Metabarcoding (amplicon sequencing)

Upstream analysis

Miguel Semedo

2024/09/04

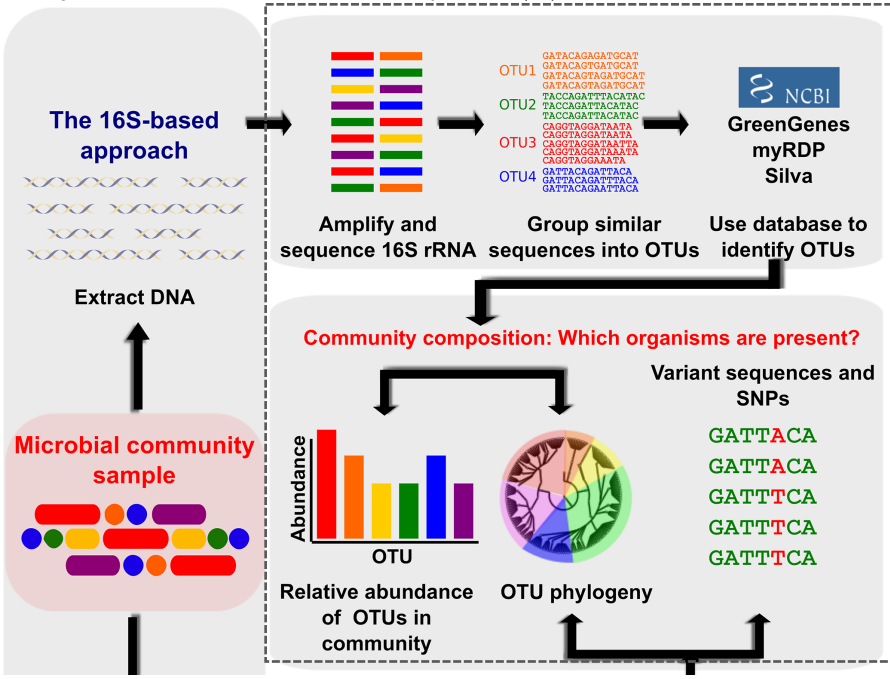
Funding

Sponsorship

Support

Amplicon sequencing – What for?

Morgan and Huttenhower, 2012. PLoS Comput Biol 8(12):e1002808.



Targeted Metagenomics / Metabarcoding / Amplicon Sequencing

Marker genes amplification (e.g. 16S rRNA, others)

- **WHO'S THERE (IN GENERAL)?**
  - With taxonomic markers: What can they possibly do (functional inference)?
  - With functional genes: Who's doing this specific function (potentially)?

16S rRNA gene sequencing

- Taxonomy marker for prokaryotes (others for eukaryotes - **can you name a few examples**?).
- How does composition and diversity vary across samples/conditions?
- What is the core microbiome of this particular sample type?
- What microorganisms are present exclusively in certain samples?
- What are the main biotic and abiotic interactions (can be specific – hypothesis driven)?
- What is the effect of a pollutant on bacterial diversity?

Why the 16S rRNA gene?

