Metabarcoding Workshop

Anna Heintz-Buschart

June 2022

University of Amsterdam
Life Sciences

“Metabarcoding”

- Coupling high-throughput sequencing with our ability to associate sequences from eDNA with a taxonomic name is called “eDNA metabarcoding” (Deiner *et al*. Mol Ecol. 2017)

- “massively parallel tag sequencing strategy” (Sogin *et al*. PNAS 2006)

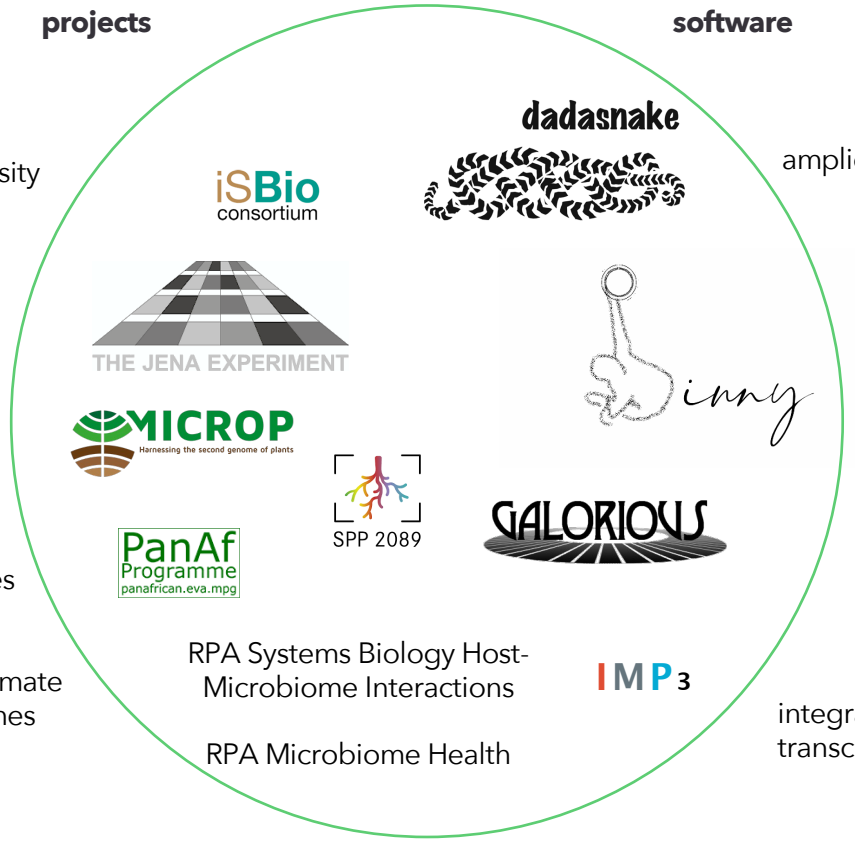
- descendant of DGGE profiles (Ferris *et al*. Appl Environ Microbiol. 1996)

Overview of today

- A look at the aims
- Overview of the method
- Limitations – from sample to sequencing data
- How do we try to deal with these limitations?
- Which problems persist?

- dadasnake – aims and realization
- dadasnake: options in detail
- Q&A

About me



2008

MSc Biology (Microbiology, Botany, Molecular & Cell Biology)

PhD: Fungal human pathogen
- compound screening, mode-of action
- gene expression analysis

2011

Postdoc: Gene regulatory network modelling

2012

Postdoc: Integrated meta-omics
- human microbiome, wastewater treatment
- metagenomics, metatranscriptomics, metaproteomics
- lab automation
- bioinformatics pipelines

2017

Metagenomics support:
- biodiversity
- soil, plants, animal microbiomes
- bioinformatics pipelines
- data integration

