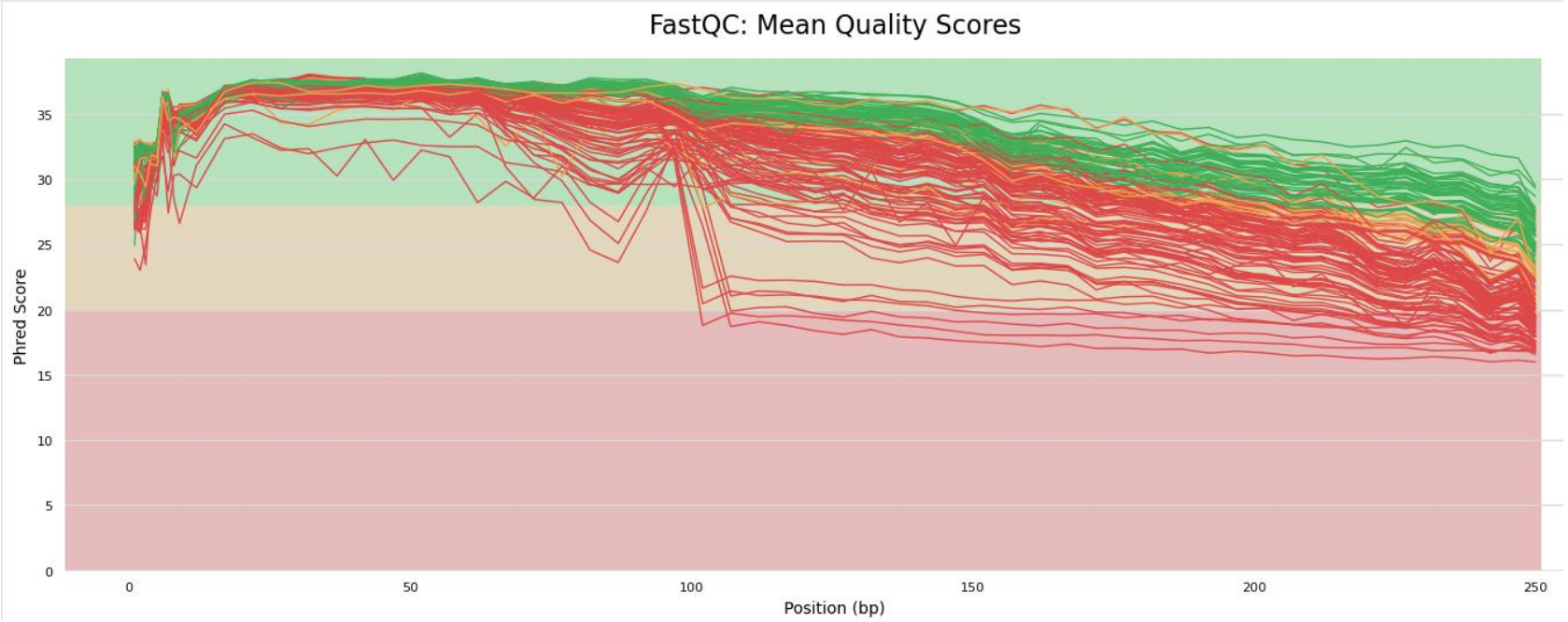
Hoang N. H. Tran[1,11], Trang Nguyen Hoang Thu[1,11], Phu Huu Nguyen[2], Chi Nguyen Vo[2,3], Khanh Van Doan[4], Chau Nguyen Ngoc Minh[1], Ngoc Tuan Nguyen[2], Van Ngoc Duc Ta[2], Khuong An Vu[2], Thanh Danh Hua[2], To Nguyen Thi Nguyen[1], Tan Trinh Van[1], Trung Pham Duc[1], Ba Lap Duong[2], Phuc Minh Nguyen[2], Vinh Chuc Hoang[2], Duy Thanh Pham[1,5], Guy E. Thwaites[1,5], Lindsay J. Hall[6,7,8], Daniel J. Slade[9], Stephen Baker[10], Vinh Hung Tran[2] and Hao Chung The[1]✉

Table 1. Baseline characteristics of patients recruited in this study.

	CRC cases ($n = 42$)	Controls ($n = 21$)	*p*-value
Age	64 [54–69]	60 [53–66]	0.359
Male sex	62%	76%	0.395
BMI	22.9 [20.85–24.95]	22.2 [21.1–23.4]	0.387
Overweight/obesity[a]	47.60%	33%	0.409
Diabetes	19%	19%	1
High blood pressure	52%	47.60%	0.79
Active smoking in the last two years	21.40%	19%	1
Oral diseases[b]	33%	38%	0.782
Family history of cancer	19%	19%	1
Location of sampled mucosa			0.533
Descending colon	7	3	
Sigmoid colon	28	12	
Rectum	7	6	
Size of tumour/polyp (cm)	5 [4–5.75]	1 [0.7–1.2]	
TNM stage of cancer	II (18), III (20), IV (4)		
Polyp dysplasia grade		Low (4), none (17)	

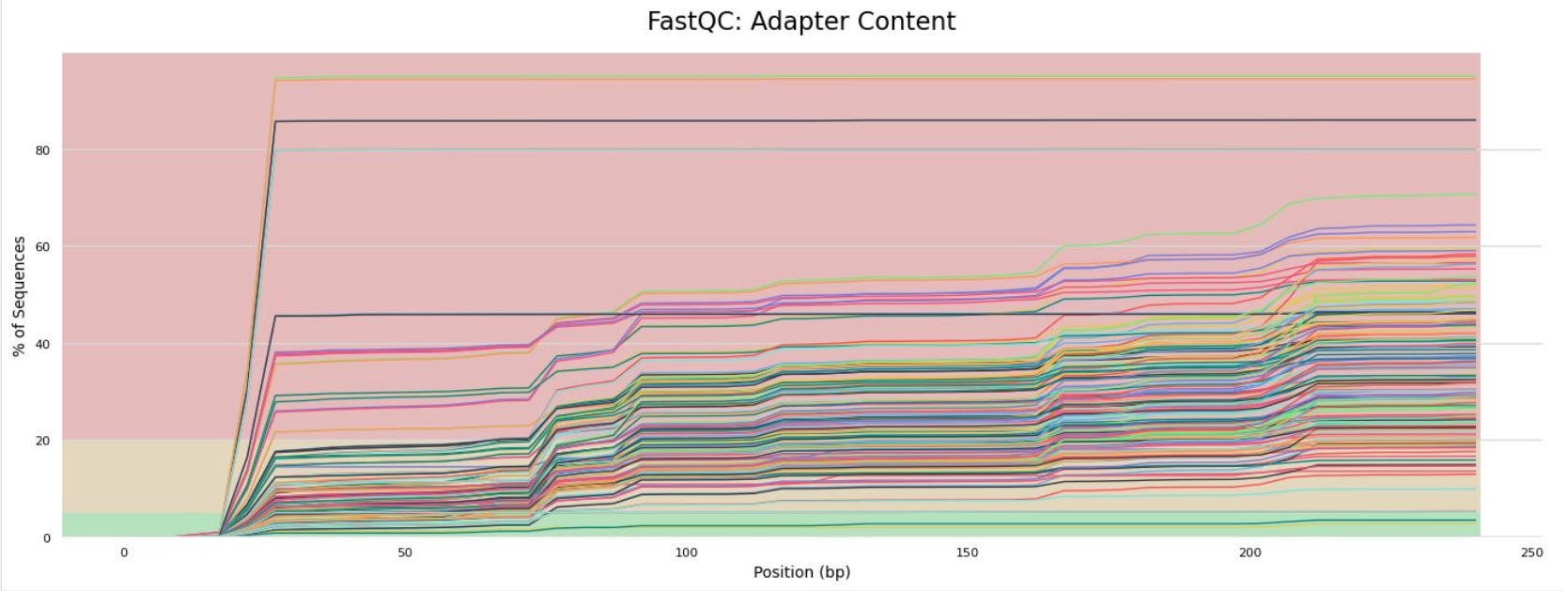
Data from 43 tumour and 25 biopsy microbiomes are included in this practical

Poor sequencing quality of 68 gut tissue microbiomes



16S rRNA sequencing:

- V4 region (315F – 806R)
- 30 cycles of primary PCR
- Include mock and negative controls
- Library sequenced on Illumina MiSeq (250bp PE)

