Metabarcoding processing pipelines and considerations

Soledad Benitez Ponce (benitezponce.1@osu.edu)
March 14, 2024

Received: 10 February 2023 | Revised: 5 June 2023 | Accepted: 6 July 2023

DOI: 10.1111/1755-0998.13847

MOLECULAR ECOLOGY RESOURCES WILEY

FROM THE COVER

A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses

Ali Hakimzadeh[1] | Alejandro Abdala Asbun[2] | Davide Albanese[3] | Maria Bernard[4,5] | Dominik Buchner[6] | Benjamin Callahan[7] | J. Gregory Caporaso[8] | Emily Curd[9] | Christophe Djemiel[10] | Mikael Brandström Durling[11] | Vasco Elbrecht[6] | Zachary Gold[12] | Hyun S. Gweon[13,14] | Mehrdad Hajibabaei[15] | Falk Hildebrand[16,17] | Vladimir Mikryukov[1] | Eric Normandeau[18] | Ezgi Özkurt[16,17] | Jonathan M. Palmer[19] | Géraldine Pascal[20] | Teresita M. Porter[15] | Daniel Straub[21] | Martti Vasar[1] | Tomáš Větrovský[22] | Haris Zafeiropoulos[23] | Sten Anslan[1,24]

Metabarcoding processing and analysis pipelines

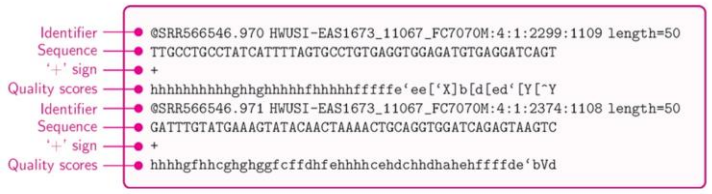
1. From raw-reads to OTU-table

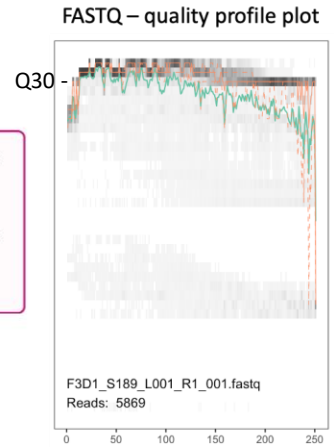
2. Statistical analysis

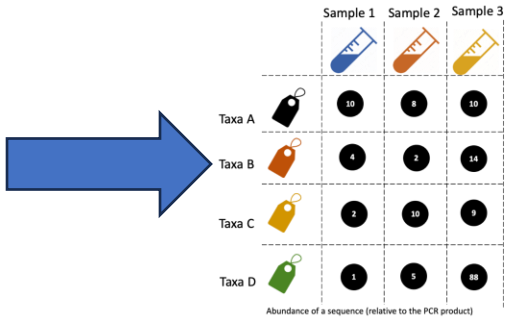
Metabarcoding processing and analysis pipelines

1. From raw-reads to OTU-table

fastq file







+

	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV_1	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA
ASV_2	k__Fungi	p__Ascomyc	c__Leotiomy	o__Helotiale	NA	NA	NA
ASV_3	k__Viridiplar	p__Anthophy	c__Monocot	o__Poales	f__Poaceae	g__Zea	s__mays
ASV_4	k__Fungi	p__Mortiere	c__Mortierel	o__Mortiere	f__Mortierel	g__Mortiere	s__minutissima
ASV_5	k__Fungi	p__Ascomyc	c__Sordarior	o__Hypocrea	f__Nectriace	g__Fusarium	NA
ASV_6	k__Fungi	p__Ascomyc	c__Dothideo	o__Pleospor	f__Phaeosph	g__Setophor	s__terrestris
ASV_7	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA

+

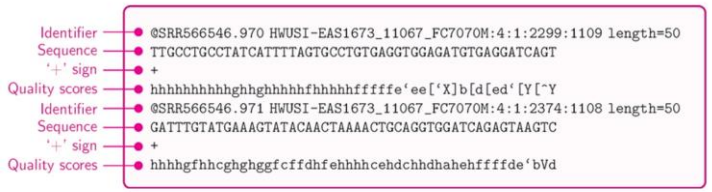
2. Statistical analysis

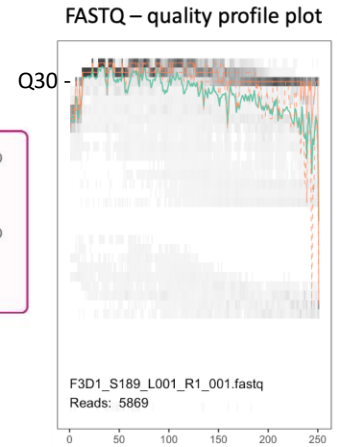
https://benjjneb.github.io/dada2/tutorial_1_6.html, Hosseini et al, 2016;

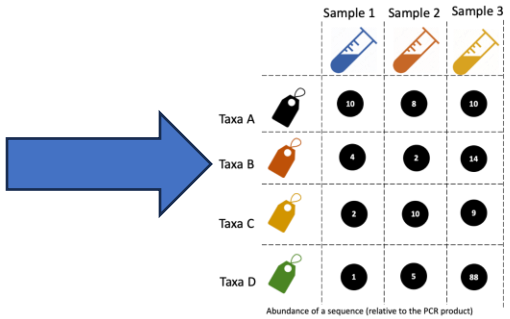
Metabarcoding processing and analysis pipelines

1. From raw-reads to OTU-table

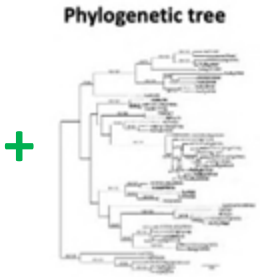
fastq file



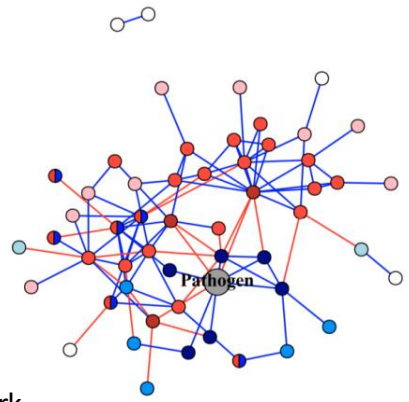




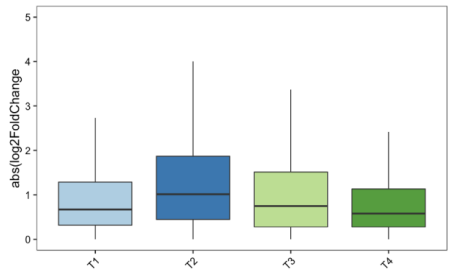
	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV_1	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA
ASV_2	k__Fungi	p__Ascomyc	c__Leotiomy	o__Helotiale	NA	NA	NA
ASV_3	k__Viridiplar	p__Anthophy	c__Monocoty	o__Poales	f__Poaceae	g__Zea	s__mays
ASV_4	k__Fungi	p__Mortiere	c__Mortierel	o__Mortiere	f__Mortierel	g__Mortiere	s__minutissima
ASV_5	k__Fungi	p__Ascomyc	c__Sordarior	o__Hypocrea	f__Nectriace	g__Fusarium	NA
ASV_6	k__Fungi	p__Ascomyc	c__Dothideo	o__Pleospor	f__Phaeosph	g__Setophor	s__terrestris
ASV_7	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA



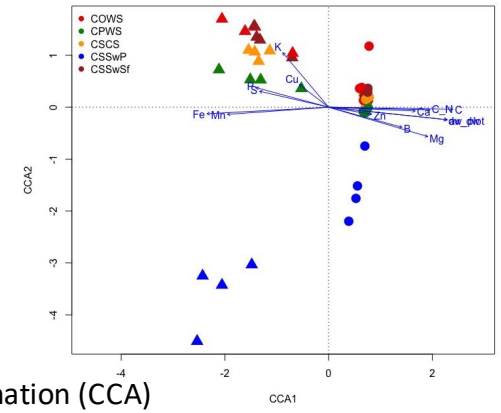
2. Statistical analysis (examples)



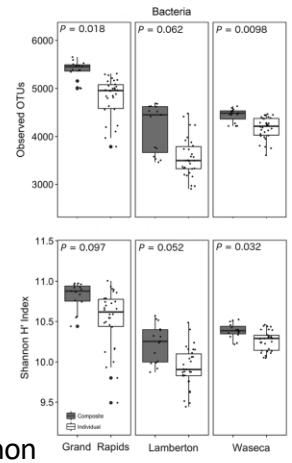
Network



Fold change



Ordination (CCA)



Richness and Shannon

https://benjjneb.github.io/dada2/tutorial_1_6.html, Hosseini et al, 2016; Poudel et al 2016; Castle et al 2019; Bullington et al 2021

From raw-reads to OTU-table

Marker gene analyses

Raw reads

INTEGRATED PIPELINES (step-by-step)

Exploration
FASTQC
SeqKit

Processing
FASTX-Toolkit
PRINSEQ
Trimmomatic
fastq-join (paired-end)
PEAR (paired-end)

OTUs/ASVs estimation
QIIME2
mothur
DADA2
Deblur
UNOISE3

ITS-extraction

Taxonomic classification
RDP classifier
q2-feature classifier (QIIME2)
UCLUST
BLAST

Decontamination

Functional prediction (optional)
PICRUSt2
Tax4Fun
PanFP

Normalization
DeseqR (R Bioconductor package)
edgeR (R Bioconductor package)
metagenomeSeq (R Bioconductor package)
R core software

Copy number correction
rrnDB (R package)
CopyRighter
pplacer
picante R package

QIIME 2
mothur
UPARSE

Samples
OTUs
Observation counts

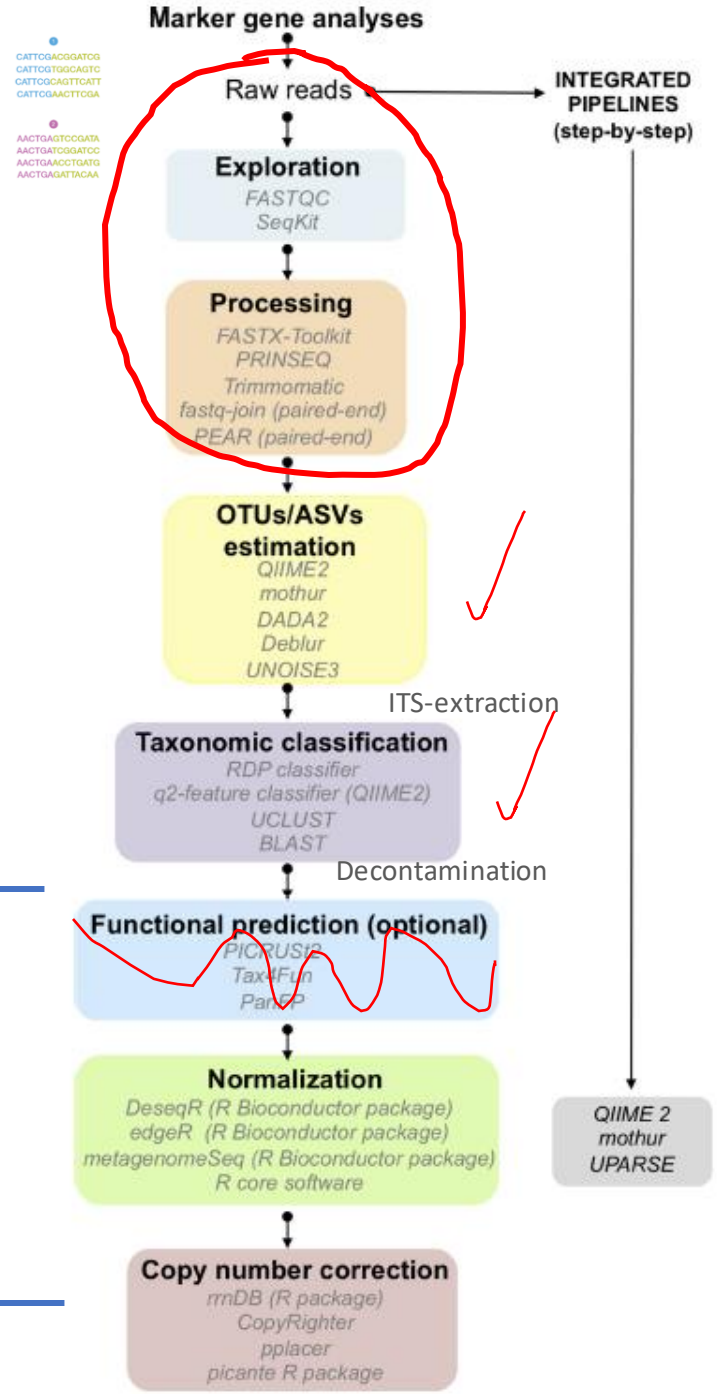
Samples
OTUs
Normalized data

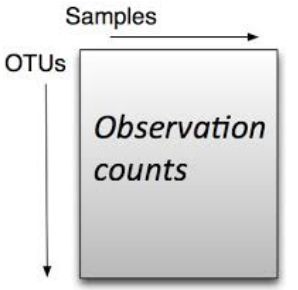
	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV_1	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA
ASV_2	k__Fungi	p__Ascomyc	c__Leotiomy	o__Helotiale	NA	NA	NA
ASV_3	k__Viridiplar	p__Anthophy	c__Monocot	o__Poales	f__Poaceae	g__Zea	s__mays
ASV_4	k__Fungi	p__Mortiere	c__Mortierel	o__Mortiere	f__Mortierel	g__Mortiere	s__minutissima
ASV_5	k__Fungi	p__Ascomyc	c__Sordarior	o__Hypocrea	f__Nectriace	g__Fusarium	NA
ASV_6	k__Fungi	p__Ascomyc	c__Dothideo	o__Pleospor	f__Phaeosph	g__Setophor	s__terrestris
ASV_7	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA