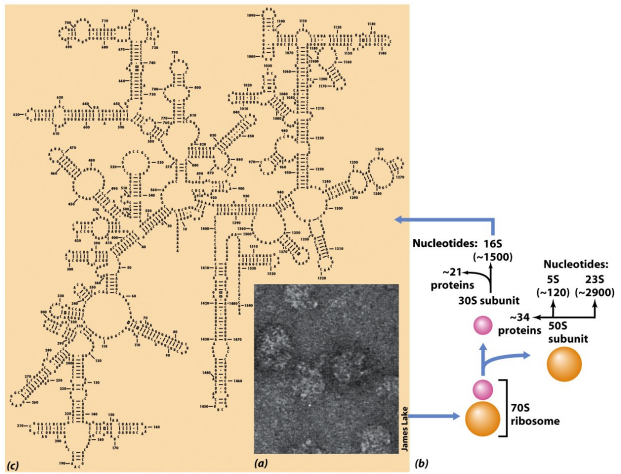
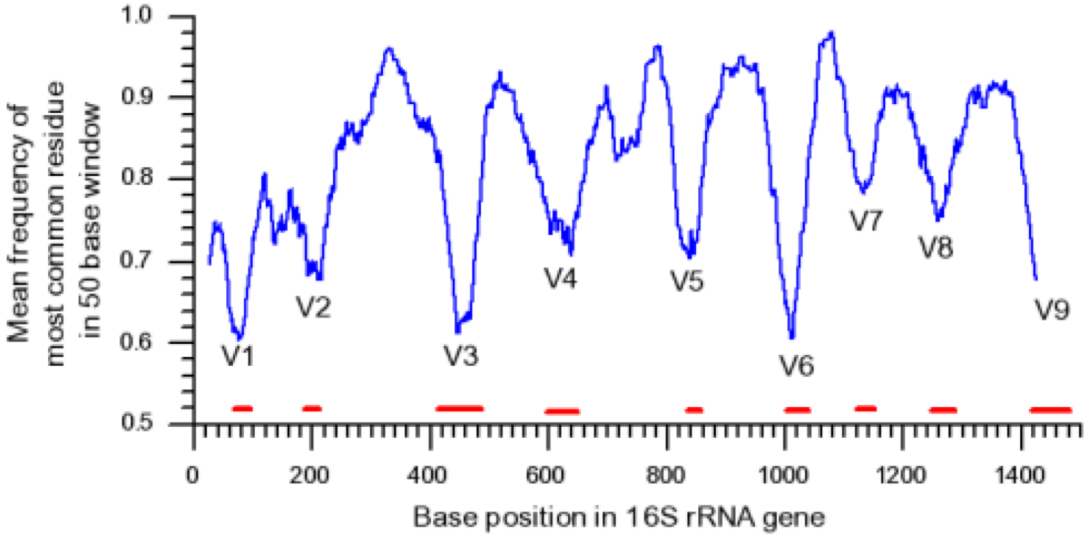
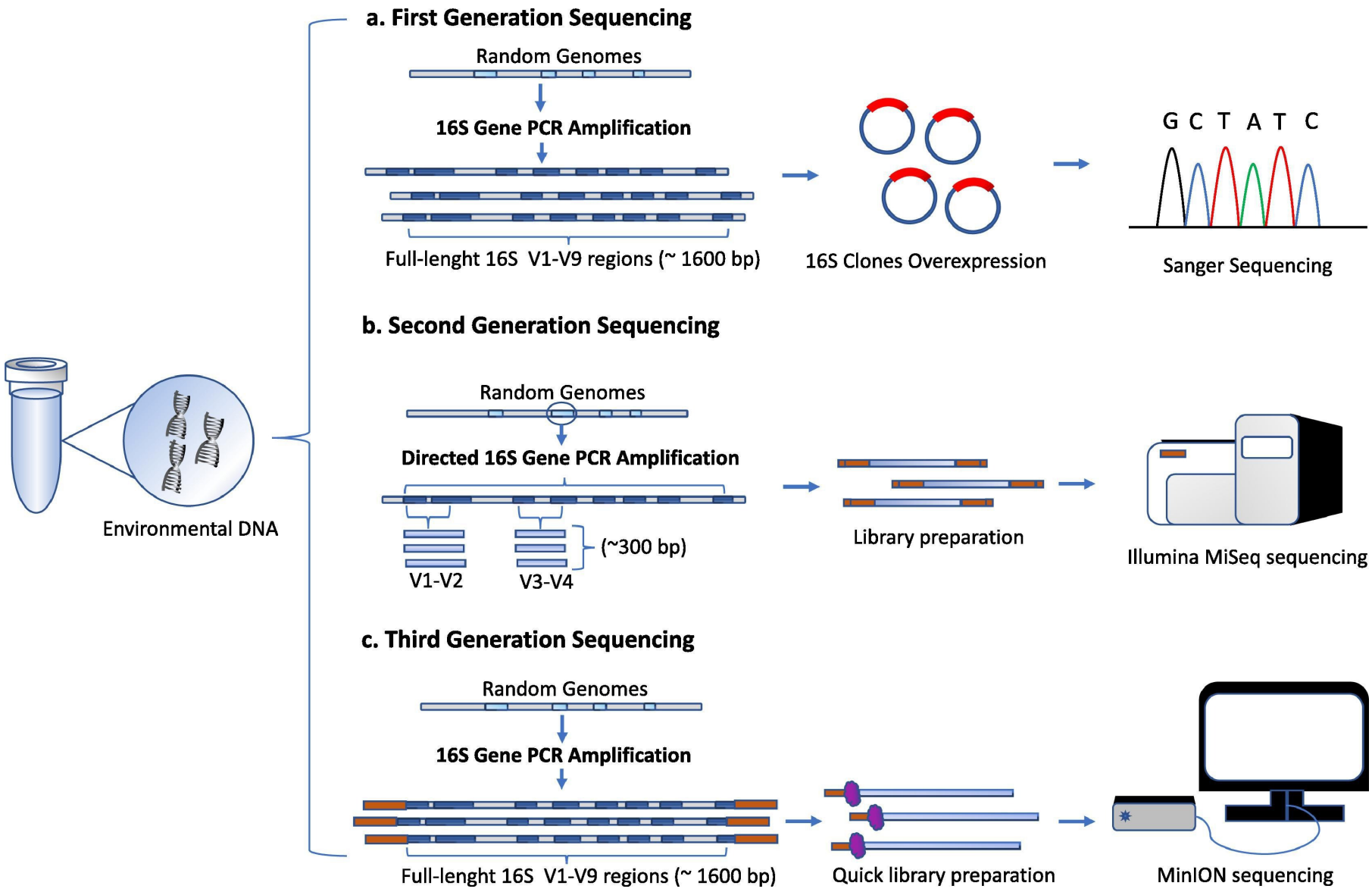


Figure 11-11 Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.