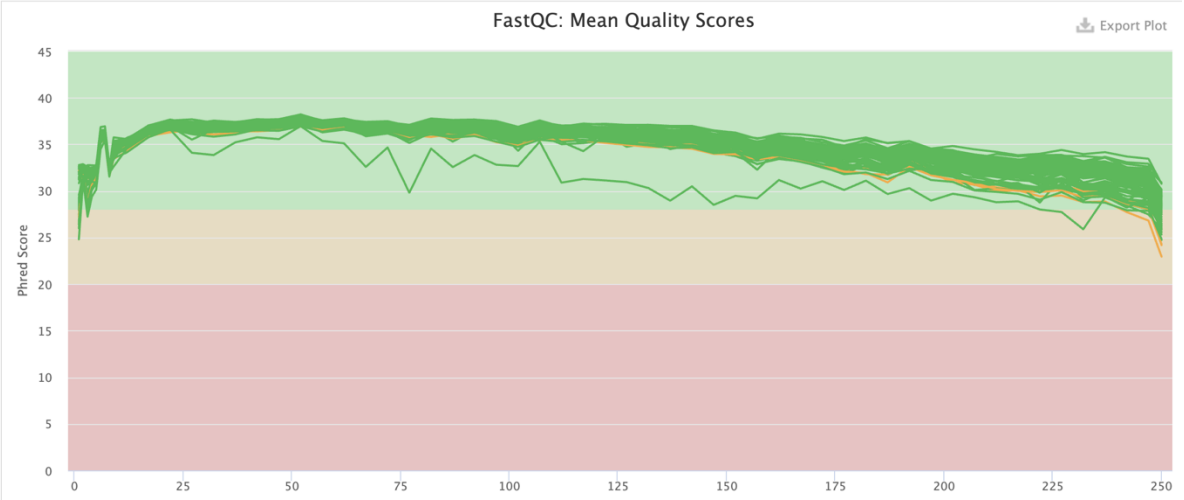


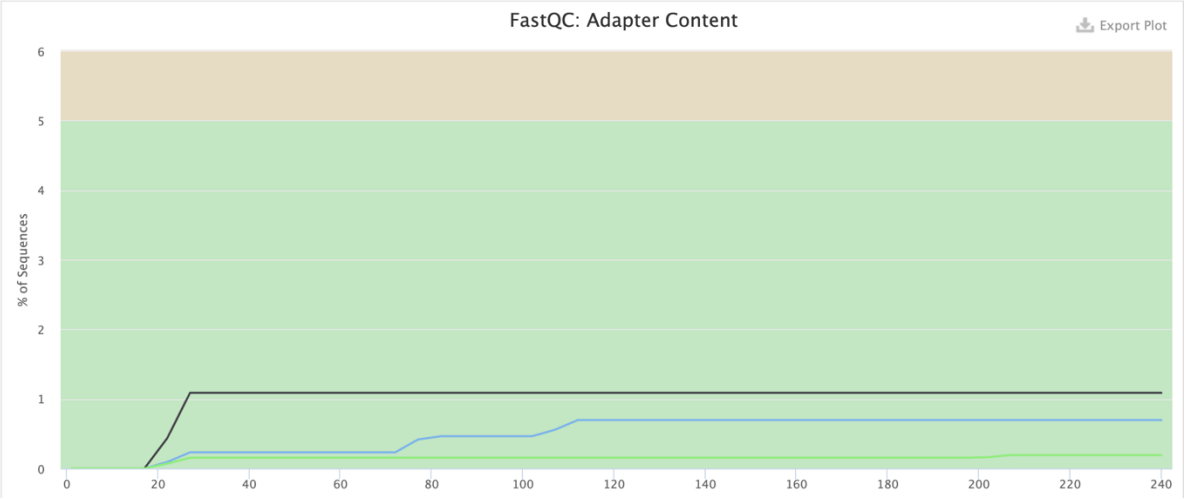
What is the likely cause of high adapter content?

To remove adapter sequences from raw fastq output

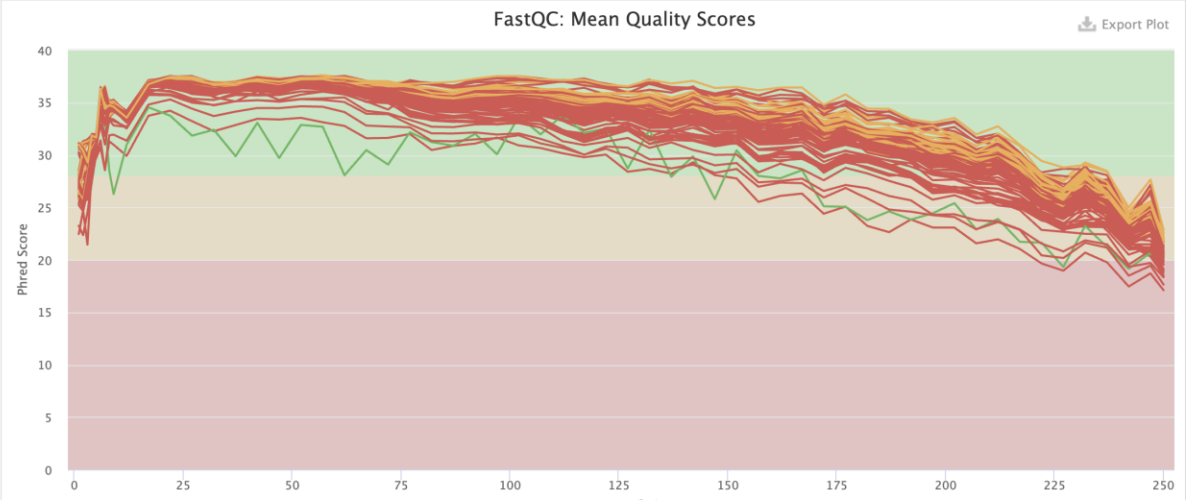
```
java -jar ~/local/Trimmomatic-0.36/trimmomatic-0.36.jar PE -phred33 ${x}_1.fastq.gz ${x}_2.fastq.gz ${x}_paired_1.fastq.gz
${x}_unpaired_1.fastq.gz ${x}_paired_2.fastq.gz ${x}_unpaired_2.fastq.gz ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:6:true
MINLEN:150
```