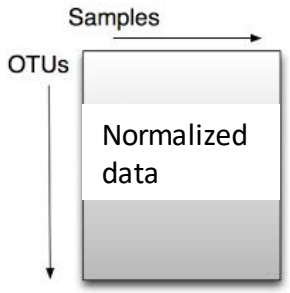
Phylogenetic tree

Perez-Cobas et al 2020

From raw-reads to OTU-table





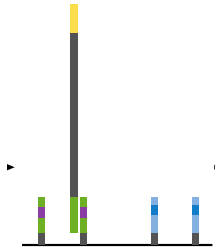


	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV_1	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA
ASV_2	k__Fungi	p__Ascomyc	c__Leotiomy	o__Helotiale	NA	NA	NA
ASV_3	k__Viridiplar	p__Anthophy	c__Monocot	o__Poales	f__Poaceae	g__Zea	s__mays
ASV_4	k__Fungi	p__Mortiere	c__Mortierel	o__Mortiere	f__Mortierel	g__Mortiere	s__minutissima
ASV_5	k__Fungi	p__Ascomyc	c__Sordarior	o__Hypocrea	f__Nectriace	g__Fusarium	NA
ASV_6	k__Fungi	p__Ascomyc	c__Dothideo	o__Pleospor	f__Phaeosph	g__Setophor	s__terrestris
ASV_7	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA

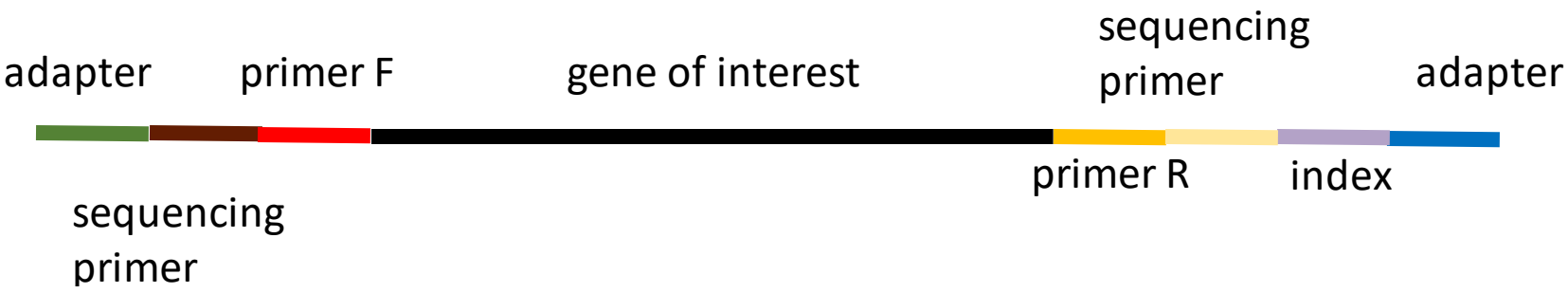
Phylogenetic tree

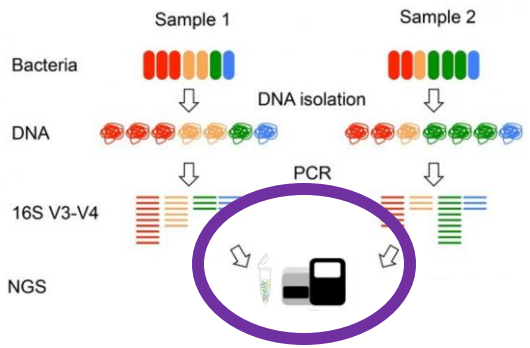
Perez-Cobas et al 2020

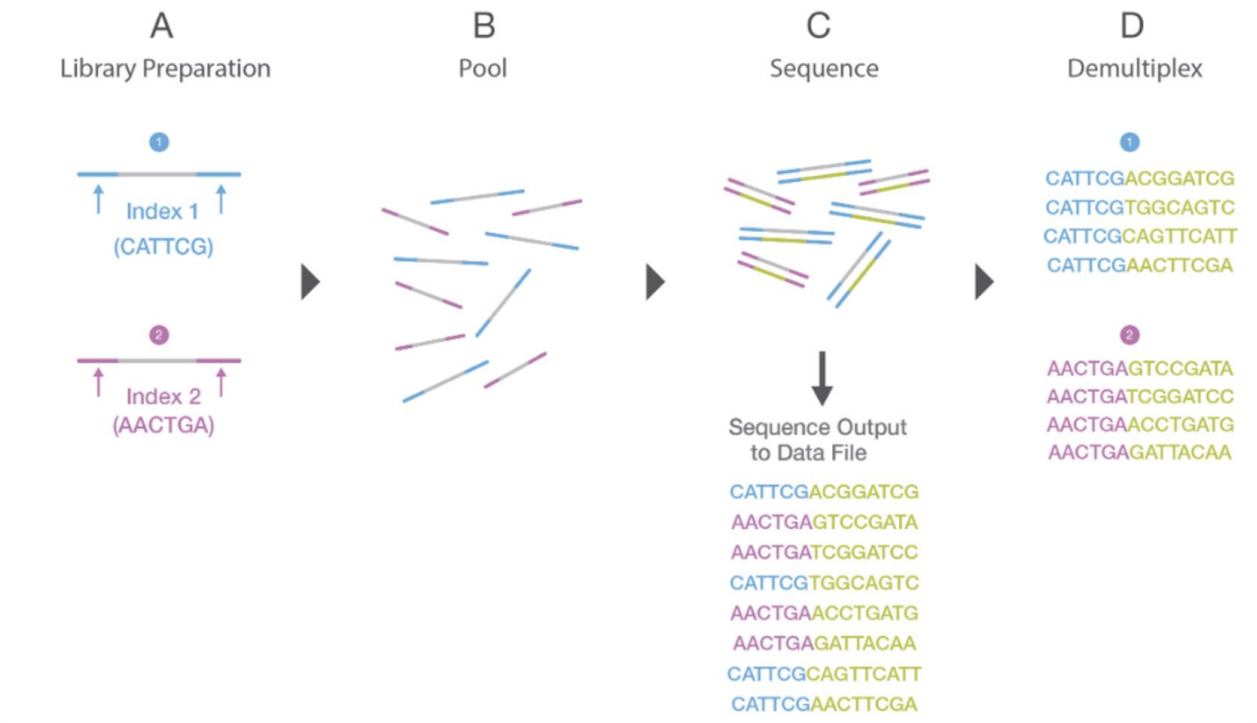
Structure of the amplicon (sequencing read)



Illumina flow cell



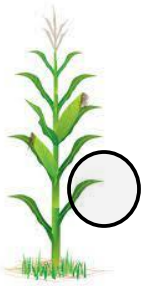


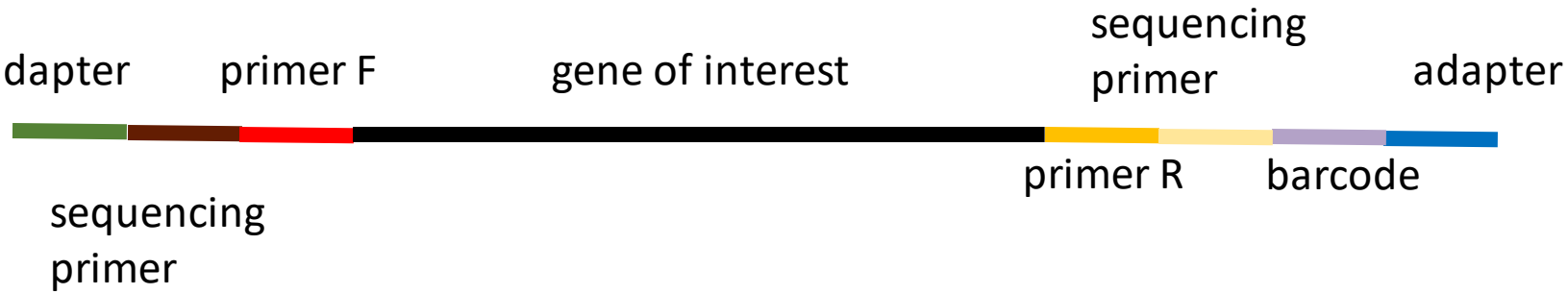


Metabarcoding – different indexes allows the sequencing of multiple samples at once

Lundberg et al 2013, Caporaso et al 2011, www.illumina.com

Exploration and processing of reads





R1

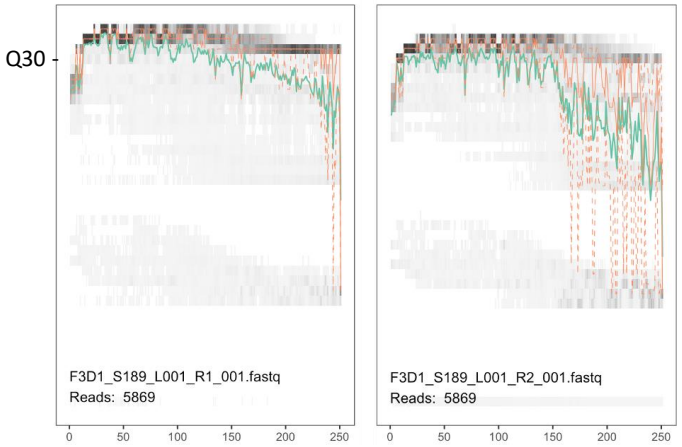
fastq file

```
Identifier      @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence        TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign        +
Quality scores  hhhhhhhhhhghghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier      @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence        GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign        +
Quality scores  hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd
```