2021

Assistant Prof Microbial Metagenomics
- meta-omics integration
- human and plant microbiomes

a.u.s.heintzbuschart@uva.nl

SP C2.205

github.com/a-h-b

twitter.com/_a_h_b_

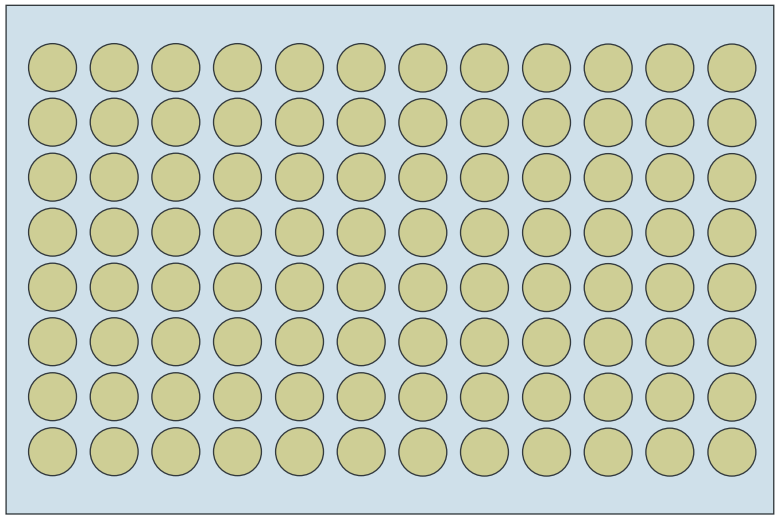
Questions

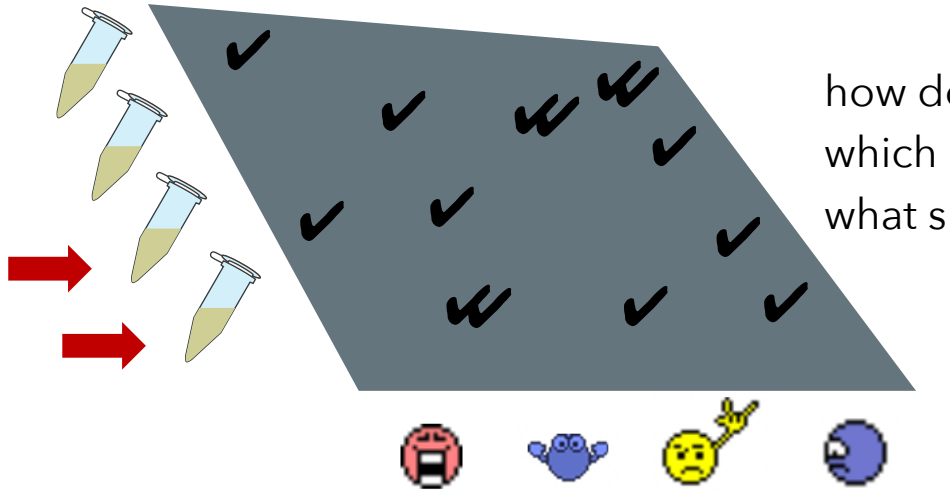
mixed community

what is in my sample?

who is in my community?

what is in my samples?





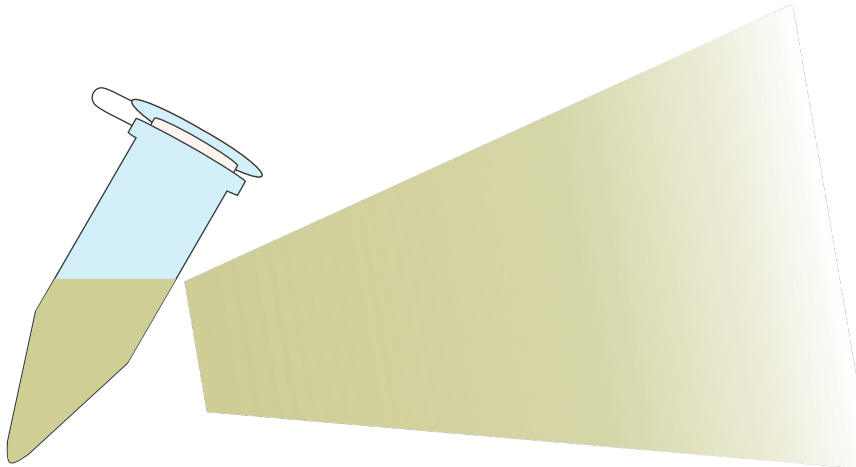
how do my samples compare?
which of my samples are similar?
what shapes my samples?

when are they there?

who is there often/in high numbers?
who is there with whom?