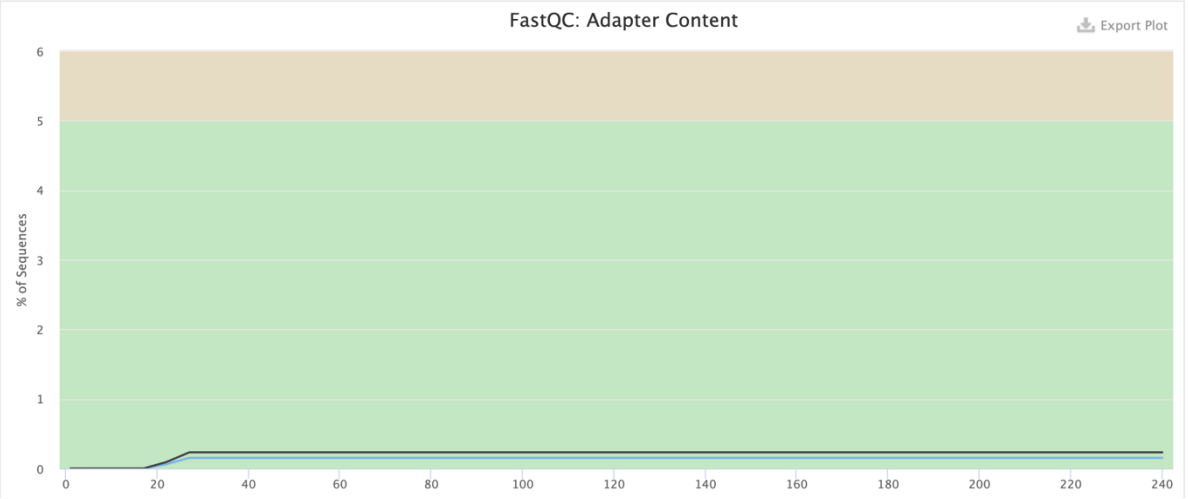
Forward reads



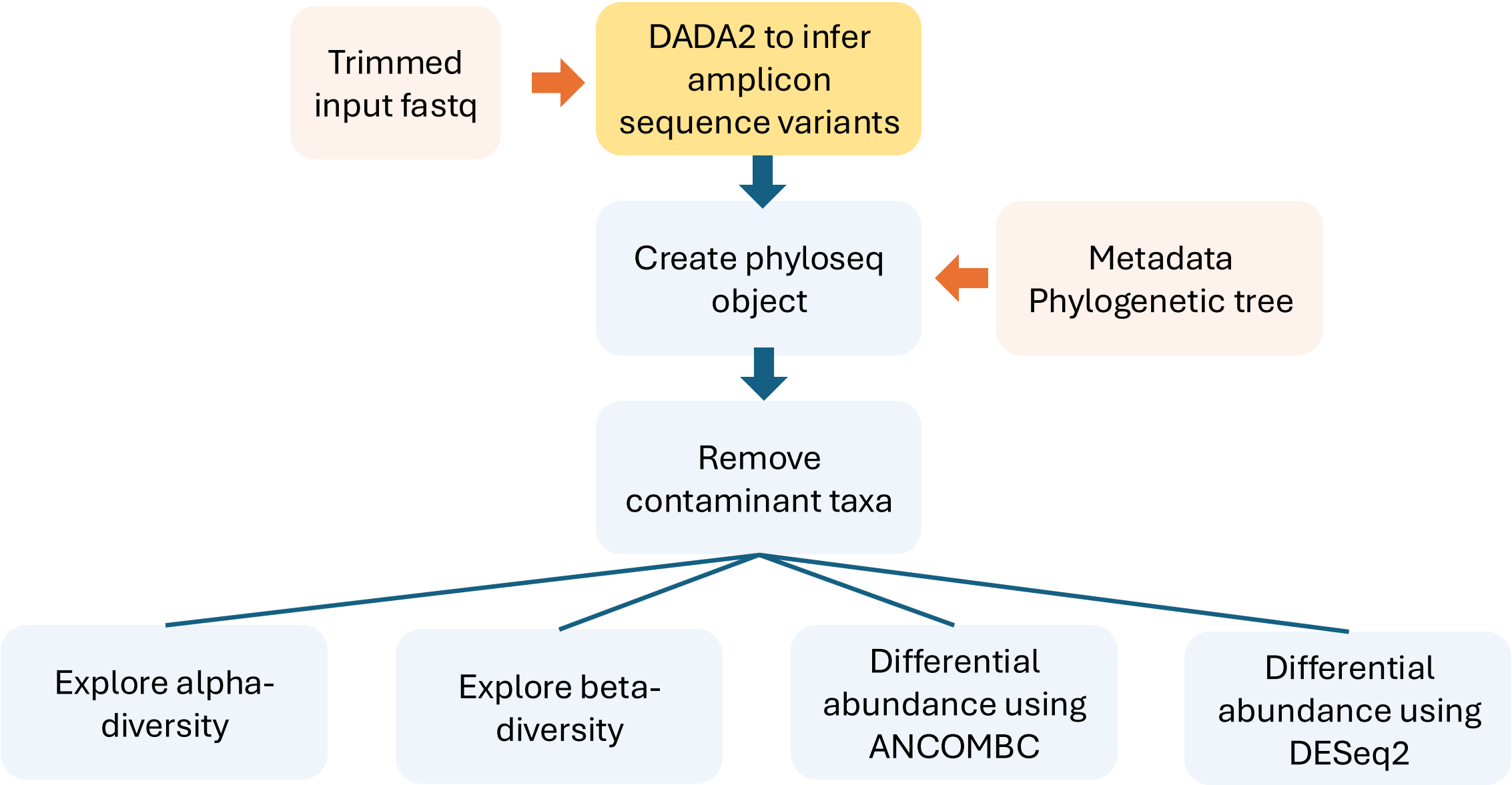


Reverse reads





Overview of tutorial data analysis



DADA2 pipeline in practice

Starting material: Paired-end fastq files

Steps:

- Filter and trim: remove low quality sections of reads, and primers
- Denoising: learn errors from own data and infer true sequences
- Merge: combine forward and reverse reads from each sample
- Chimera detection and removal

DADA2 Pipeline Tutorial (1.16)

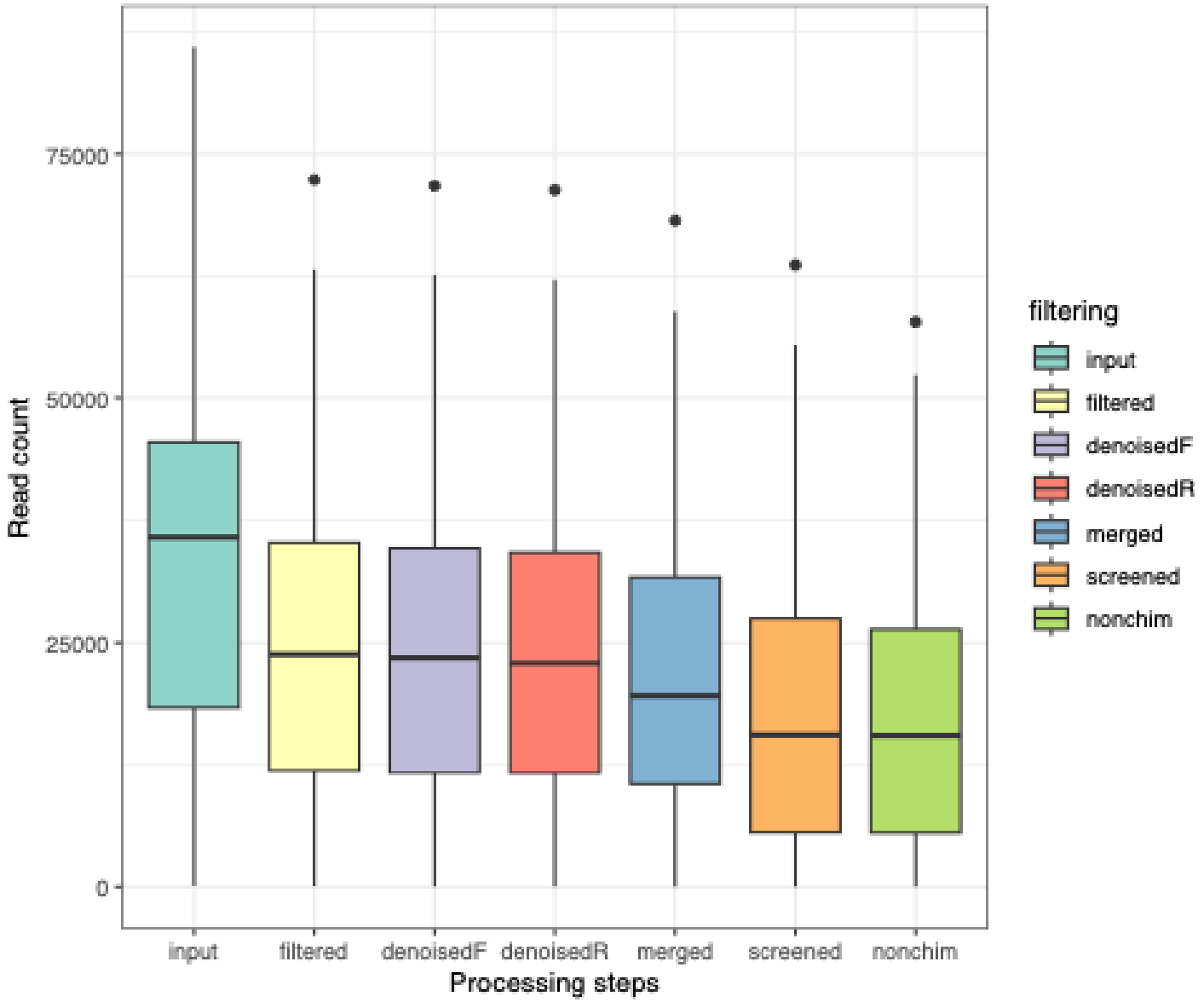
Here we walk through version 1.16 of the DADA2 pipeline on a small multi-sample dataset. Our starting point is a set of Illumina-sequenced paired-end fastq files that have been split (or "demultiplexed") by sample and from which the barcodes/adapters have already been removed. The end product is an **amplicon sequence variant (ASV) table**, a higher-resolution analogue of the traditional OTU table, which records the number of times each exact amplicon sequence variant was observed in each sample. We also assign taxonomy to the output sequences, and demonstrate how the data can be imported into the popular phyloseq R package for the analysis of microbiome data.

https://benjjneb.github.io/dada2/tutorial.html

Practical starts now

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

After DADA2, what your output should be



Following each filtering step, the library size is getting shrunk

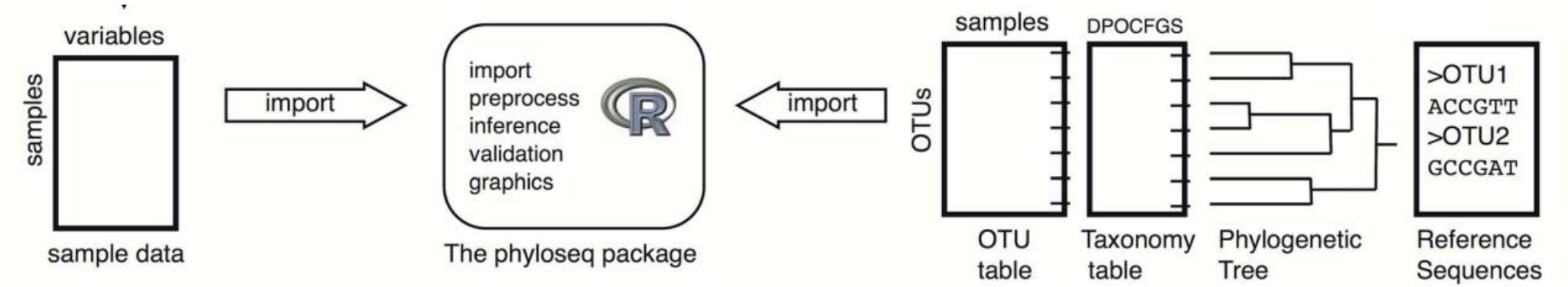
Filtered sequences are:

- Low sequencing quality
- Forward and reverse reads do not have sufficient overlapping
- Non-bacteria taxa
- PCR chimera

Phyloseq integrates microbiome related data

What should be input:

- OTU table: directly from ASV table after removal of nonbacterial taxa
- Taxonomy table: output from ‘assignTaxonomy’ function
- Phylogenetic tree: Read in tree built separately
- Sample metadata: csv file with rownames() matching sample_names()

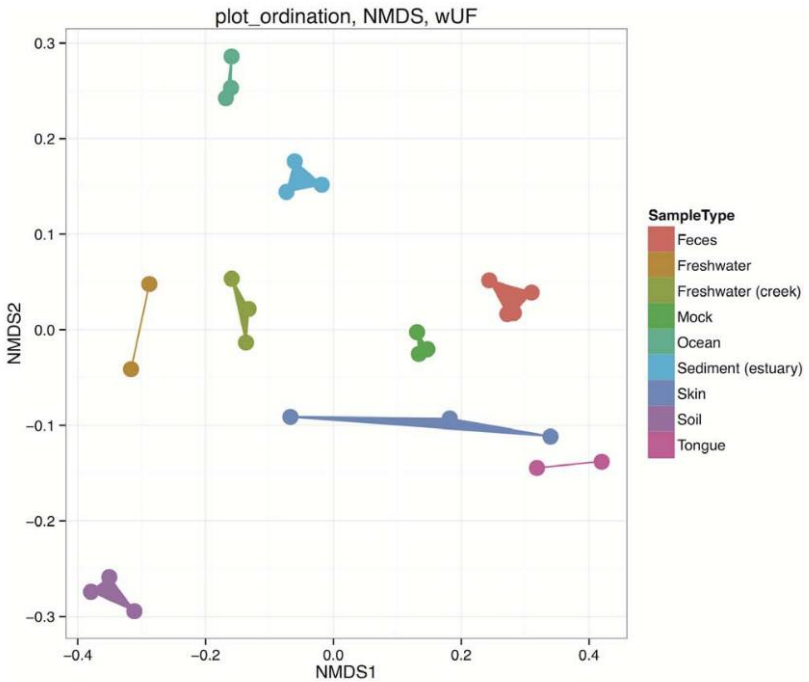


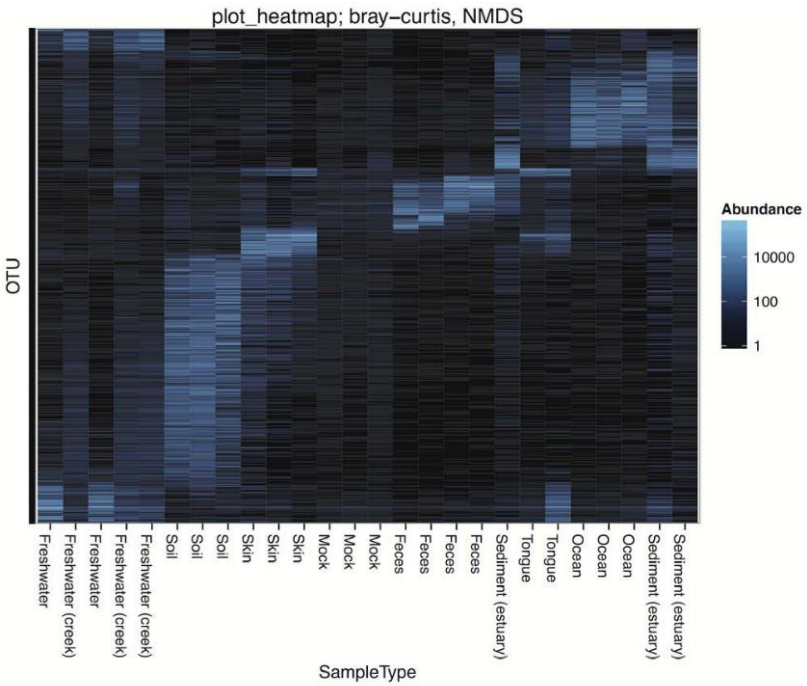
phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data

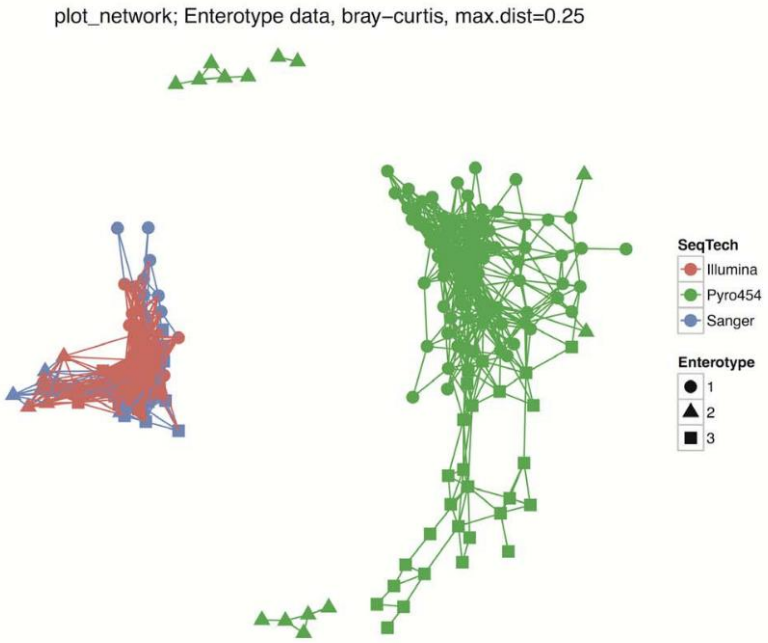
Paul J. McMurdie, Susan Holmes*

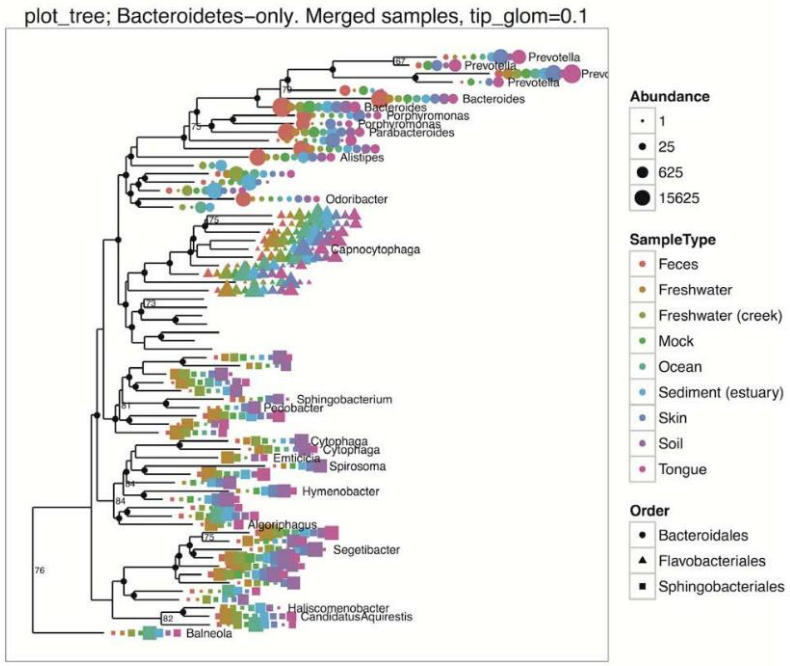
McMurdie and Holmes, 2013, PLoS One – >19,900 citations