R1

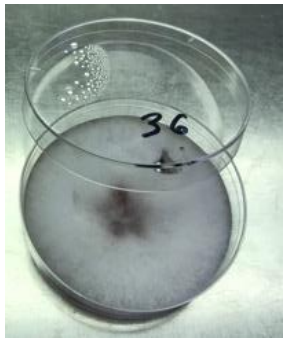
R2

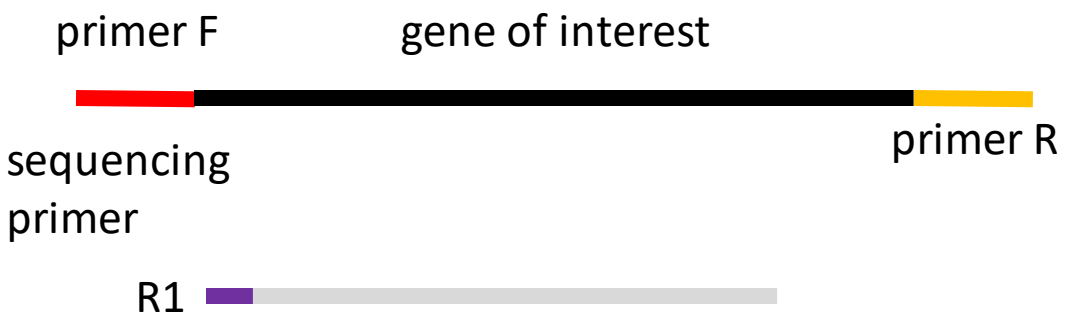
FASTQ – quality profile plot

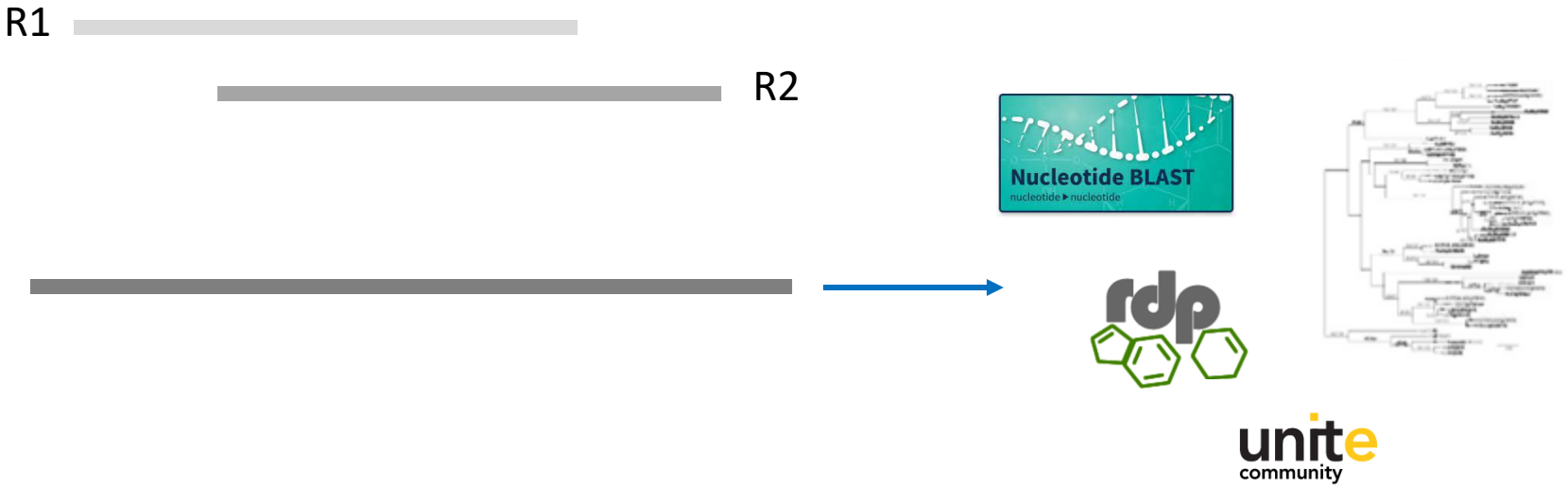


https://benjjneb.github.io/dada2/tutorial_1_6.html, Hosseini et al, 2016; Lundberg et al 2013, Caporaso et al 2011,www.illumina.com

Comparison with Sanger sequencing

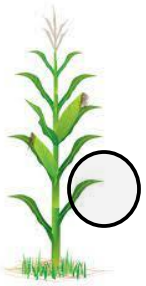


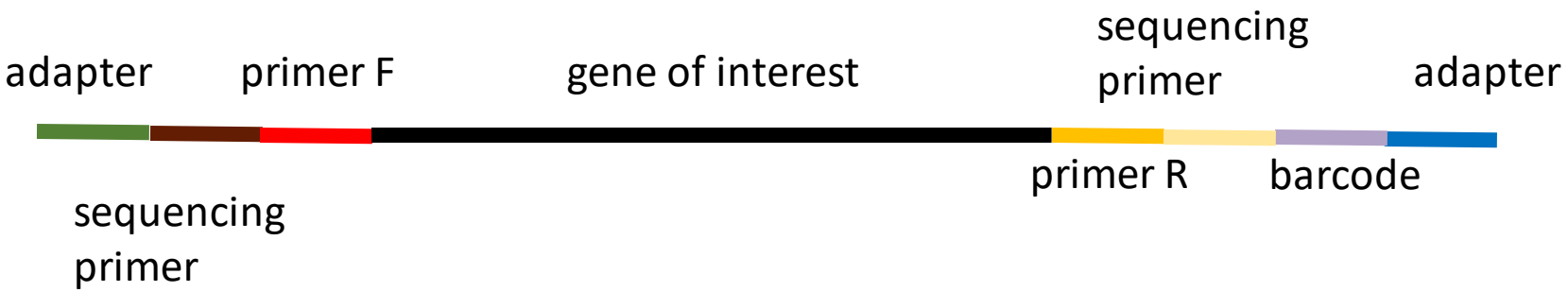




https://benjjneb.github.io/dada2/tutorial_1_6.html, Hosseini et al, 2016; Lundberg et al 2013, Caporaso et al 2011,www.illumina.com

Exploration and processing of reads





R1

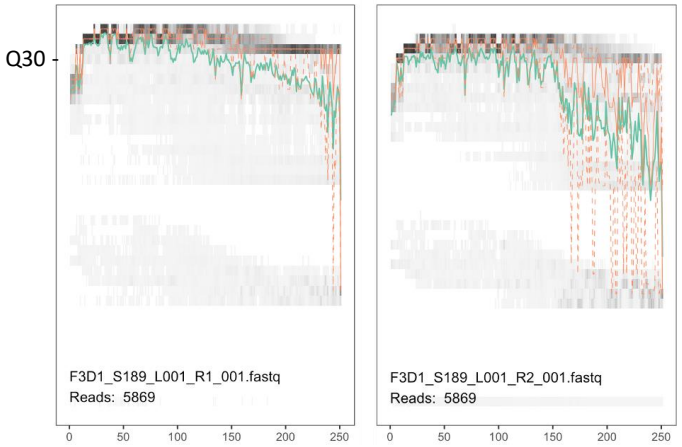
fastq file

```
Identifier      @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence        TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign        +
Quality scores  hhhhhhhhhhghghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier      @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence        GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign        +
Quality scores  hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd
```

R1

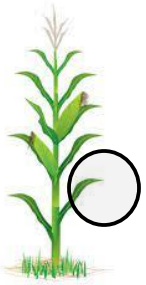
R2

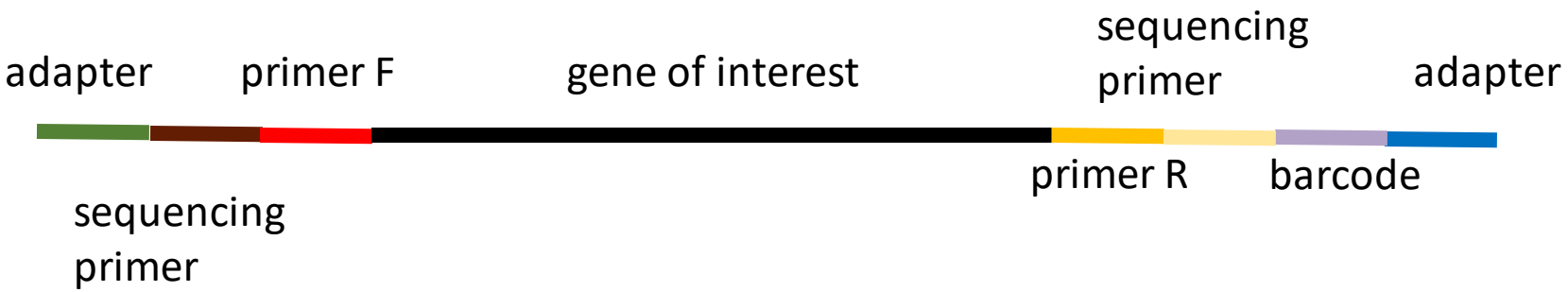
FASTQ – quality profile plot



https://benjjneb.github.io/dada2/tutorial_1_6.html, Hosseini et al, 2016; Lundberg et al 2013, Caporaso et al 2011,www.illumina.com

Exploration and processing of reads





R1

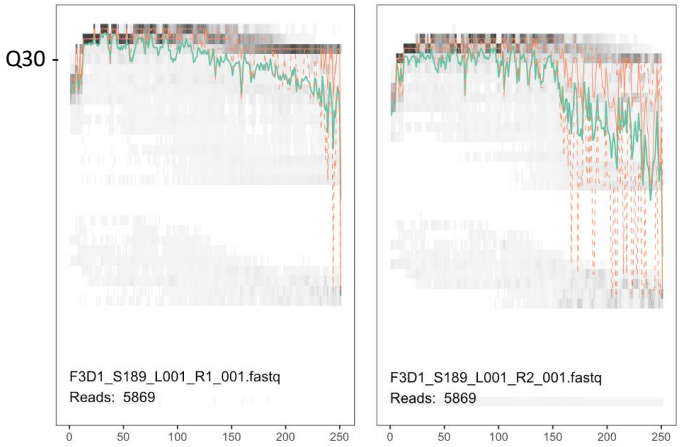
fastq file

```
Identifier      @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence        TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign        +
Quality scores  hhhhhhhhhhghghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier      @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence        GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign        +
Quality scores  hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd
```

R1

R2

FASTQ – quality profile plot

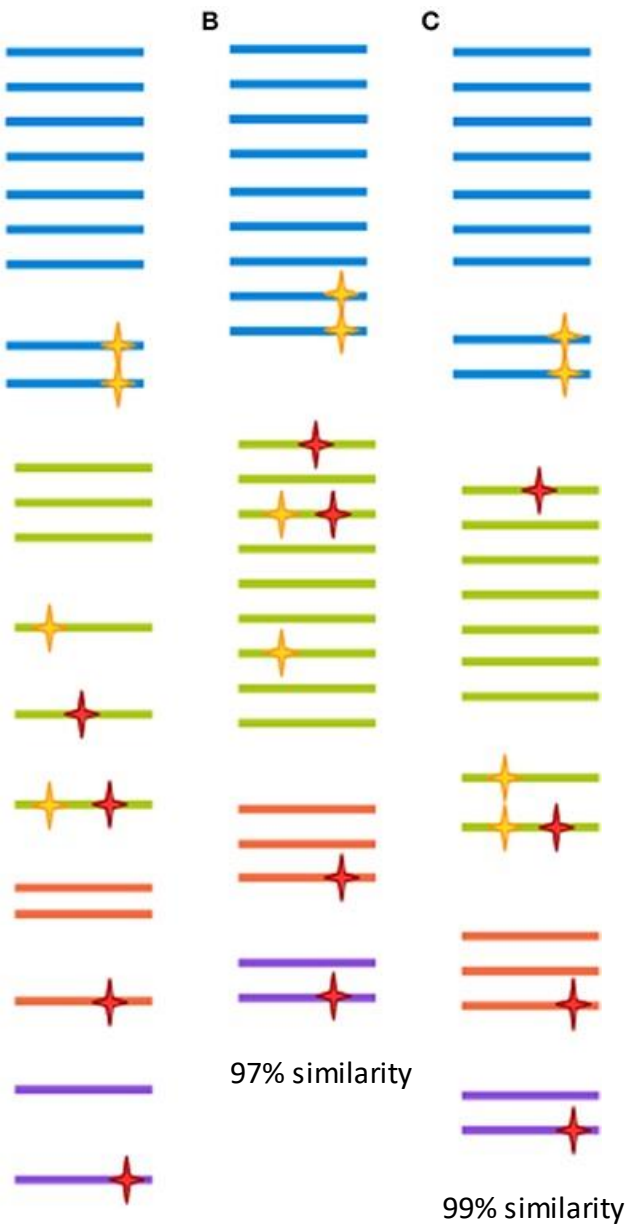


Taxonomic unit of analysis

https://benjjneb.github.io/dada2/tutorial_1_6.html, Hosseini et al, 2016; Lundberg et al 2013, Caporaso et al 2011,www.illumina.com

Estimation of the taxonomic unit of analysis (*De novo*)

Algorithm:



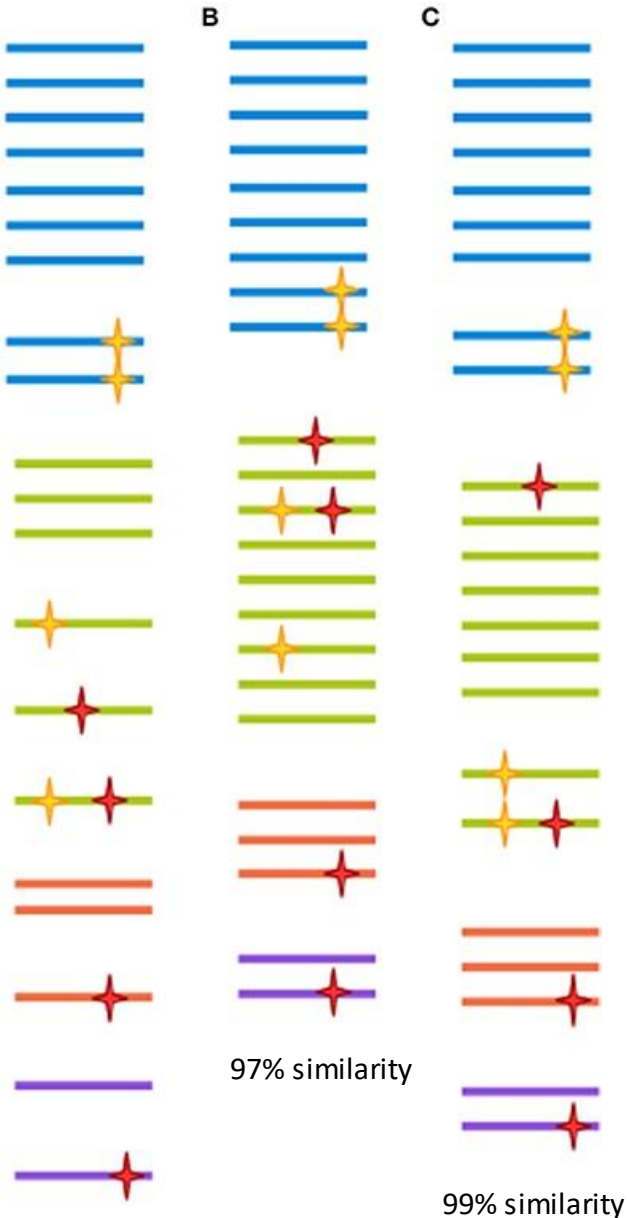
yellow star: single nucleotide difference
red star: sequencing error