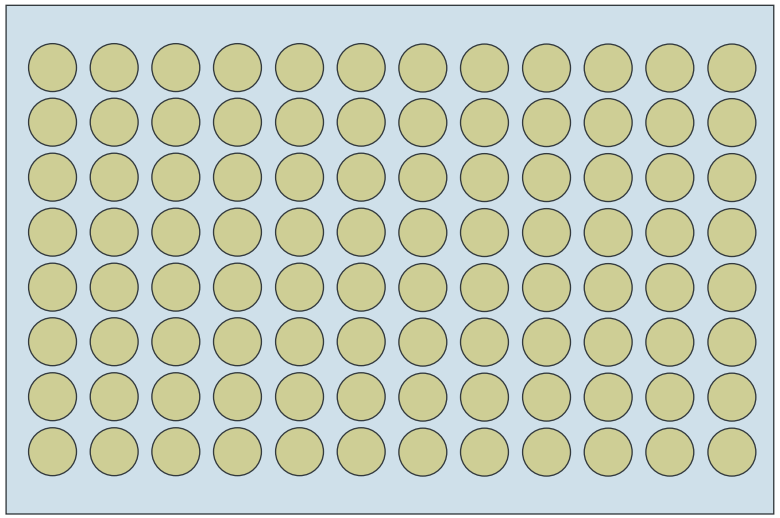
Questions

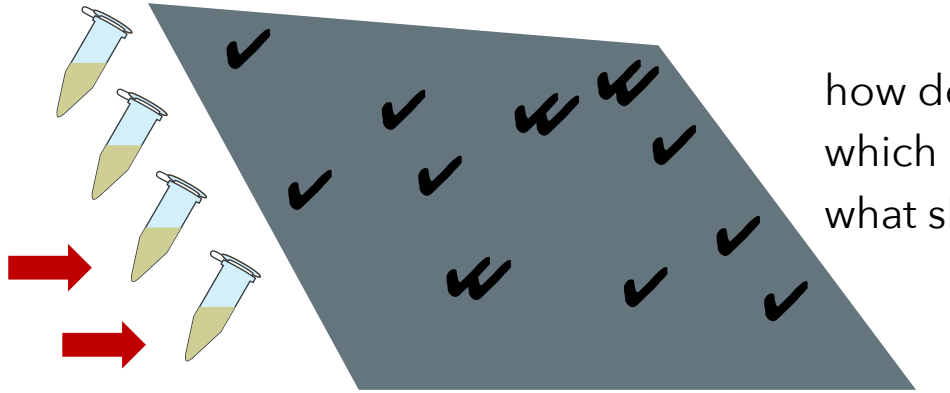
what is in my sample?



mixed community

what is in my samples?





how do my samples compare?
which of my samples are similar?
what shapes my samples?

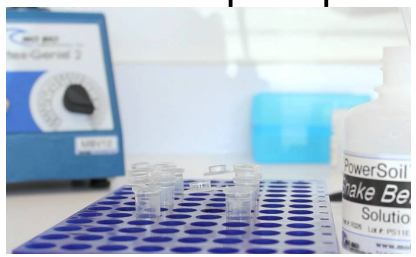
when are they there?

who is there often/in high numbers?
who is there with whom?

Measuring microbiomes: DNA based methods

sampling + preservation

sample processing



Measuring microbiomes: marker genes

- classical target: 16S rRNA gene

- pre-requisites:
  - conserved regions for primers to bind
  - variable regions with suitable phylogenetic resolution
  - similar mutation rates across all measurable taxa
  - no horizontal gene transfer
  - suitable length