Built-in plotting functions

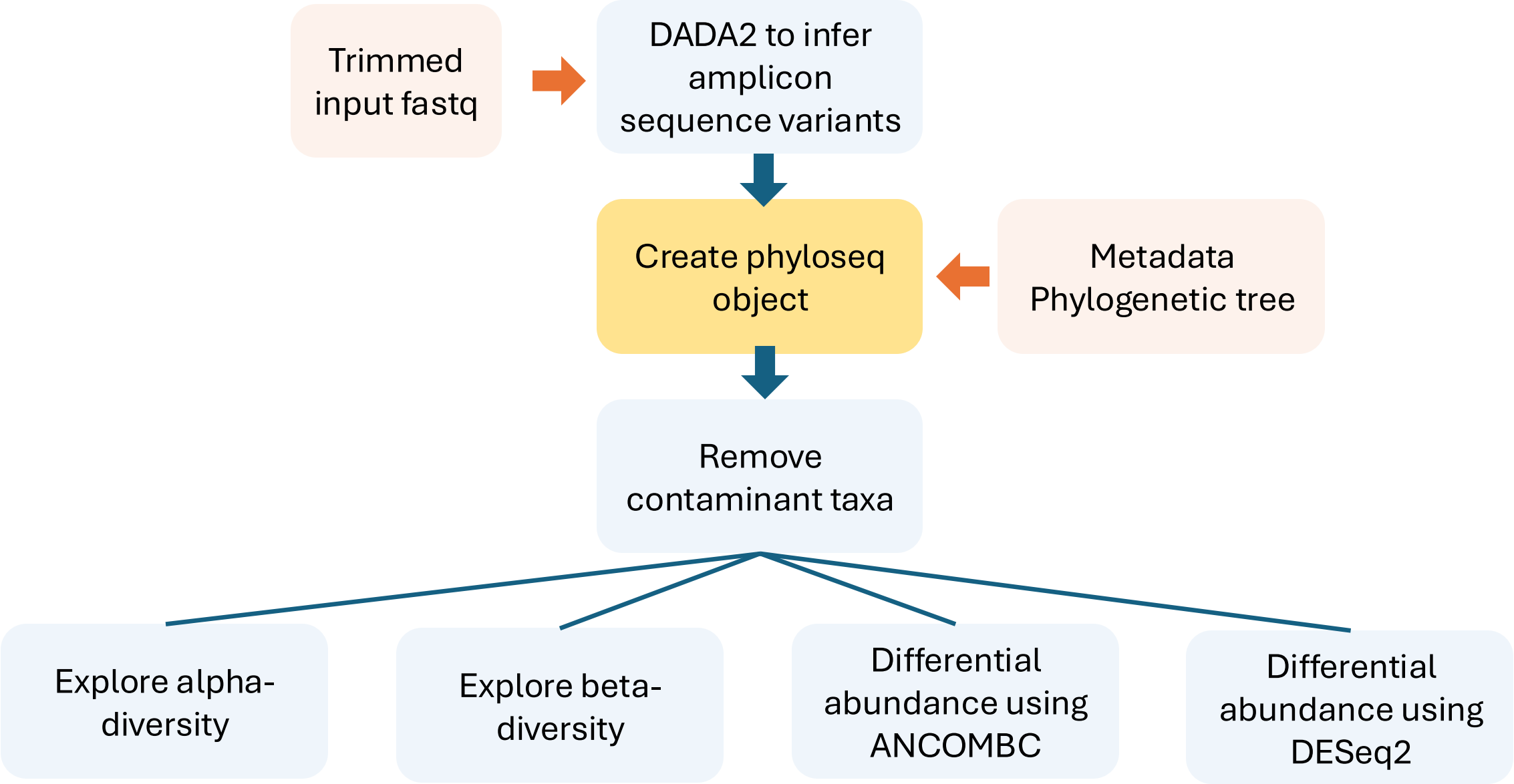








Overview of tutorial data analysis



After phyloseq, your output should be

otu_table()

```
OTU Table:          [5 taxa and 5 samples]
                     taxa are columns
    seq00001 seq00002 seq00003 seq00004 seq00005
B1       811       16        0       155       174
B10    18108        0        0         6        11
B11      658        0        0        12       558
B12     1978       14        0      8480       117
B13      688       13        0       698        13
```

sample_data()

	A	B	C	D	E	F	G
1	sample_ID	Group	sample_type	Patient	enrol_year	Age	sex
2	B1	control	biopsy	27EN_0001	NA	NA	NA
3	B10	control	biopsy	27EN_0010	2019	73	F
4	B11	control	biopsy	27EN_0011	2019	75	M
5	B12	control	biopsy	27EN_0012	2019	35	M
6	B13	control	biopsy	27EN_0013	2019	67	M

phy_tree()

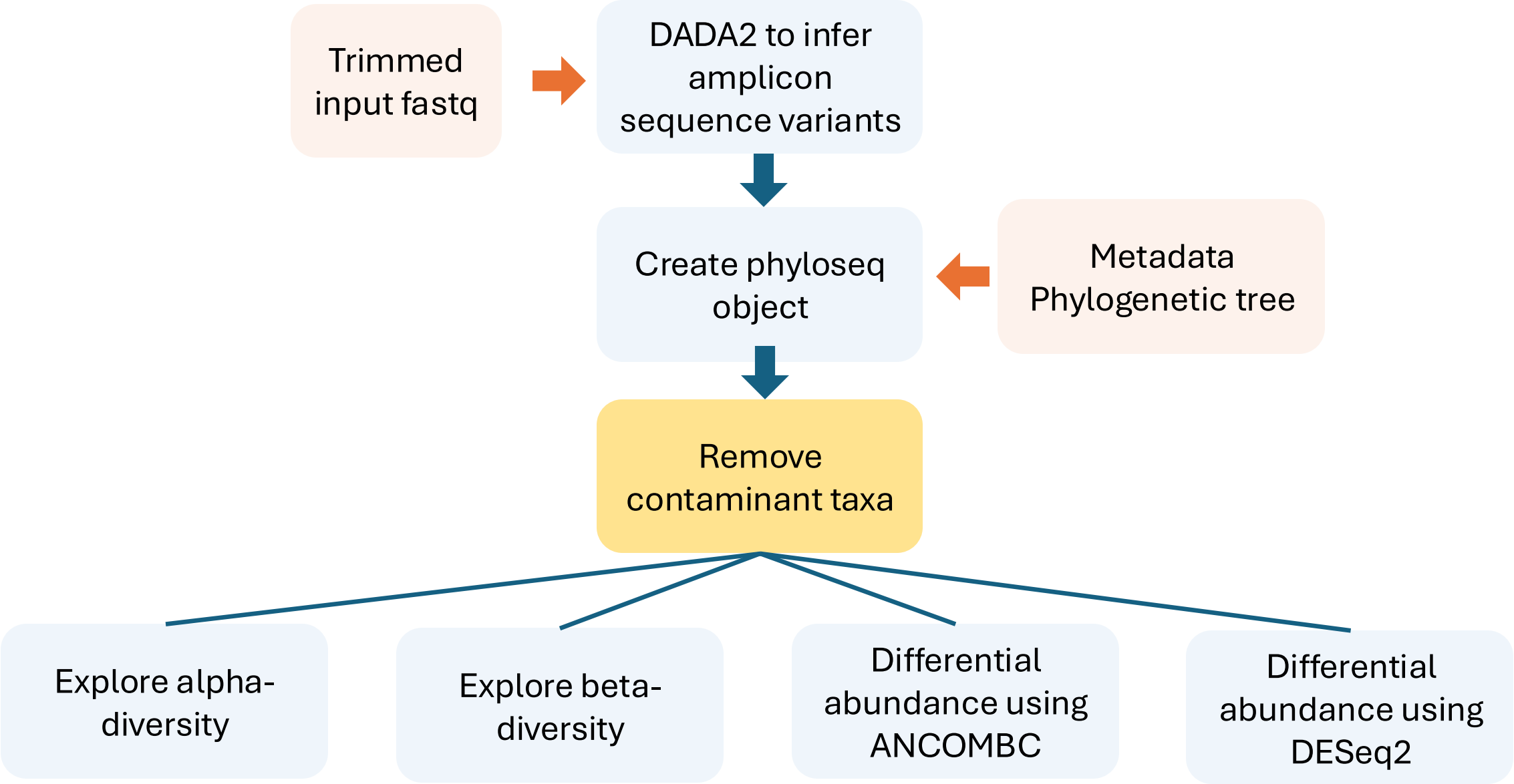
?

tax_table()

ID	Kingdom	Phylum	Class	Order	Family	Genus	Species
seq00001	Bacteria	Pseudomonadota	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Escherichia-Shigella	coli
seq00002	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	fragilis
seq00003	Bacteria	Fusobacteriota	Fusobacteriia	Fusobacteriales	Leptotrichiaceae	Leptotrichia	NA
seq00004	Bacteria	Fusobacteriota	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium	mortiferum
seq00005	Bacteria	Actinomycetota	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella	aerofaciens
seq00006	Bacteria	Bacillota	Negativicutes	Veillonellales-Selenon	Selenomonadaceae	Megamonas	NA
seq00007	Bacteria	Pseudomonadota	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Klebsiella	pneumoniae
seq00008	Bacteria	Bacillota	Bacilli	Staphylococcales	Gemellaceae	Gemella	NA
seq00009	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	vulgatus

```
> ps
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 1856 taxa and 75 samples ]
sample_data() Sample Data:       [ 75 samples by 37 sample variables ]
tax_table()   Taxonomy Table:    [ 1856 taxa by 7 taxonomic ranks ]
```