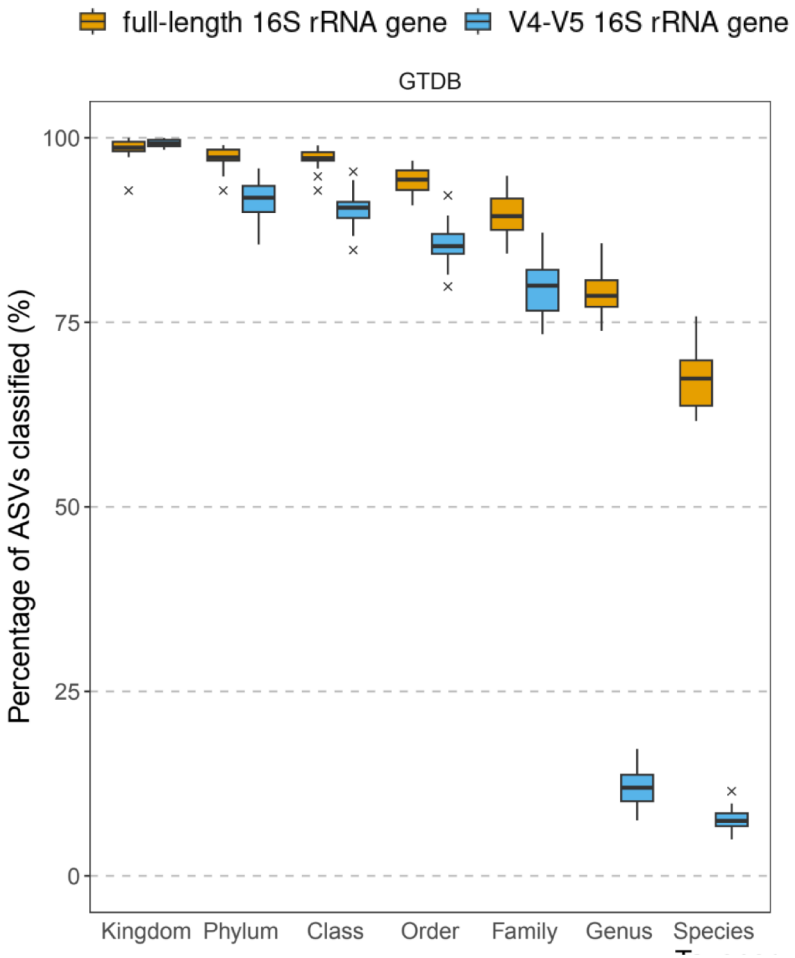
- RNA component of the small ribosomal unit in prokaryotes
- Universal marker for bacterial taxonomy
- Found in all bacteria and archaea (not single copy-gene in all cases – careful!). It can also be detected in eukaryotes. **This advantage can also be a problem, which one?**
- Different regions with different variability levels (highly conserved across all taxa and highly variable)
- Phylogenies derived from 16S rRNA gene tend to agree well with other conserved genes (evolution)

16S rRNA gene sequencing – *"3 technology generations"*



Santos et al., 2020. Comput. Structu. Biotech.

16S rRNA gene sequencing – short vs. long-read



Pascoal et al. 2024
https://annalsmicrobiology.biomedcentral.com/articles/10.1186/s13213-024-01767-6

Long-read sequencing (PacBio or ONT) tend to improve taxonomic assignment, especially at the lowest ranks (Genus, Species). Why do we still use short-read sequencing (Illumina)?

- High yield (number of reads per run)
- Low cost (vs PacBio, this may change o/t)
- Low error rate (vs ONT)
- Comparability with past studies and large datasets
- Difficulty to have long DNA fragments (sample quality)

Ultimately, it depends on your research needs...and sample type + opportunities/access to sequencing facilities.