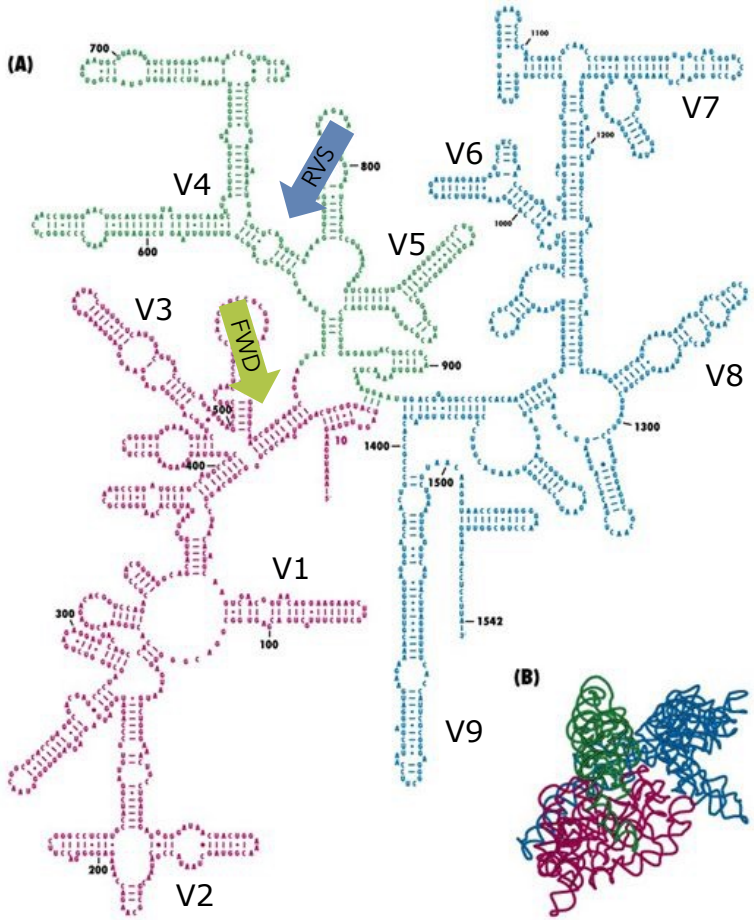


Figure 30-14
Biochemistry, Sixth Edition
© 2007 W.H. Freeman and Company

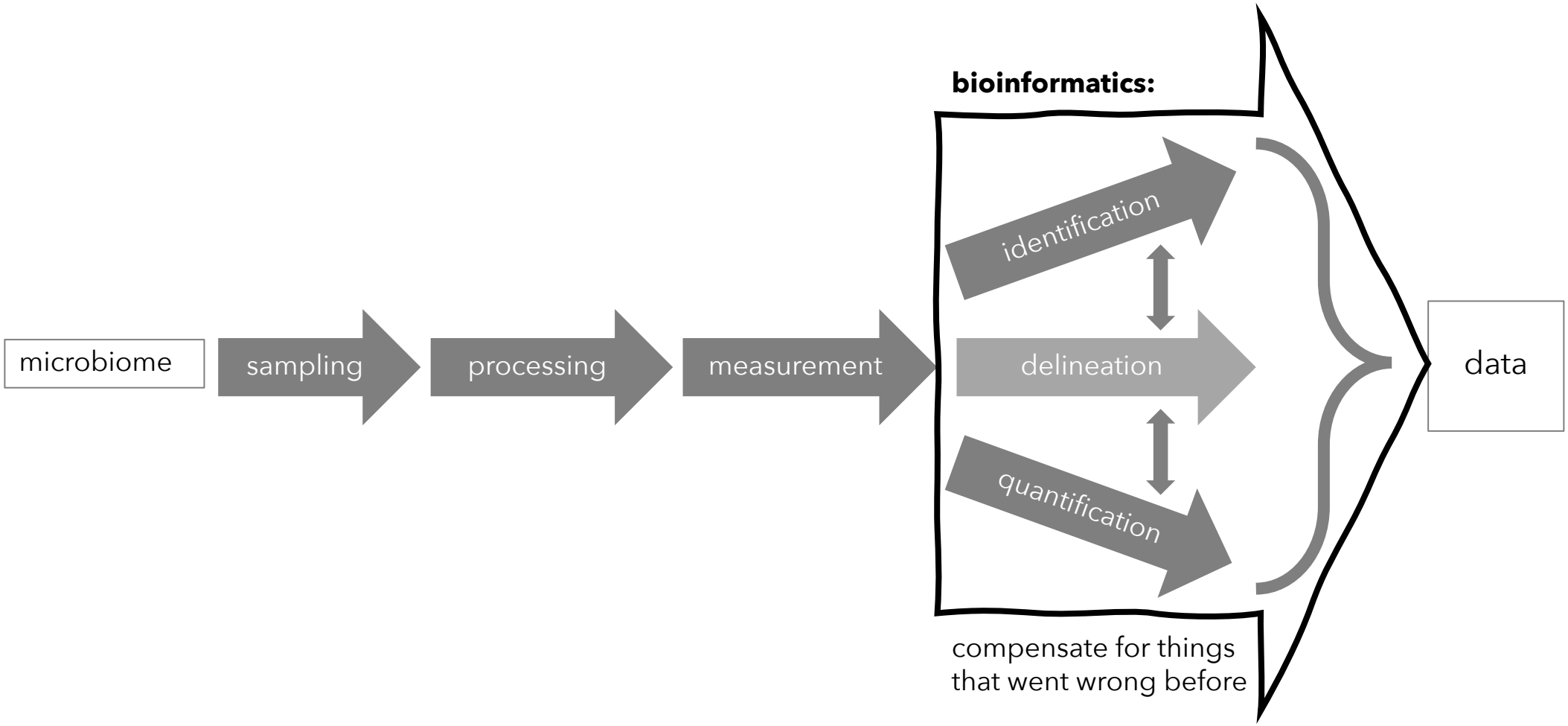
Measuring microbiomes: marker genes

- classical target: 16S rRNA gene