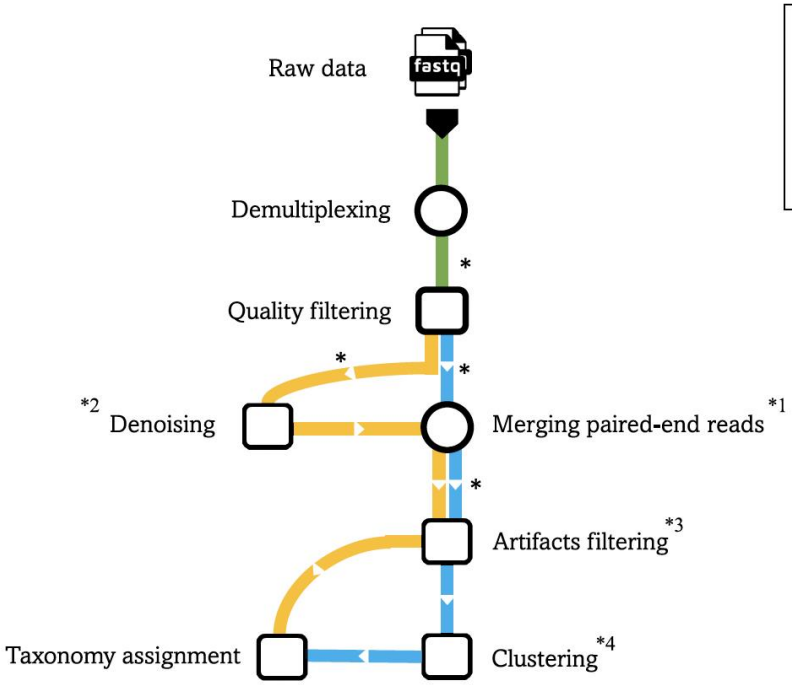
Single nucleotide difference

Hugerth and Andersson 2017

Estimation of the taxonomic unit of analysis (*De novo*)

Algorithm:





yellow star: single nucleotide difference
red star: sequencing error

Single nucleotide difference

Hakimzadeh et al 2023

Hugerth and Andersson 2017

Denoising and error correction

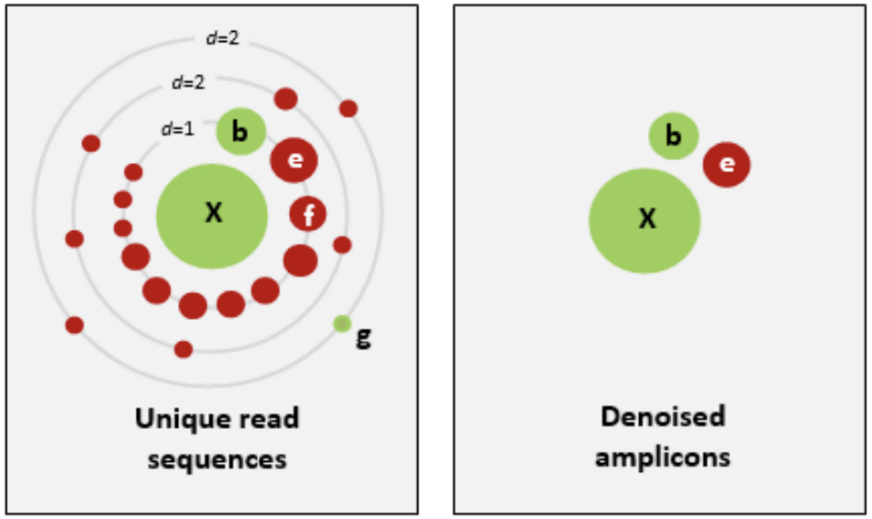


Figure 1. Schematic of the UNOISE2 denoising strategy. The left panel shows the neighborhood close to a high-abundance unique read sequence **X**, grouped by the number of sequence differences (*d*). Dots are unique sequences, the size of a dot indicates its abundance. Green dots are correct biological sequences; red dots have one or more errors. Neighbors with small numbers of differences and small abundance compared to **X** are predicted to be bad reads of **X.** The right panel shows the denoised amplicons. Here, **X** and **b** were correctly predicted, **e** is an error with anomalously high abundance that was wrongly predicted to be correct, **f** is an error that was correctly discarded but has an abundance almost high enough to be a false positive, and **g** is a low-abundance correct amplicon that was wrongly discarded. The abundances of **b**, **e**, and **f** are similar, illustrating the fundamental challenge in denoising: how to set an abundance threshold that distinguishes correct sequences from errors. *zOTUs (zero radius) or ESV (exact sequence variant)

Edgar, 2016 https://doi.org/10.1101/081257

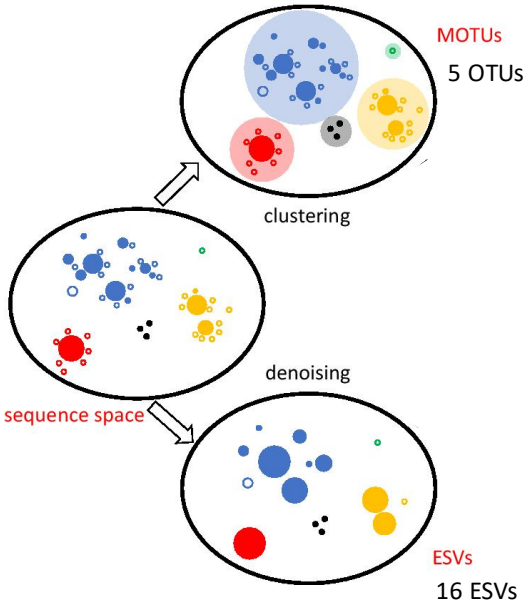
ASVs or OTUs?

~~ASVs or OTUs?~~

Does the pipeline incorporate denoising and error correction?

"ASVs are identical denoised reads with as few as 1 base pair difference between variants, representing an inference of the biological sequences prior to amplification and sequencing errors (Callahan et al., 2017)".

OTUs represent clusters of sequences based on a specified similarity threshold. One sequence representative is chosen for further analysis.



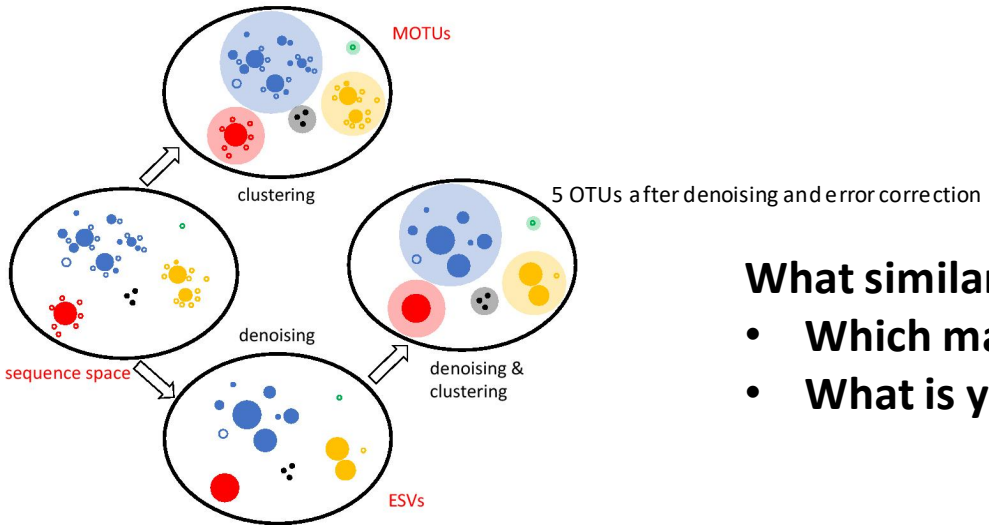
Antich etal 2021, Hakimzadeh et al 2023

~~ASVs or OTUs?~~

Does the pipeline incorporate denoising and error correction?

"ASVs are identical denoised reads with as few as 1 base pair difference between variants, representing an inference of the biological sequences prior to amplification and sequencing errors (Callahan et al., 2017)".

OTUs represent clusters of sequences based on a specified similarity threshold. One sequence representative is chosen for further analysis.



What similarity threshold to cluster with?

- **Which marker gene did you use?**
- **What is your desired taxonomic resolution?**

Antich etal 2021, Hakimzadeh et al 2023

Relationship between gene region variability, % similarity, and ‘genome splitting’

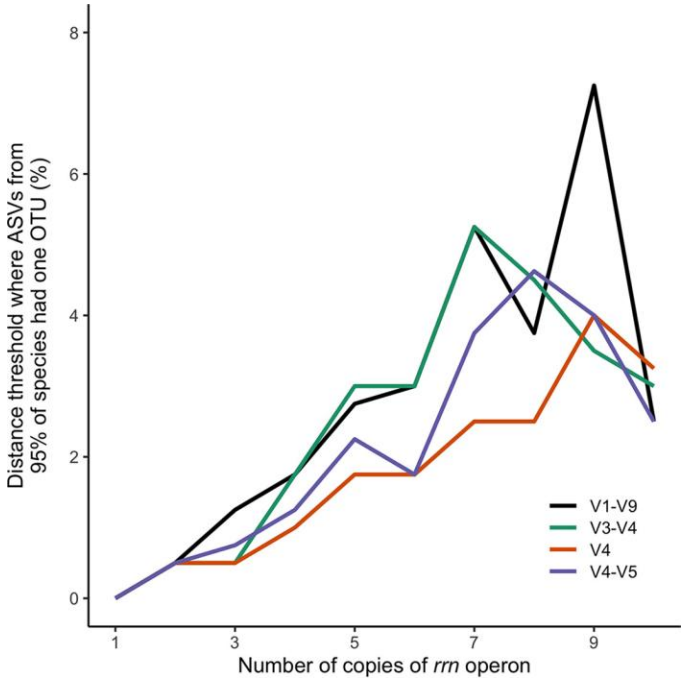


FIG 1 The distance threshold required to prevent the splitting of genomes into multiple OTUs increased as the number of *rrn* operons in the genome increased. Each line represents the median distance threshold for each region of the 16S rRNA gene that is required for 95% of the genomes with the indicated number of *rrn* operons to cluster their ASVs to a single OTU. The median distance threshold was calculated across 100 randomizations in which one genome was sampled from each species. Only those numbers of *rrn* operons that were found in more than 100 species are included.