16S rRNA gene sequencing – typical workflow

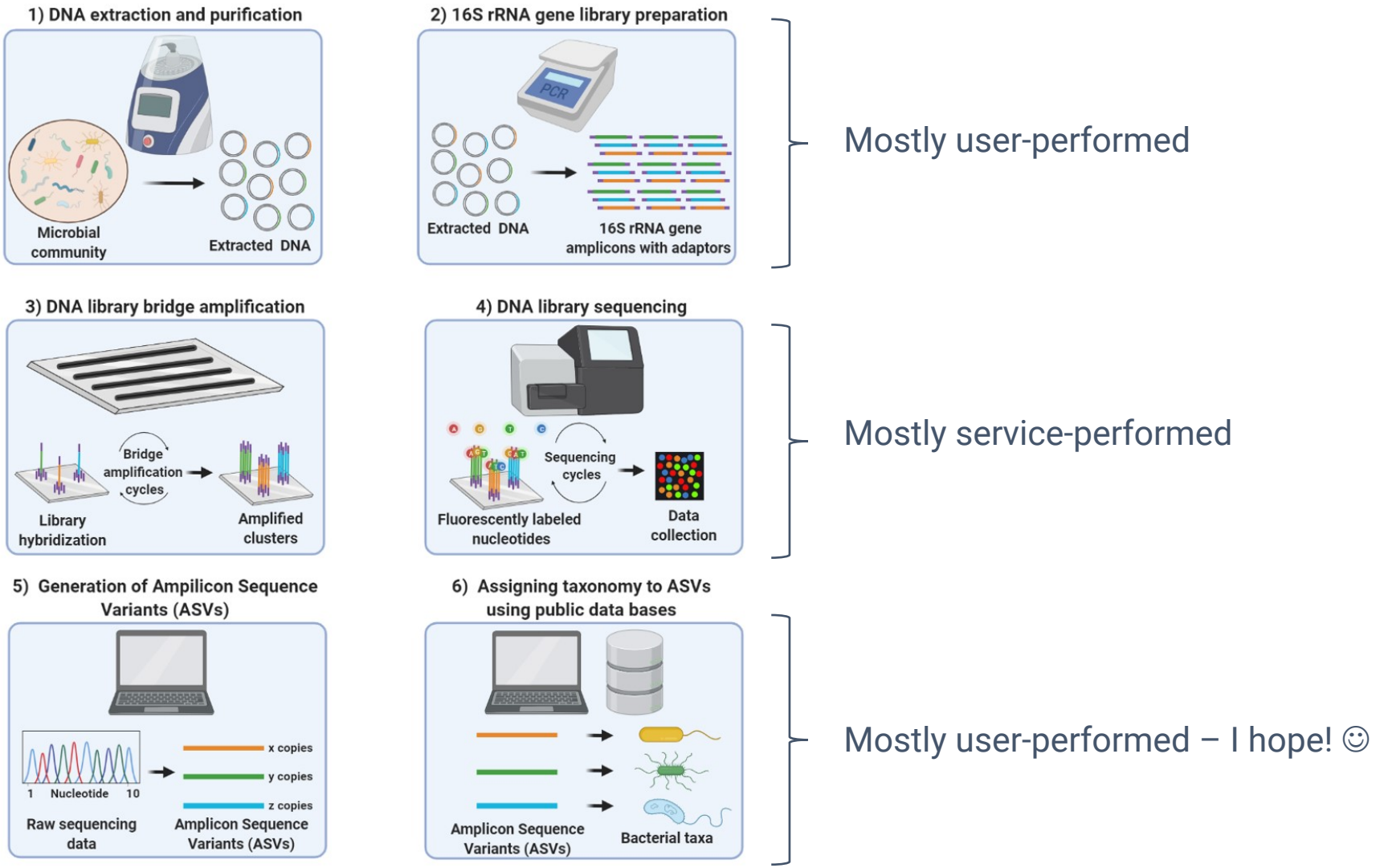


Diagram source: https://www.ufz.de/index.php?en=48523
(UFZ Helmoltz Center for Environmental Research)

16S rRNA gene sequencing – What degree of classification?

WHO'S THERE?

"The bacterial species problem"

Impossibility of using the traditional biological species concept (no sexual reproduction).

Prokaryotic "species" has been defined as:

- Strains with ≥ **70% DNA-DNA hybridization** and a difference in melting temperature of < 5C (Wayne et al., 1987).
- Same species if their **16S rRNA genes are ≥ 97% identical** (Stackebrandt and Goebel, 1994). Later reevaluated at 98.7%.
  - Observation: Different species with 99% similarity in the 16S rRNA gene may have very different physiologies (e.g. *Bacillus* species).

Still a "definition in progress"...

Circunventing the problem

Other terms: ecotype, phylotype, ribotype, oligotype, clade (different meaning), etc.

Operational Taxonomic Unit (OTU): usually defined as a cluster of organisms with **16S rRNA genes that are ≥ 97% similar** (threshold may vary). Two main methods: closed-reference and *de novo* clustering.

Limitations of the OTU-based approach (clustering):

- Arbitrary dissimilarity threshold.
- No consistency converting to taxonomic levels.
- Distances within a taxonomic group may not be evenly distributed.
- Dataset-dependent (de novo method).
- Clustering is computationally intensive.

Circunventing the problem

More recently: **Amplicon Sequence Variants (ASV).**

OPEN

The ISME Journal (2017) 11, 2639–2643
www.nature.com/ismej

PERSPECTIVE

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

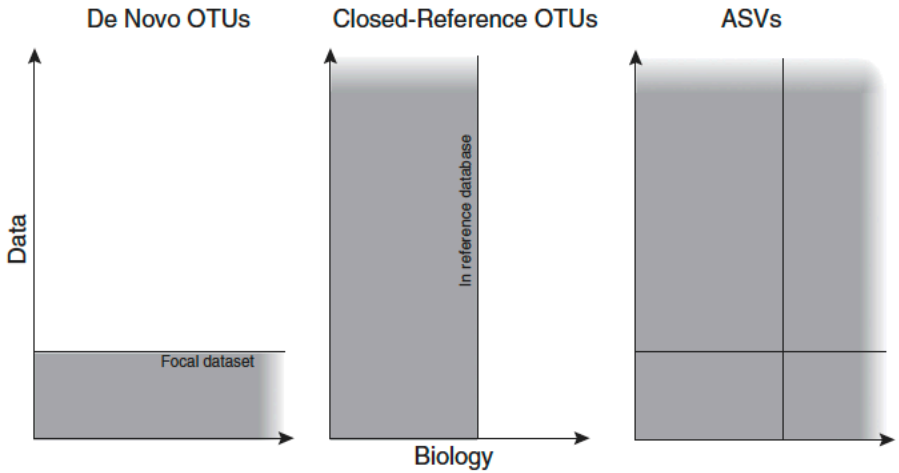
Benjamin J Callahan[1], Paul J McMurdie[2] and Susan P Holmes[3]
[1]*Department of Population Health and Pathobiology, NC State University, Raleigh NC, USA;* [2]*Whole Biome Inc, San Francisco CA, USA and* [3]*Department of Statistics, Stanford University, Stanford CA, USA*