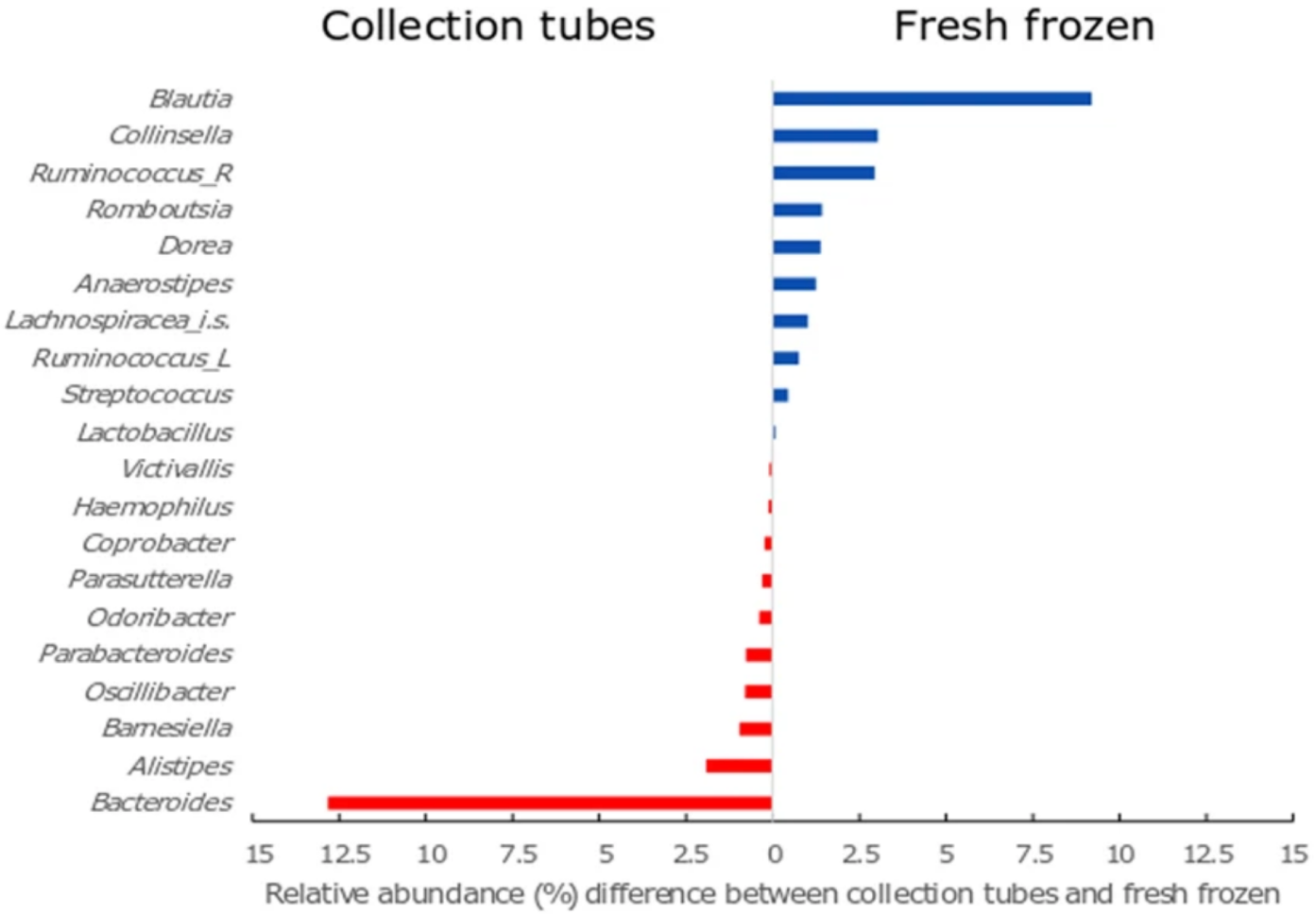
- pre-requisites:
  - conserved regions for primers to bind
  - variable regions with suitable phylogenetic resolution
  - similar mutation rates across all measurable taxa
  - no horizontal gene transfer
  - suitable length