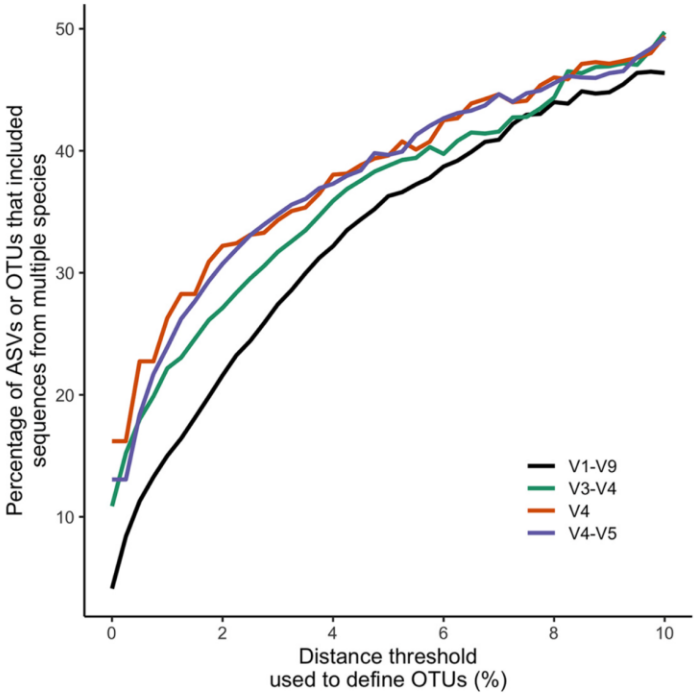
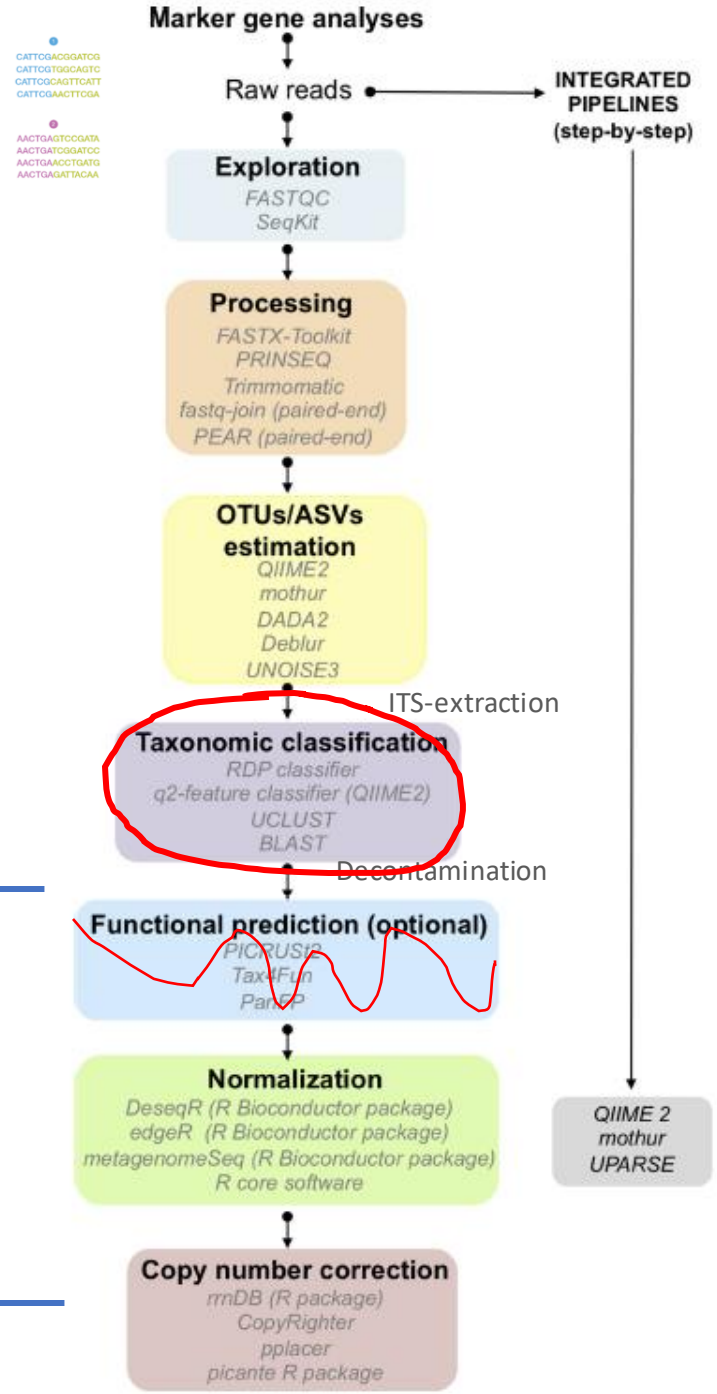
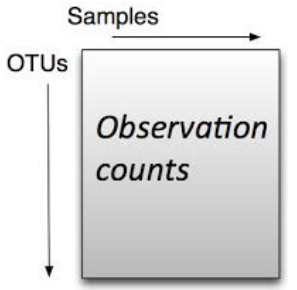


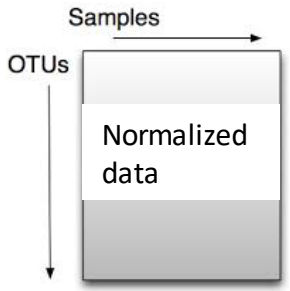
FIG 2 As the distance threshold used to define an OTU increased, the percentages of ASVs and OTUs representing multiple species increased. These data represent the median fractions for both measurements across 100 randomizations. In each randomization, one genome was sampled from each species.

Schloss, 2021

From raw-reads to OTU-table







	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV_1	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA
ASV_2	k__Fungi	p__Ascomyc	c__Leotiomy	o__Helotiale	NA	NA	NA
ASV_3	k__Viridiplar	p__Anthophy	c__Monocot	o__Poales	f__Poaceae	g__Zea	s__mays
ASV_4	k__Fungi	p__Mortiere	c__Mortierel	o__Mortiere	f__Mortierel	g__Mortiere	s__minutissima
ASV_5	k__Fungi	p__Ascomyc	c__Sordarior	o__Hypocrea	f__Nectriace	g__Fusarium	NA
ASV_6	k__Fungi	p__Ascomyc	c__Dothideo	o__Pleospor	f__Phaeosph	g__Setophor	s__terrestris
ASV_7	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA

Phylogenetic tree

Perez-Cobas et al 2020

Taxonomic classification and databases

Example

>ASV1
GAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAACGGTGAAGCAG
GAGCTTGCTCTTGTGGATCAGTGGCGAACGGGTGAGTAACACGTGAGCAACCTGCCCCGAACTCTGGGA
TAAGCGCTGGAAACGGCGTCTAATACTGGATATGCACCAGGGAGGCATCTTCACTGGTGGGAAAGATTTT
TTGGTTCGGGATGGGCTCGCGGCCTATCAGCTTGTTGGTGAGGTAACGGCTCACCAAGGCGTCGACGGG
TAGCCGGCCTGAGAGGGTGACCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAG
CAGTGGGGAATATTGCACAATGGGCGGAAGCCTGATGCAGCAACGCCGC

Classification

Bacteria

Firmicutes

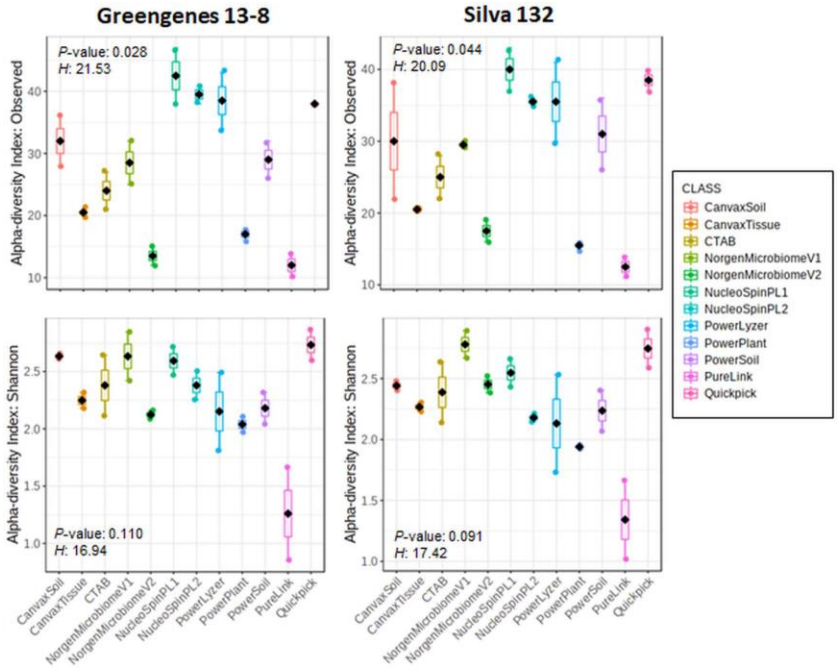
Bacilli

Bacillales

Bacillaceae

Bacillus

Bacillus subtilis



*depends on the taxonomic resolution of your marker gene (and its length)

Taxonomic classification and databases

Example

>ASV1
GAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAACGGTGAAGCAG
GAGCTTGCTCTTGTGGATCAGTGGCGAACGGGTGAGTAACACGTGAGCAACCTGCCCCGAACTCTGGGA
TAAGCGCTGGAAACGGCGTCTAATACTGGATATGCACCAGGGAGGCATCTTCACTGGTGGGAAAGATTTT
TTGGTTCGGGATGGGCTCGCGGCCTATCAGCTTGTTGGTGAGGTAACGGCTCACCAAGGCGTCGACGGG
TAGCCGGCCTGAGAGGGTGACCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAG
CAGTGGGGAATATTGCACAATGGGCGGAAGCCTGATGCAGCAACGCCGC

Classification

Bacteria

Firmicutes

Bacilli

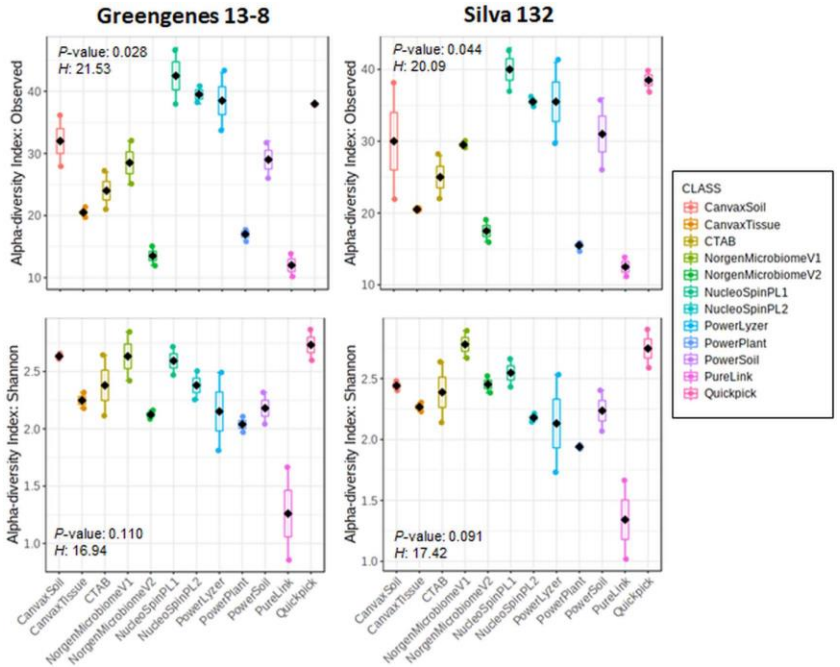
Bacillales

Bacillaceae

Bacillus

Bacillus subtilis

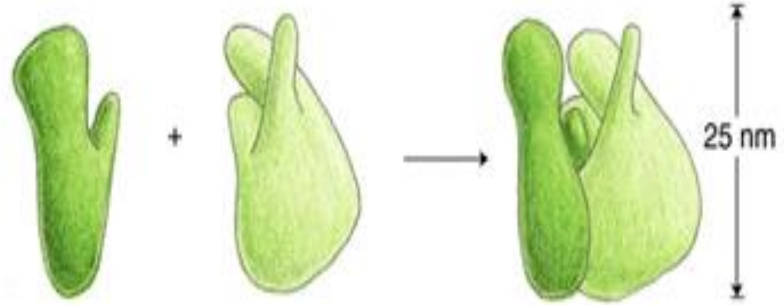
Tree building (all ASVs)



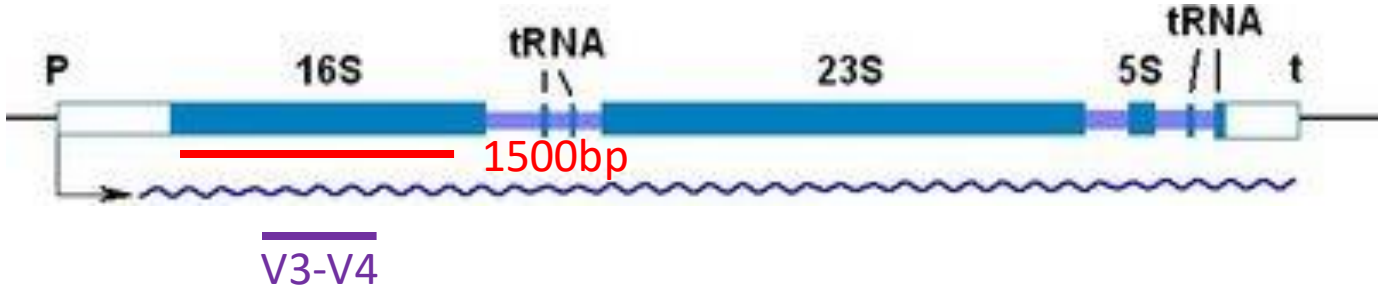
*depends on the taxonomic resolution of your marker gene (and its length)