ASV methods infer the biological sequences in the sample prior to the introduction of amplification and sequencing errors, and ***distinguish sequence variants differing by as little as one nucleotide****.* (In Callahan et al., 2017)

Circunventing the problem

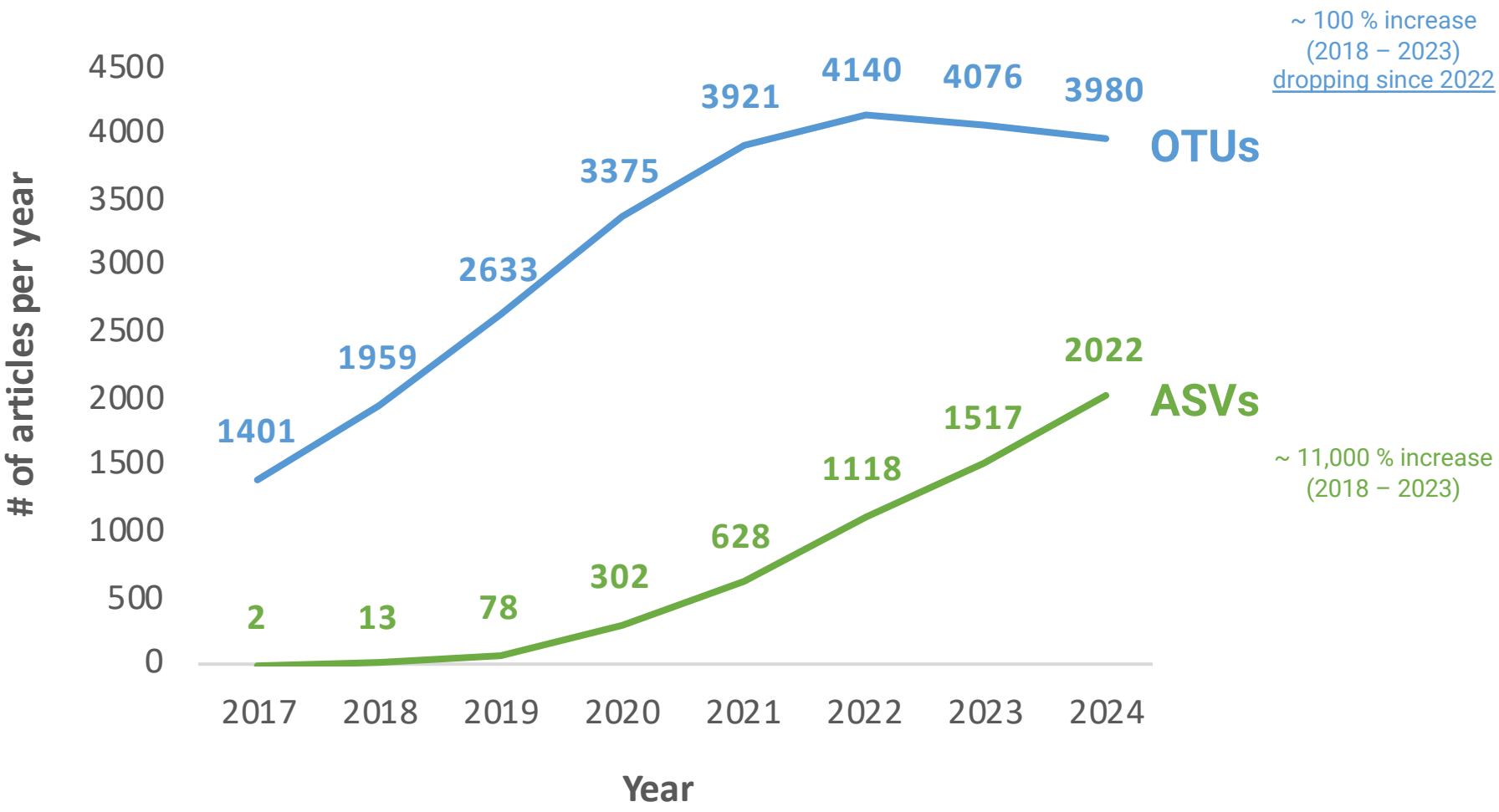
Main advantages of ASV vs. OTU:

- Higher resolution (obvious).
- Intrinsic biological meaning.
- Independent from reference database (compared to closed-reference OTU methods).
- Reusability (across different datasets – compared to *de novo* OTU methods).
- Reproducible in the future.



Callahan et al., 2017, The ISME Journal.

Current scientific use



Data source: Science Direct (Elsevier)

16S sequencing analysis workflow (from short reads)

Different softwares/pipelines/workflows have been used to process amplicon NGS data, such as Mothur, QIIME, and more recently DADA2 (R package), that applies the concept of "**amplicon sequence variants (ASVs)**".