Overview of tutorial data analysis



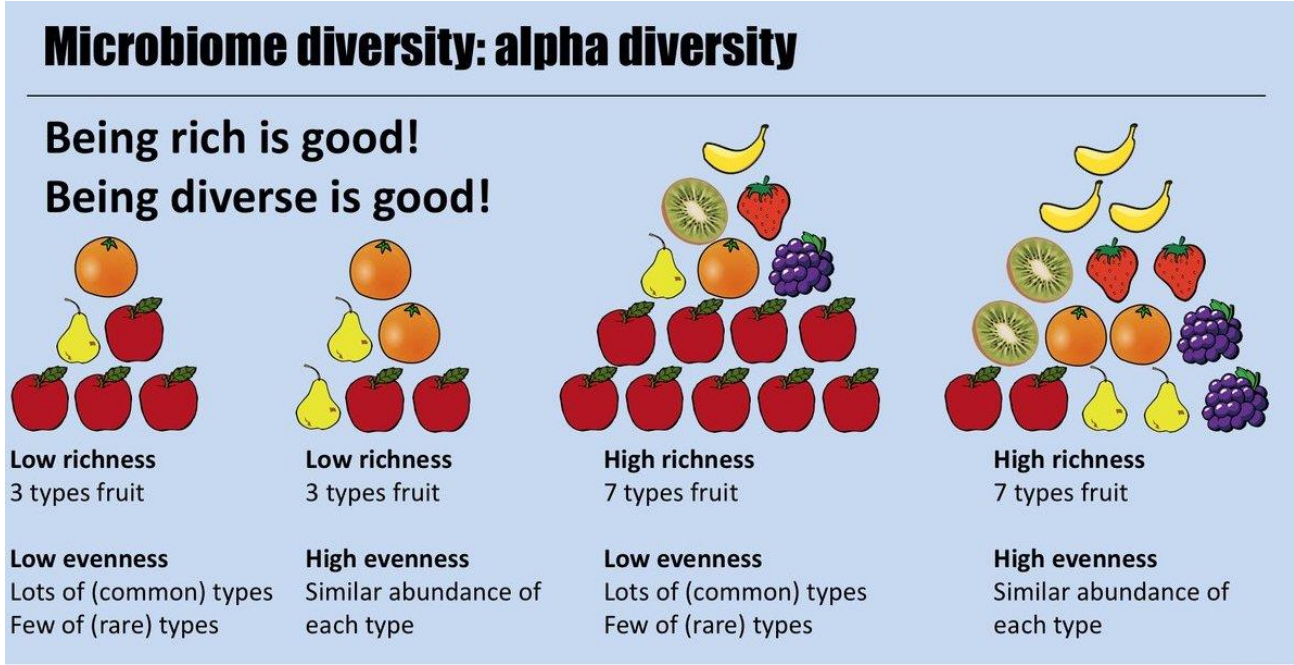
Diversity indices

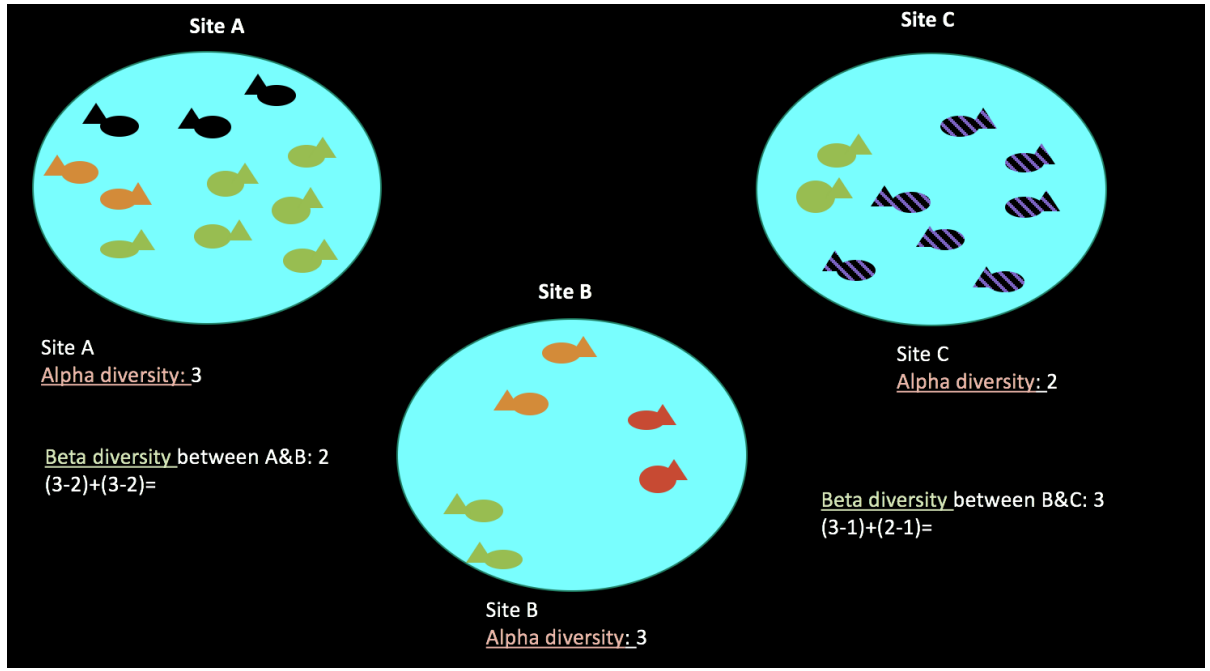
Alpha-diversity: for within-sample calculation

- Richness: number of taxa present in the sample
- Shannon index: combining richness with evenness

Beta-diversity: between-sample calculation

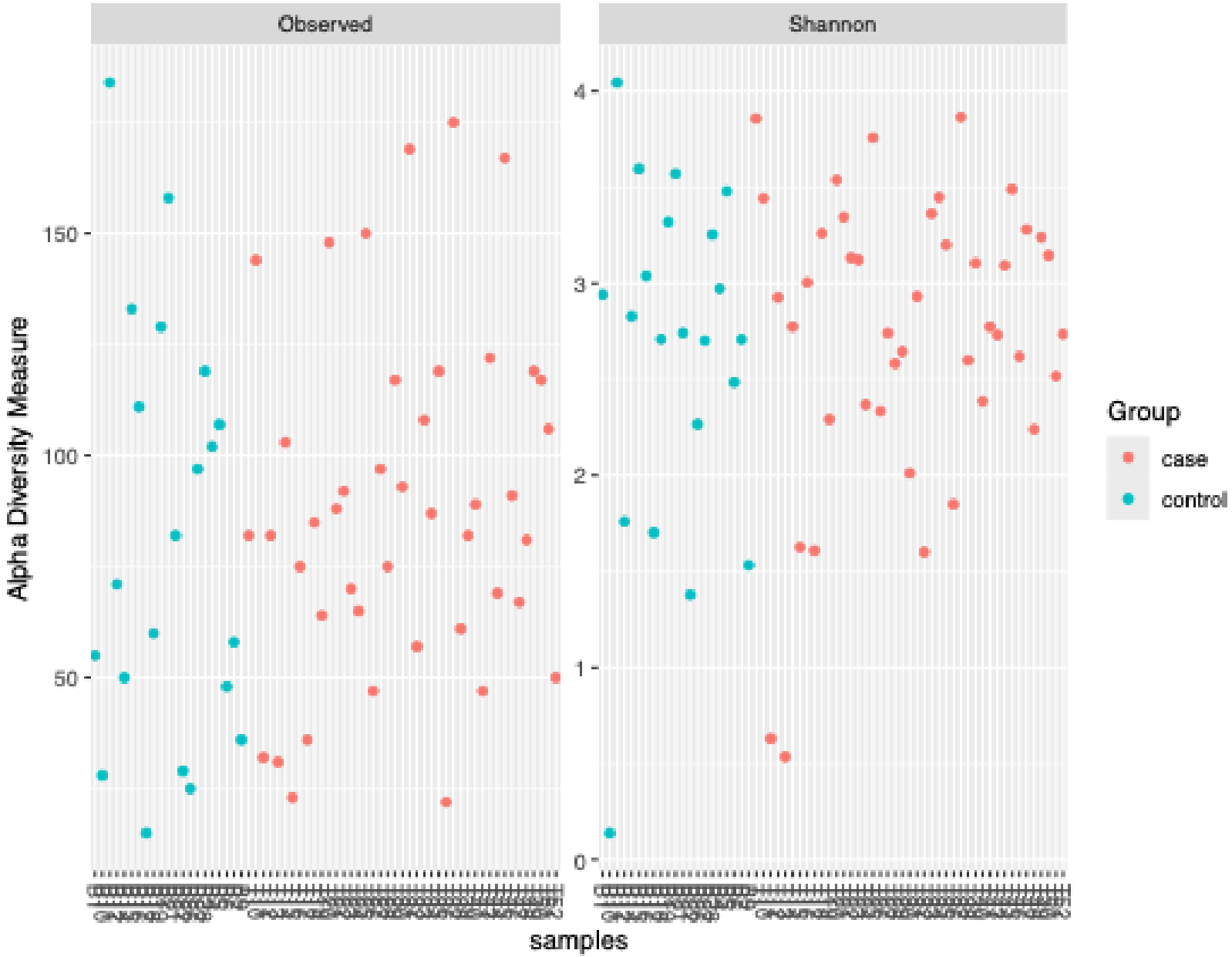
- Phylogenetic unaware: Bray-Curtis
- Phylogenetic aware: Unifrac, Weighted Unifrac, phylogenetic ILR (PhILR)





Credit: VMI Bootcamp
https://awbrooks19.github.io/vmi_microbiome_bootcamp/rst/4_concepts_of_community_analysis.html