Ribosomal markers as taxonomic barcodes

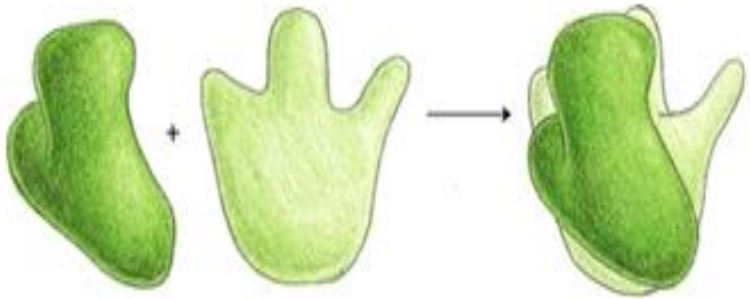
30S 50S



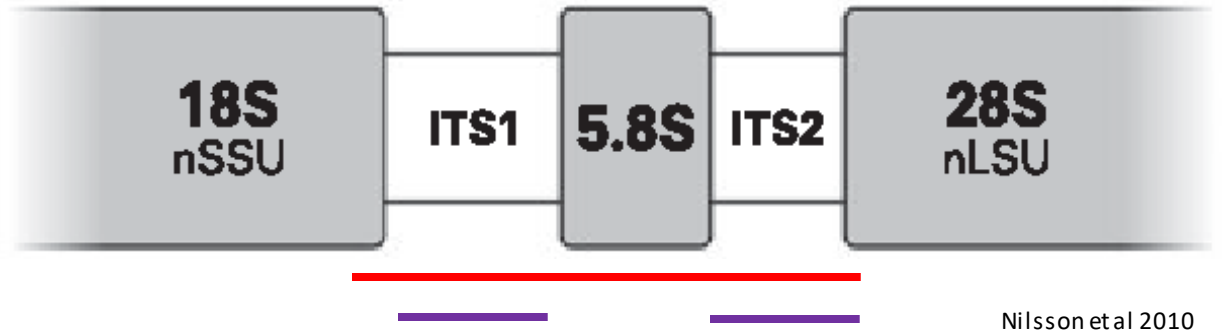
Bacteria (chloroplast, mitochondria)



40S 60S



Eukarya (Fungi, Nematodes, Protists, Plants)



Nilsson et al 2010

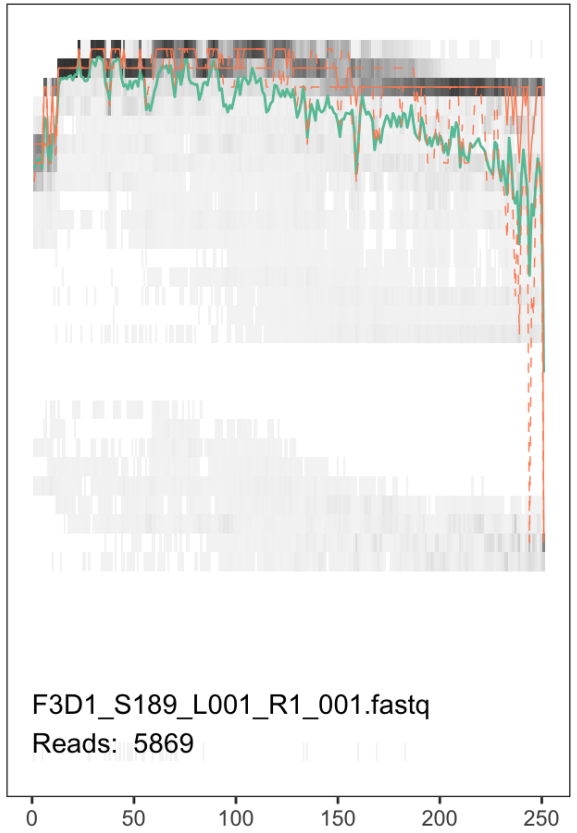
Things to look for in a pipeline

- Different quality control steps:
  - Sequence length (trimming)
  - Low quality reads (filtering)
  - Deal with sequencing errors (denoising/error correction)
  - Homopolymers
  - Chimeras

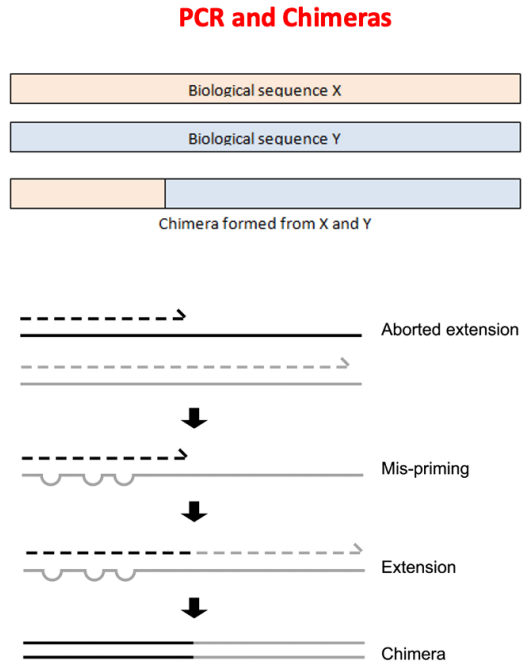
- OTU vs SV
- Deal with controls

- Documentation
- Format/compatibility with downstream analysis

FASTQ – quality profile plot



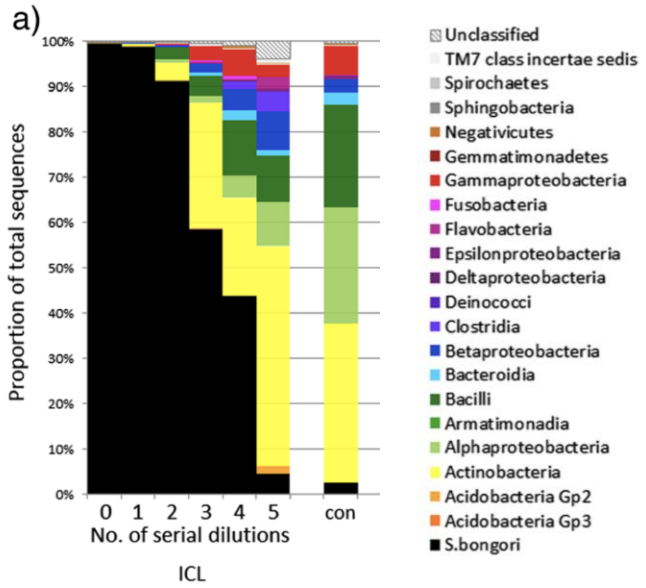
https://benjjneb.github.io/dada2/tutorial_1_6.html



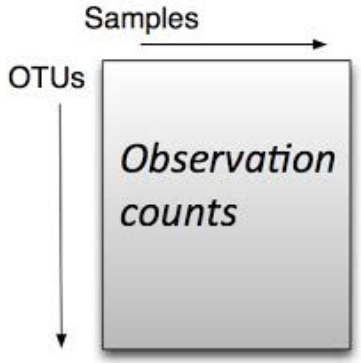
Bahram et al (2019) doi:10.1111/1758-2229.12684; Haas et al 2011

Negative controls and identifying contaminants

Contaminants in low biomass samples



con=template free PCR
http://www.biomedcentral.com/1741-7007/12/87



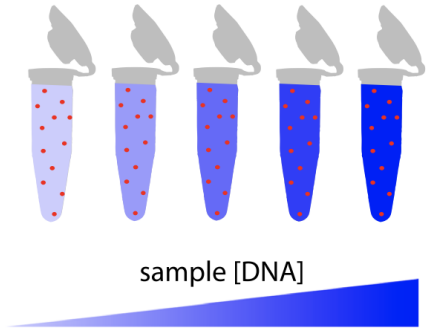
Davis *et al. Microbiome* (2018) 6:226
https://doi.org/10.1186/s40168-018-0605-2

Microbiome

METHODOLOGY Open Access

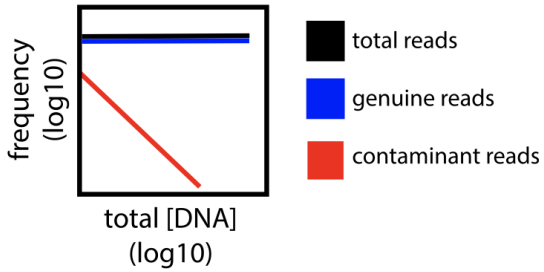
Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data

Nicole M. Davis[1], Diana M. Proctor[2,3], Susan P. Holmes[4], David A. Relman[1,2,5] and Benjamin J. Callahan[6,7*]



equal, low-level contaminating DNA

sequence equimolar amounts well-mixed total DNA



contaminant DNA correlates inversely with total DNA

Technical replicates or other standards

Phytobiomes • 2018 • 2:165-170 https://doi.org/10.1094/PBIOMES-09-17-0041-R

OPEN ACCESS phytobiomesjournal.org

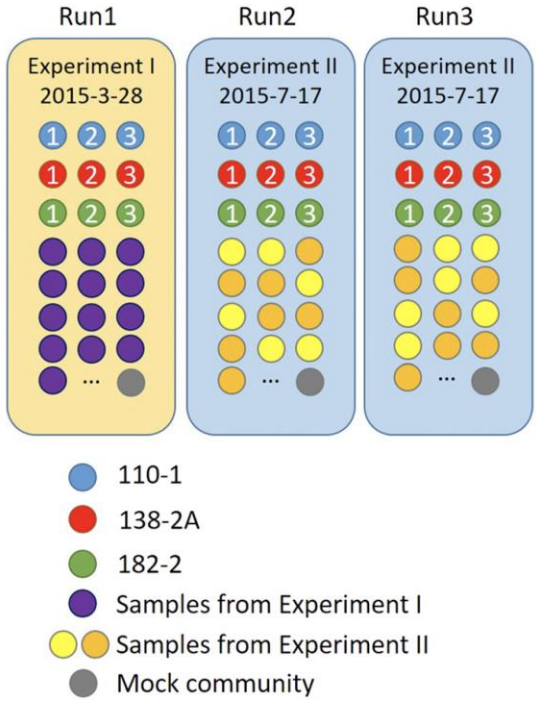
Phytobiomes Journal A Transdisciplinary Journal of Sustainable Plant Productivity

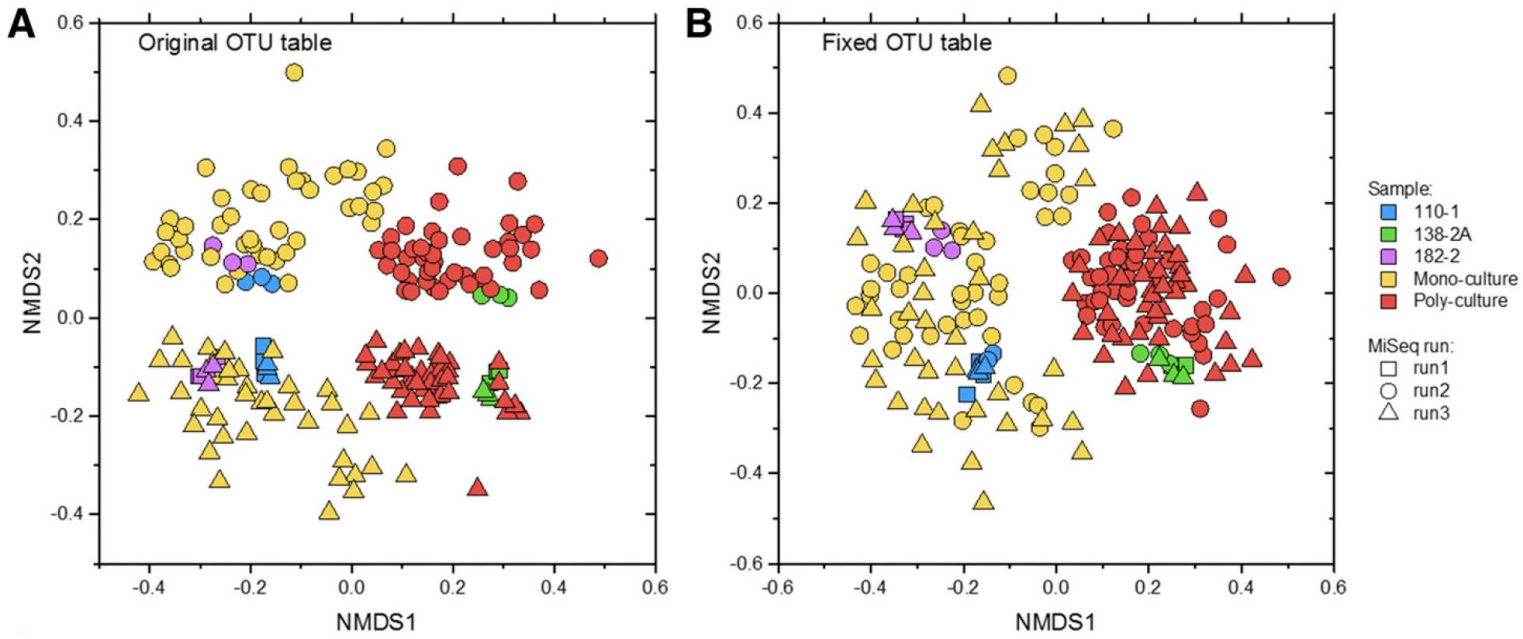
RESEARCH e-Xtra*

Run-to-Run Sequencing Variation Can Introduce Taxon-Specific Bias in the Evaluation of Fungal Microbiomes