Package 'dada2'

May 13, 2018

Type Package

Title Accurate, high-resolution sample inference from amplicon sequencing data

Description The dada2 package infers exact amplicon sequence variants (ASVs) from high-throughput amplicon sequencing data, replacing the coarser and less accurate OTU clustering approach. The dada2 pipeline takes as input demultiplexed fastq files, and outputs the sequence variants and their sample-wise abundances after removing substitution and chimera errors. Taxonomic classification is available via a native implementation of the RDP naive Bayesian classifier, and genus-species assignment by exact matching.

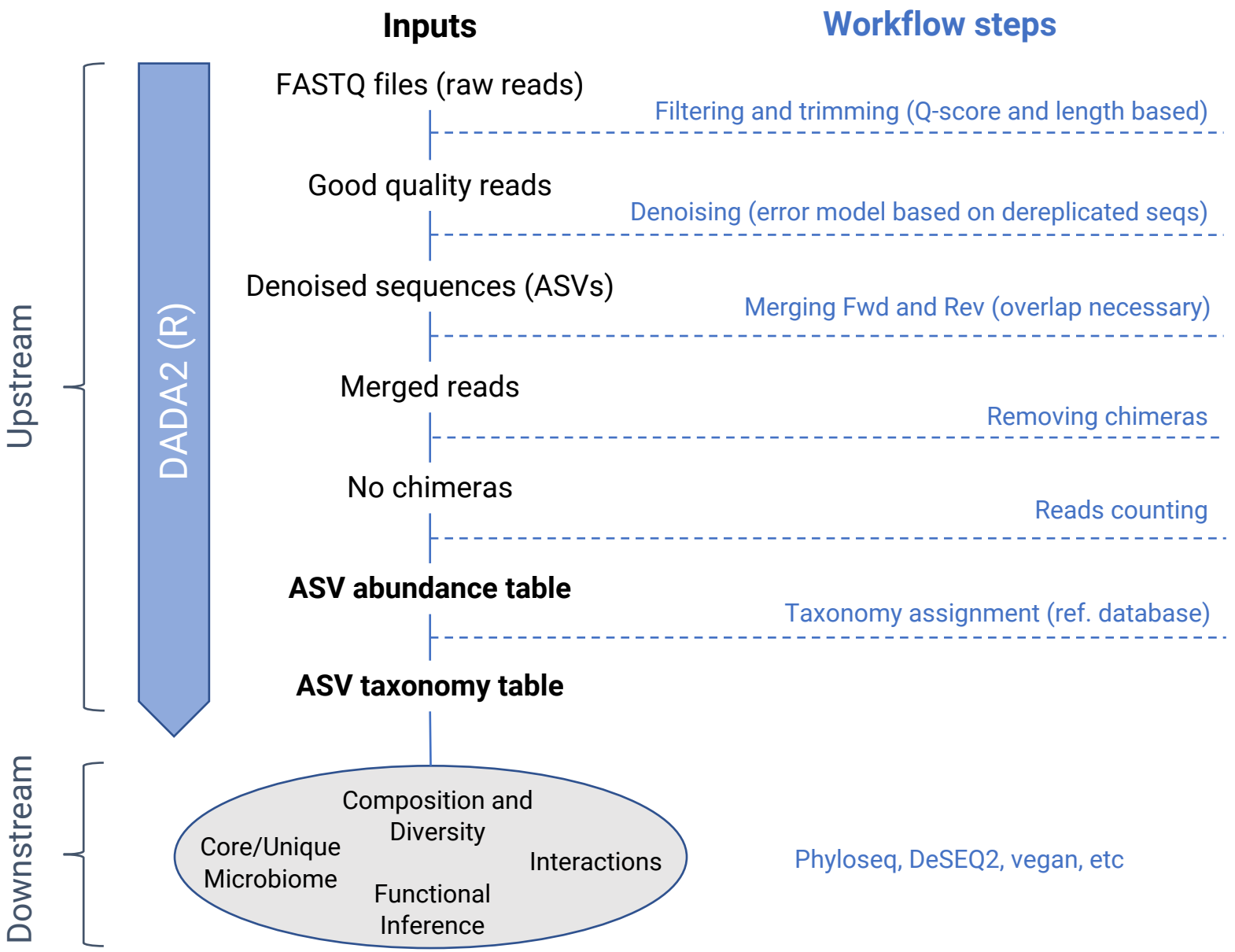
Version 1.9.0

Date 2018-1-30

Maintainer Benjamin Callahan <benjamin.j.callahan@gmail.com>

Author Benjamin Callahan <benjamin.j.callahan@gmail.com>, Paul McMurdie, Susan Holmes

16S sequencing analysis workflow (from short reads)



16S sequencing analysis workflow (from short reads)

WORKFLOW SUMMARY REPORT (UPSTREAM)

33 libraries	Input RAW sequences	Filtered + Denoised	Merged	non chimeras	Total Yield
Total	**5,734,065**	5,080,131	2,939,179	**1,972,593**	**34.4%**
Step Yield		88.6%	57.9%	67.1%	
Average/sample	173,760 ± 55,517	153,943 ± 49,754	89,066 ± 32,413	59,776 ± 20,666	

The balance between merging overlap and chimera removal is often sensitive (play with it!).