Inspecting alpha-diversity

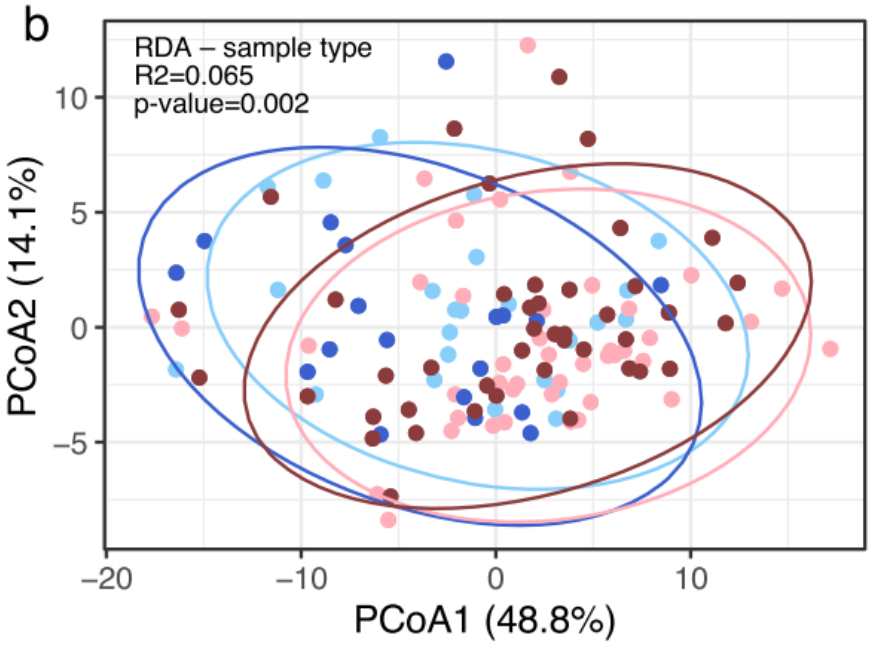


Beta-diversity

Bray-Curtis distance

Range [0 – 1]
No phylogenetic information

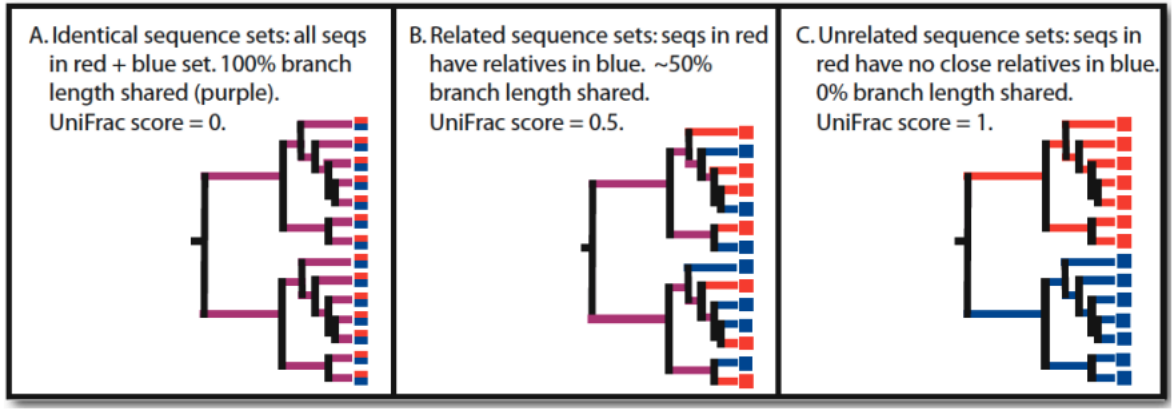
$$BC_d = \frac{\sum \left| x_i - x_j \right|}{\sum \left(x_i + x_j \right)}$$

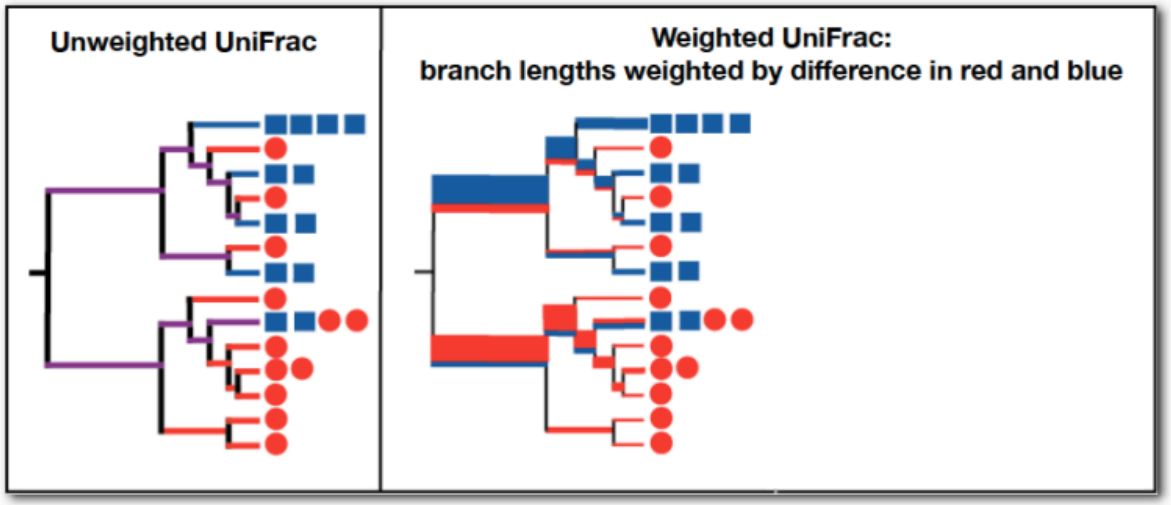


Usually presented as a principal component plot

Unifrac distance (unweighted & weighted)

Range [0 – 1]
Accounting for phylogenetic relatedness
Distance is fraction of the total branch length that is unique to any sample

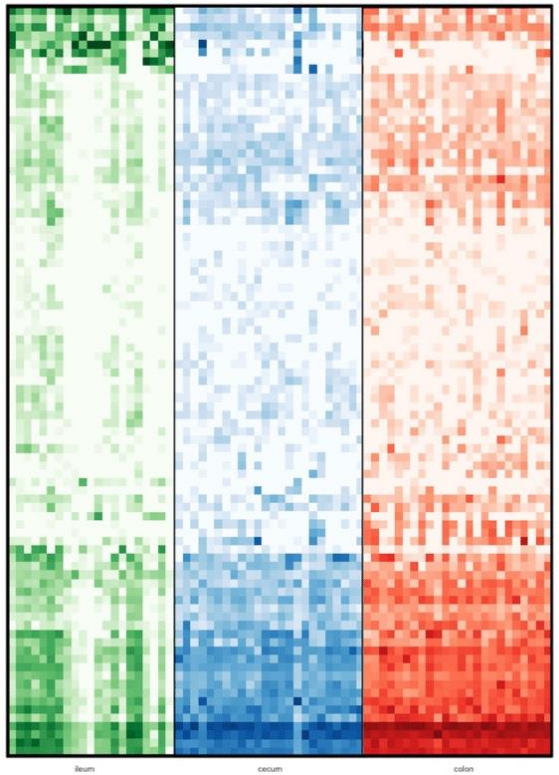


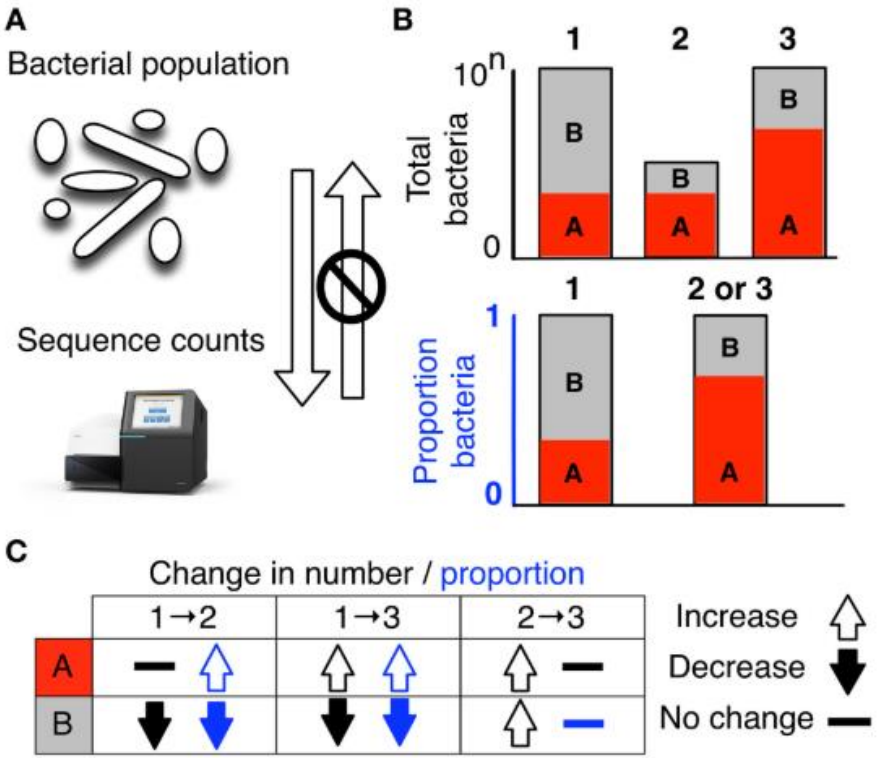


Microbiome data is compositional

Nature of 16S (microbiome) data:

- **Sparsity**: too many zero, usually >90% of the OTU table
- **Compositional**: sequenced reads represent proportion of the sampled microbiome, not absolute abundances
- **High-dimensional**: Many more taxa compared to samples





Gloor et al., Front Microb, 2017

Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	CLR ILR ALR
Distance	Bray-Curtis UniFrac Jenson-Shannon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perManova ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi ϕ ρ
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM

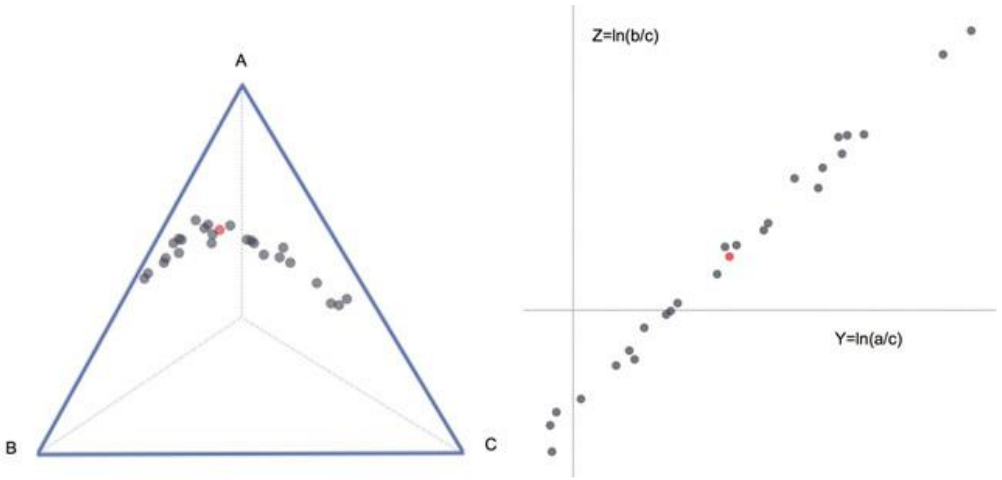
Different tools under the two frameworks

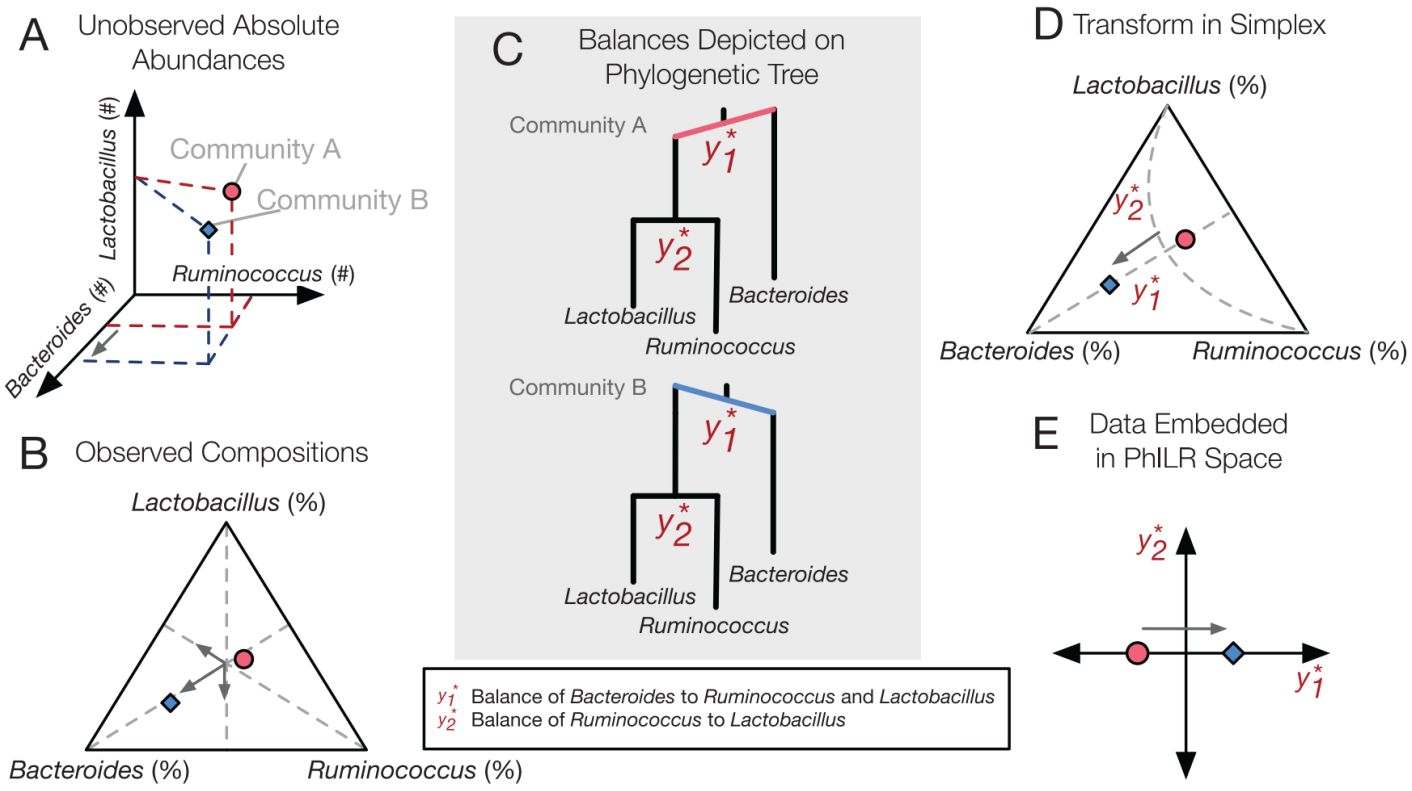
Compositional data analysis

Credit: https://qedinsight.com/2016/03/28/log-ratio-transformation/

- Compositional data are summed to a fixed number
- Common statistical methods are not fit to be applied to compositional data, thus giving spurious results

Transform data into Euclidean space, suitable for standard statistical inferences





Phylogenetic Isometric log-ratio (PhILR) transform

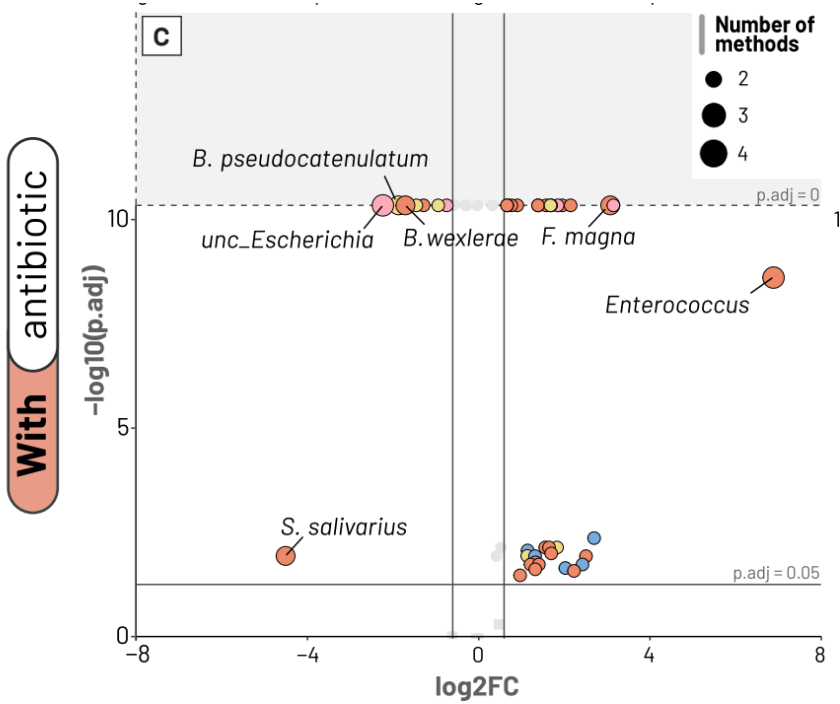
Silverman et al., Elife, 2017

Differential abundance analysis

- No best method. Each approach take assumptions and model/handle data differently
- Some methods can incorporate models to test for confounders
- Recommended to combine/intersect output from different methods for robustness

Feature	ANCOM-BC	DESeq2
Design	Compositional data	Originally for RNA-Seq
Bias correction	Correct for sampling bias	Count is abundance
False discovery rate	Better control	Prone to have inflated false positives
Model flexibility	Moderate	High with complex design
Sparsity handling	Moderately well	Option to include zero-inflated model (scRNA)

Revised from ChatGPT output



Le et al., npj AMR, 2024

ARTICLE

https://doi.org/10.1038/s41467-022-28034-z OPEN

Microbiome differential abundance methods produce different results across 38 datasets

Nearing et al., Nat Comms, 2022

Other tools to consider:

- ALDEx2 – conservative, compositional
- MaAsLin2 – complex study design

Q&A