Zewei Song† and **Dan Schlatter,** Department of Plant Pathology, University of Minnesota, Saint Paul; **Daryl M. Gohl**, University of Minnesota Genomics Center, Minneapolis; and **Linda L. Kinkel,** Department of Plant Pathology, University of Minnesota, Saint Paul

Experiment design





Mock communities and positive controls

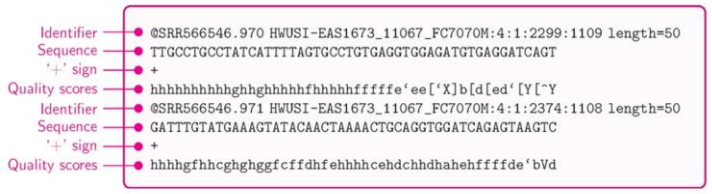
Product	Catalog #	Composition	Format
Microbial Community Standard	D6300	Even Distribution	Microbial
Microbial Community DNA Standard	D6305/6306	Even Distribution	Isolated DNA
Microbial Community Standard II	D6310	Log Distribution	Microbial
Microbial Community DNA Standard II	D6311	Log Distribution	Isolated DNA
Spike-in Control I (High Microbial Load)	D6320/D6320-10	Even Distribution	Microbial
Spike-in Control II (Low Microbial Load)	D6321/D6321-10	Log Distribution	Microbial
HMW DNA Standard	D6322	Even Distribution	Isolated DNA
Gut Microbiome Standard	D6331	Staggered Abundance	Microbial

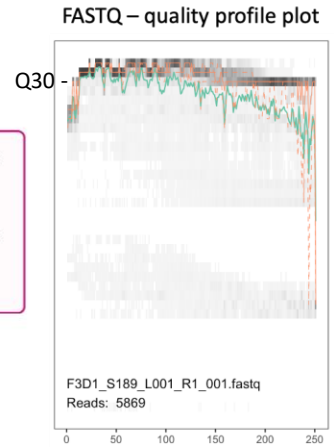
https://www.zymoresearch.com/pages/microbiome-standards

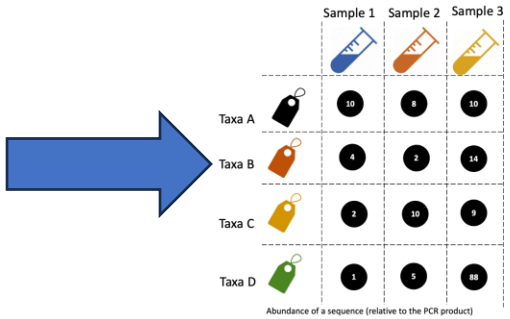
Metabarcoding processing and analysis pipelines

1. From raw-reads to OTU-table

fastq file







+

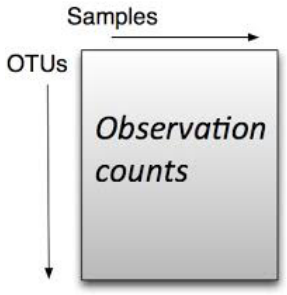
	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV_1	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA
ASV_2	k__Fungi	p__Ascomyc	c__Leotiomy	o__Helotiale	NA	NA	NA
ASV_3	k__Viridiplar	p__Anthophy	c__Monocot	o__Poales	f__Poaceae	g__Zea	s__mays
ASV_4	k__Fungi	p__Mortiere	c__Mortierel	o__Mortiere	f__Mortierel	g__Mortiere	s__minutissima
ASV_5	k__Fungi	p__Ascomyc	c__Sordarior	o__Hypocrea	f__Nectriace	g__Fusarium	NA
ASV_6	k__Fungi	p__Ascomyc	c__Dothideo	o__Pleospor	f__Phaeosph	g__Setophor	s__terrestris
ASV_7	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA

+

2. Statistical analysis

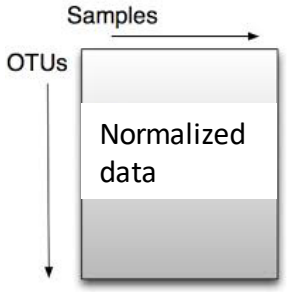
https://benjjneb.github.io/dada2/tutorial_1_6.html, Hosseini et al, 2016;

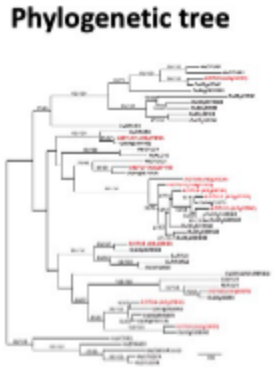
Analysing and interpreting your data



	A	B	C	D	E	F
1		sample01	sample01	sample02	sample02	sample03
2	OTU0001	528	68	1755	773	2167
3	OTU0002	138	68	559	2588	1198
4	OTU0003	36	533	673	351	815
5	OTU0004	1	2618	5	17	19
6	OTU0005	224	237	81	271	313

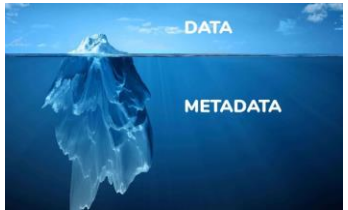
	Kingdom	Phylum	Class	Order	Family	Genus	Species
ASV_1	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA
ASV_2	k__Fungi	p__Ascomyc	c__Leotiomy	o__Helotiale	NA	NA	NA
ASV_3	k__Viridiplar	p__Anthophy	c__Monocoty	o__Poales	f__Poaceae	g__Zea	s__mays
ASV_4	k__Fungi	p__Mortiere	c__Mortierel	o__Mortiere	f__Mortierel	g__Mortiere	s__minutissima
ASV_5	k__Fungi	p__Ascomyc	c__Sordarior	o__Hypocrea	f__Nectriace	g__Fusarium	NA
ASV_6	k__Fungi	p__Ascomyc	c__Dothideo	o__Pleospora	f__Phaeosph	g__Setophor	s__terrestris
ASV_7	k__Viridiplar	p__Anthophy	c__Eudicotyl	o__Fabales	f__Fabaceae	g__Glycine	NA





**phyloseq* object

Metadata



Phytobiomes Journal • 2020 • 4:115-121 https://doi.org/10.1094/PBIOMES-09-19-0051-P

OPEN ACCESS phytobiomesjournal.org APS

Phytobiomes Journal A Transdisciplinary Journal of Sustainable Plant Productivity

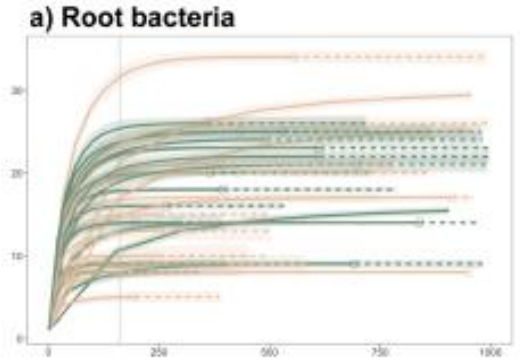
PERSPECTIVE

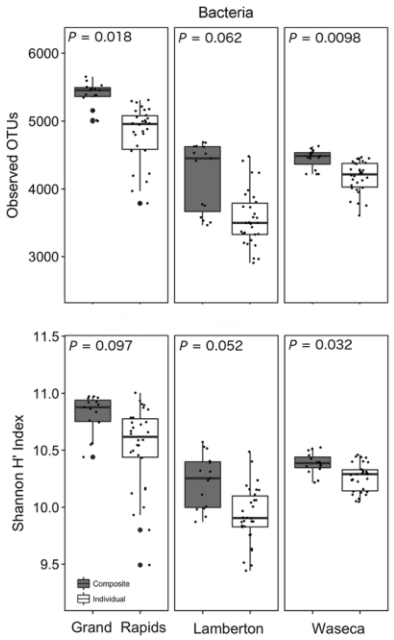
Community-Driven Metadata Standards for Agricultural Microbiome Research

J. P. Dundore-Arias,[1,†] E. A. Eloe-Fadrosh,[2] L. M. Schriml,[3] G. A. Beattie,[4] F. P. Brennan,[5] P. E. Busby,[6] R. B. Calderon,[7] S. C. Castle,[8] J. B. Emerson,[9] S. E. Everhart,[10] K. Eversole,[11] K. E. Frost,[12] J. R. Herr,[13] A. I. Huerta,[14] A. S. Iyer-Pascuzzi,[15] A. K. Kalil,[16] J. E. Leach,[17] J. Leonard,[18] J. E. Maul,[19] B. Prithiviraj,[20] M. Potrykus,[21] N. R. Redekar,[22] J. A. Rojas,[23] K. A. T. Silverstein,[24] D. J. Tomso,[25] S. G. Tringe,[26] B. A. Vinatzer,[27] and L. L. Kinkel[28]

Which analyses to apply?

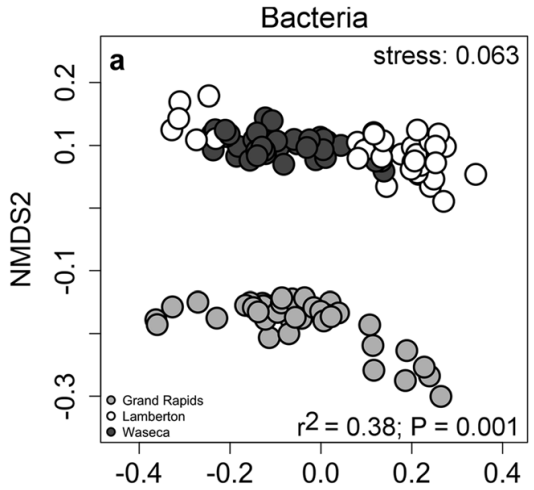
Alpha-diversity: within a community





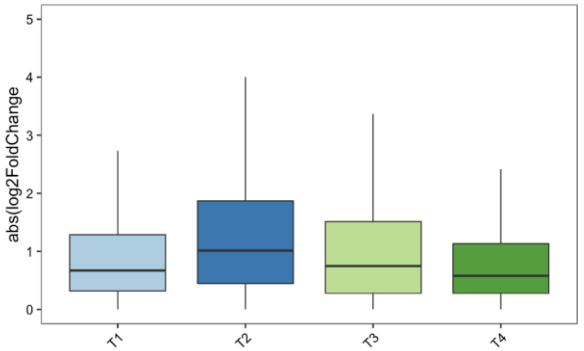
Richness
Evenness
Phylogenetic diversity
…

Beta-diversity: between communities

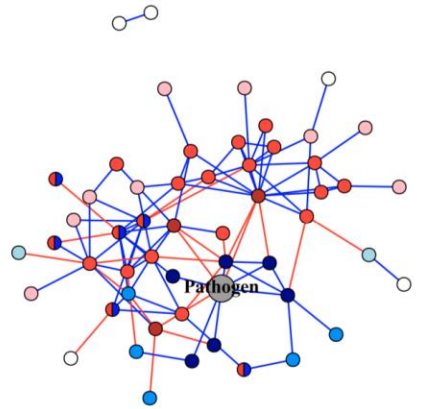


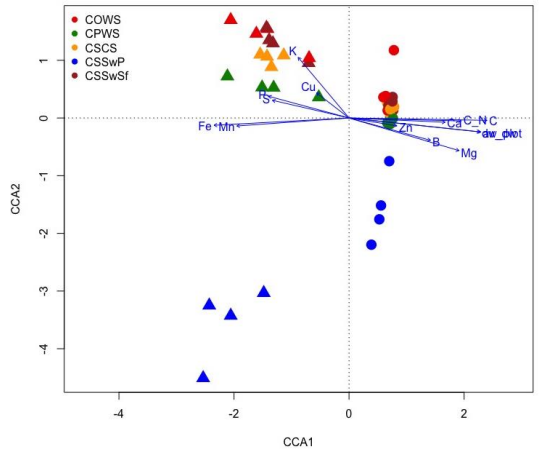
Similarity/dissimilarity indices
E.g: Bray-Curtis, Unifrac…

Differential abundance tests



Co-occurrence and interactions between microbes; and correlation with environmental variables