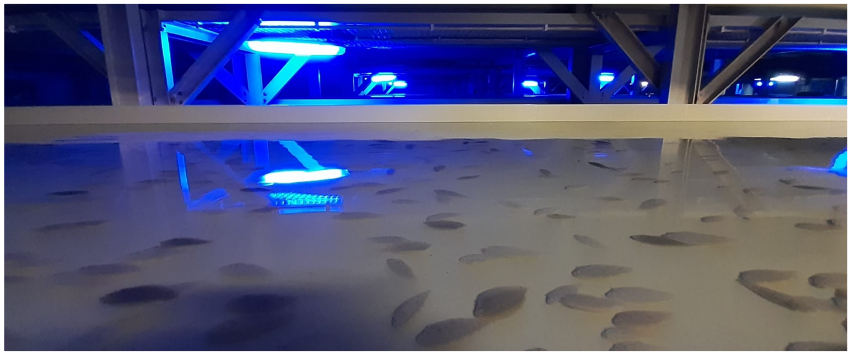
I like to work with a minimum of 30% total yield and 20,000 final number of high-quality sequences, but this is highly dependent on the study and research question. If some sample is under these criteria, I consider removing it from the downstream analysis.

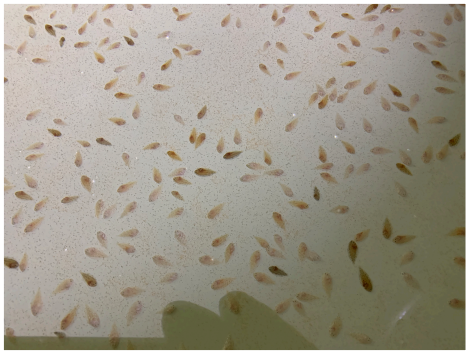
Questions?

16S sequencing analysis workflow (practical example)

Microbial Community Structure of Juvenile Sole (Solea senegalensis) Fed with Different Diets (manuscript in preparation)

1) Investigate the microbial community structure across different fish body parts
2) Investigate the effects of diet composition in the microbial community structure of the different matrices

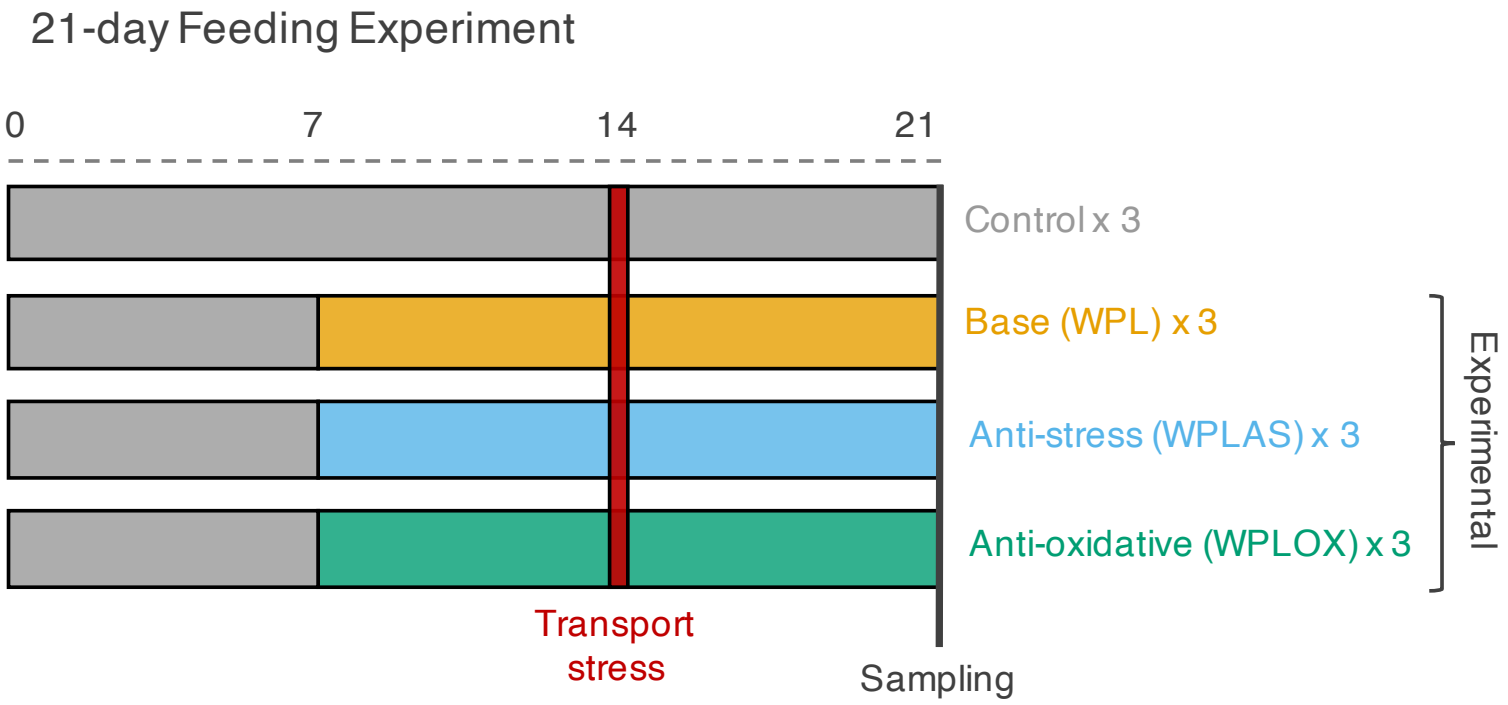




16S sequencing analysis workflow (practical example)

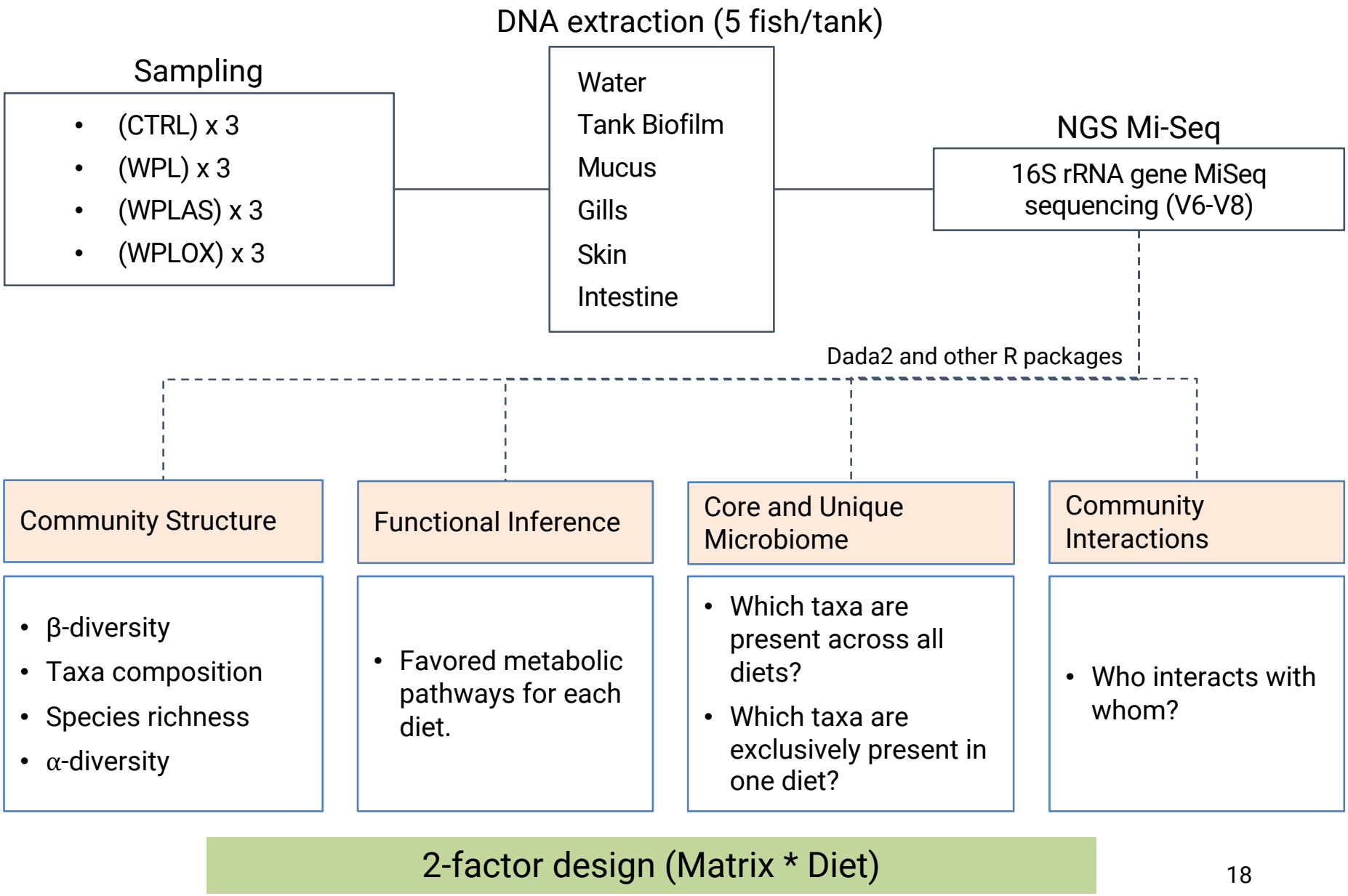
Microbial Community Structure of Juvenile Sole (Solea senegalensis) Fed with Different Diets (*manuscript in preparation*)

Experimental Design



12 tanks (4 diets x 3 replicates)

16S sequencing analysis workflow (practical example)



16S sequencing analysis workflow (practical example)

DADA2 (just the intestines for the workshop, n= 12)

Before start (checklist)

1. Check service provider report (raw, filtered, merged, etc.). Is the initial number of sequences good enough? Are adapters and primers removed?
2. Sequence files (chose which ones to use)
3. Metadata table (“map object”)
4. Get Log text file
5. Your taxonomy reference database (Silva / GTDB / Greengenes)
6. Start a new R script (or open a trusted one)
7. Don’t just trust (blindly) someone else’s script
8. Start a new RStudio session (set working directory)

16S sequencing analysis workflow (practical example)

DADA2 run (just the intestines, n= 12)