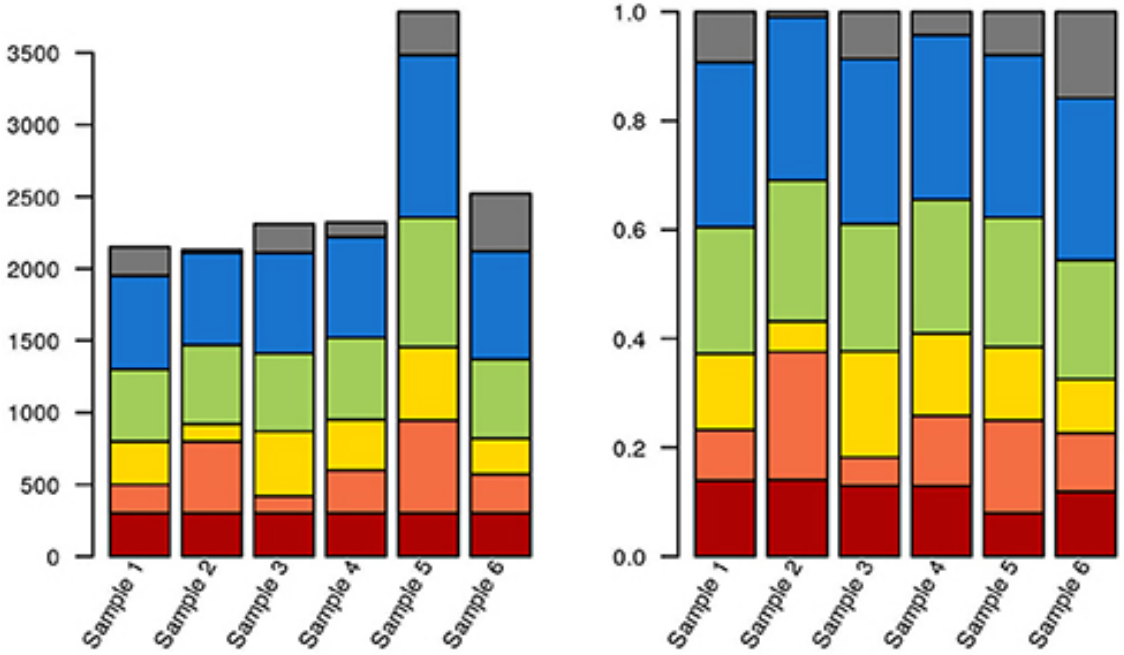




Poudel et al 2016; Castle et al 2019; Bullington et al 2021

To normalize or not?

What is your question and hypothesis?



McMurdie, Paul J., and Susan Holmes. “Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible.” *PLoS Comput Biol* 10, no. 4 (2014): e1003531.

Weiss, Sophie, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, et al. “Normalization and Microbial Differential Abundance Strategies Depend upon Data Characteristics.” *Microbiome* 5, no. 1 (March 3, 2017): 27. https://doi.org/10.1186/s40168-017-0237-y.

Schloss PD. 2024. Waste not, want not: revisiting the analysis that called into question the practice of rarefaction. mSphere 9:e00355-23.https://doi.org/10.1128/msphere.00355-23

Variation in RNA operon copy number

rrndb: the Ribosomal RNA Operon Copy Number Database

Table 1.
Intra-genomic 16S rRNA variability for Bacteria and Archaea with full-genome sequence availability

Organism	No. rRNA[a] operons	Diff. (nt)[b]	% difference[c]
Aquifex aeolicus VF5	2	–	–
Bacillus subtilis ATCC 23857	10	1–15	0.97
Campylobacter jejuni ATCC 700819	3	–	–
Deinococcus radiodurans ATCC 13939	3	0–2	0.13
Escherichia coli ATCC 10798	7	0–19	1.23
Haemophilus influenzae ATCC 51907	6	–	–
Helicobacter pylori 26695	2	–	–
Methanococcus jannaschii DSMZ 2661	2	3	0.20
Methanococcus themoautotrophicum ATCC 29096	2	2	0.14
Neisseria menigitidis MC 58	4	–	–
Treponema pallidum ATCC 25870	2	–	–
Ureaplasma urealyticum serovar 3	2	1	0.07
Vibrio cholerae ATCC 39315	8	0–14	0.91
Xyella fastidosa 9a5c	2	–	–

[a]Number of rRNA operons per genome.

[b]Pairwise difference range between 16S rRNA genes per genome.

[c]Pairwise difference range between 16S rRNA genes per genome calculated as a percentage. –, no nucleotide differences.

Klappenbach et al 2000

RESEARCH ARTICLE

Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold

Marlène Chiarello[1]*, Mark McCauley[1], Sébastien Villéger[2], Colin R. Jackson[1]

1 Department of Biology, University of Mississippi, University, MS, United States of America, **2** MARBEC, University of Montpellier, CNRS, Ifremer, IRD, Montpellier, France

* marlene.chiarello@gmail.com

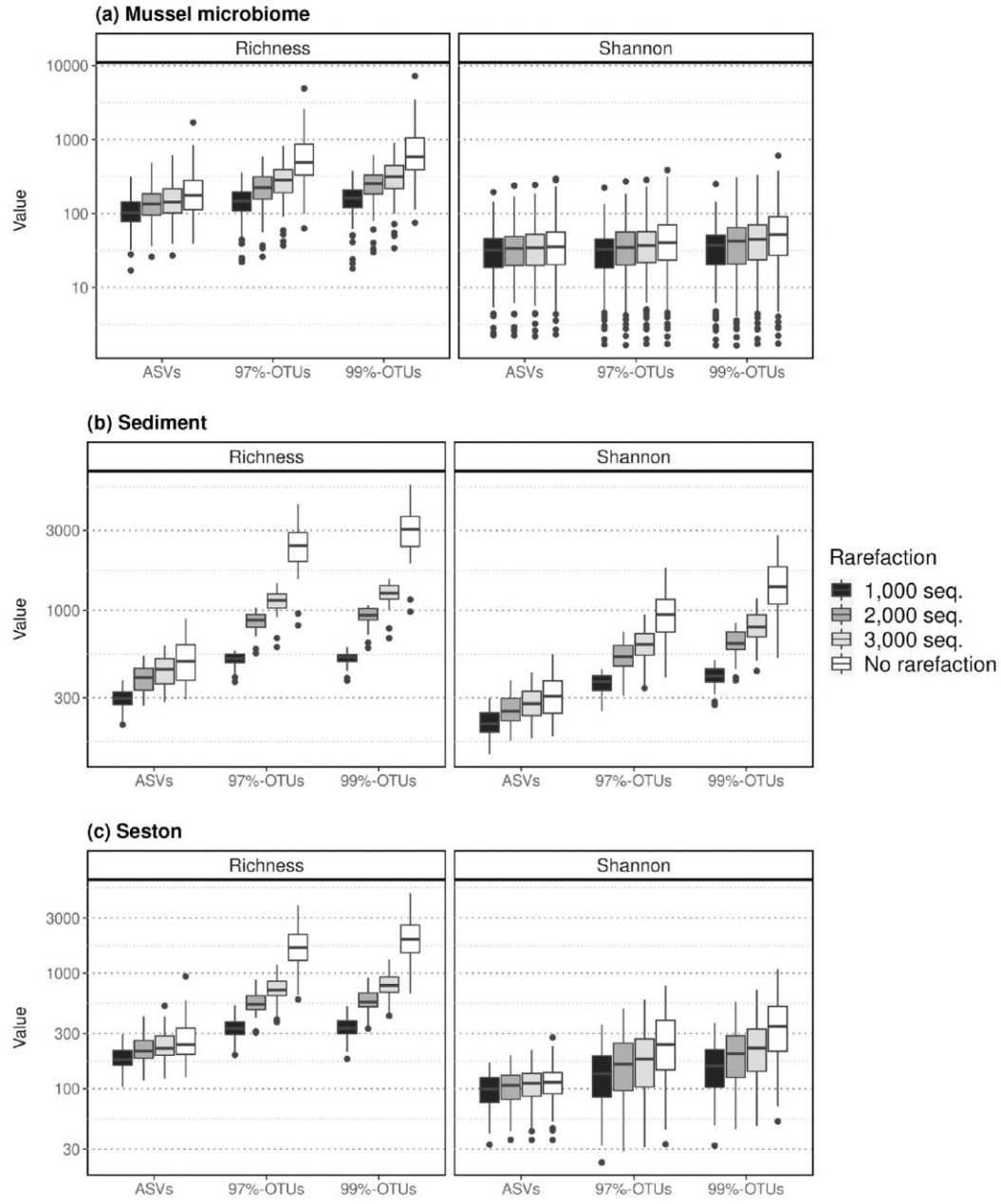
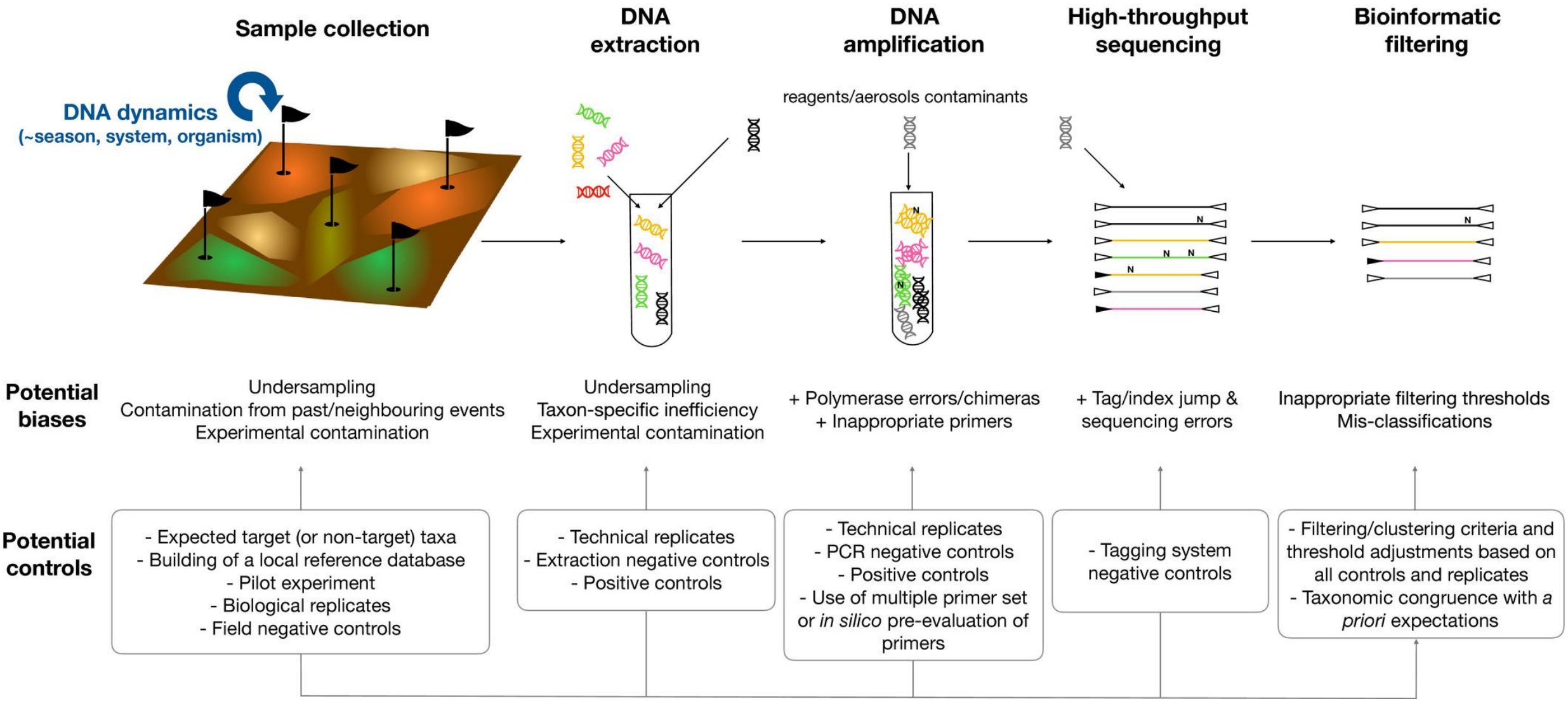


Fig 2. Effect of sequence processing methodology on alpha diversity metrics for microbiome analyses. Measure of taxonomic richness (expressed as the number of OTUs or ASVs and Shannon alpha diversity depending on the methodology used, namely ASV, 97%-OTU, and 99%-OTU, at three different levels of rarefaction (1,000; 2,000 and 3,000 sequences per sample) within each sample type studied (a-c). To ease the visualization of differences across methods and rarefaction levels, the y-axis of plots has been log transformed. The effect of methodological choices on other alpha diversity metrics is available in S3 Fig in S1 File.

https://doi.org/10.1371/journal.pone.0264443.g002



Zinger etal 2019