- other targets:
  - 18S rRNA genes
  - internal transcribed spacers (ITS)
  - 28S rRNA genes
  - 12S rRNA genes
  - cytochrome *c* oxidase subunit 1 gene (COI)
  - RuBisCO large chain (*rbc*L)
  - $tRNA^{Leu}$ intron (*trn*L)
  - RNA polymerase (*rpo*B)

Omics paradigm

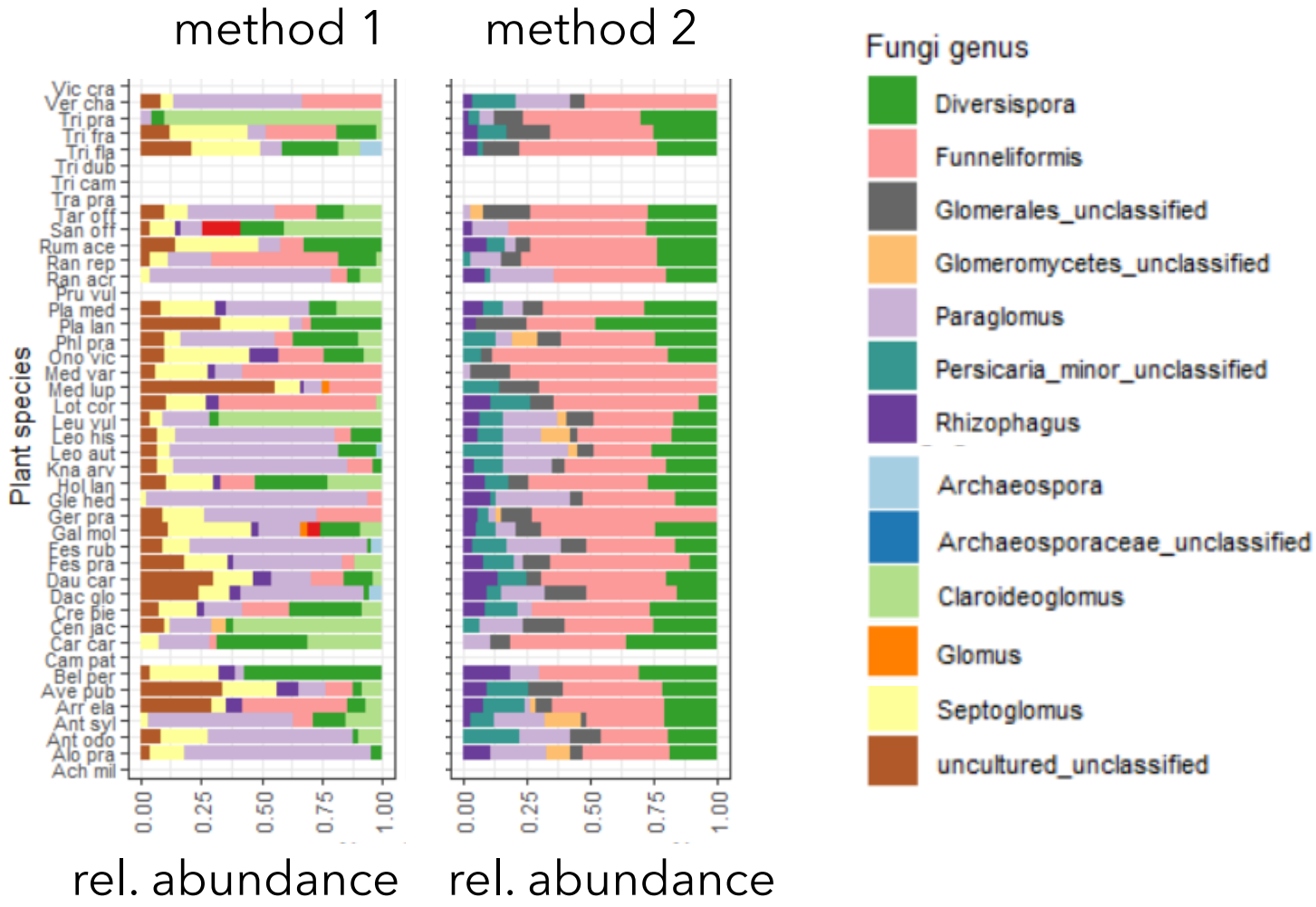


What could go wrong? Sample storage/processing



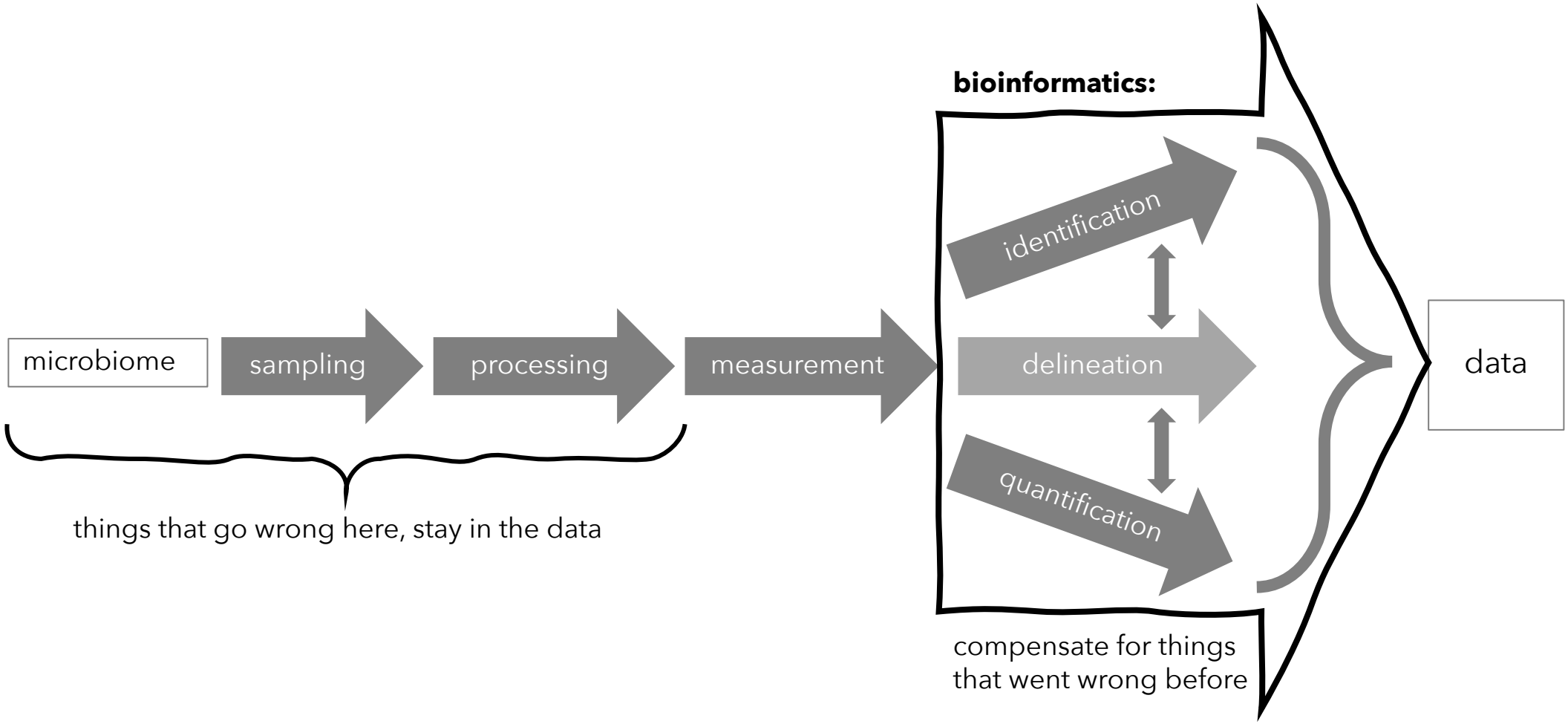
Watson *et al.*, 2019, Scientific Reports

What could go wrong? Amplification

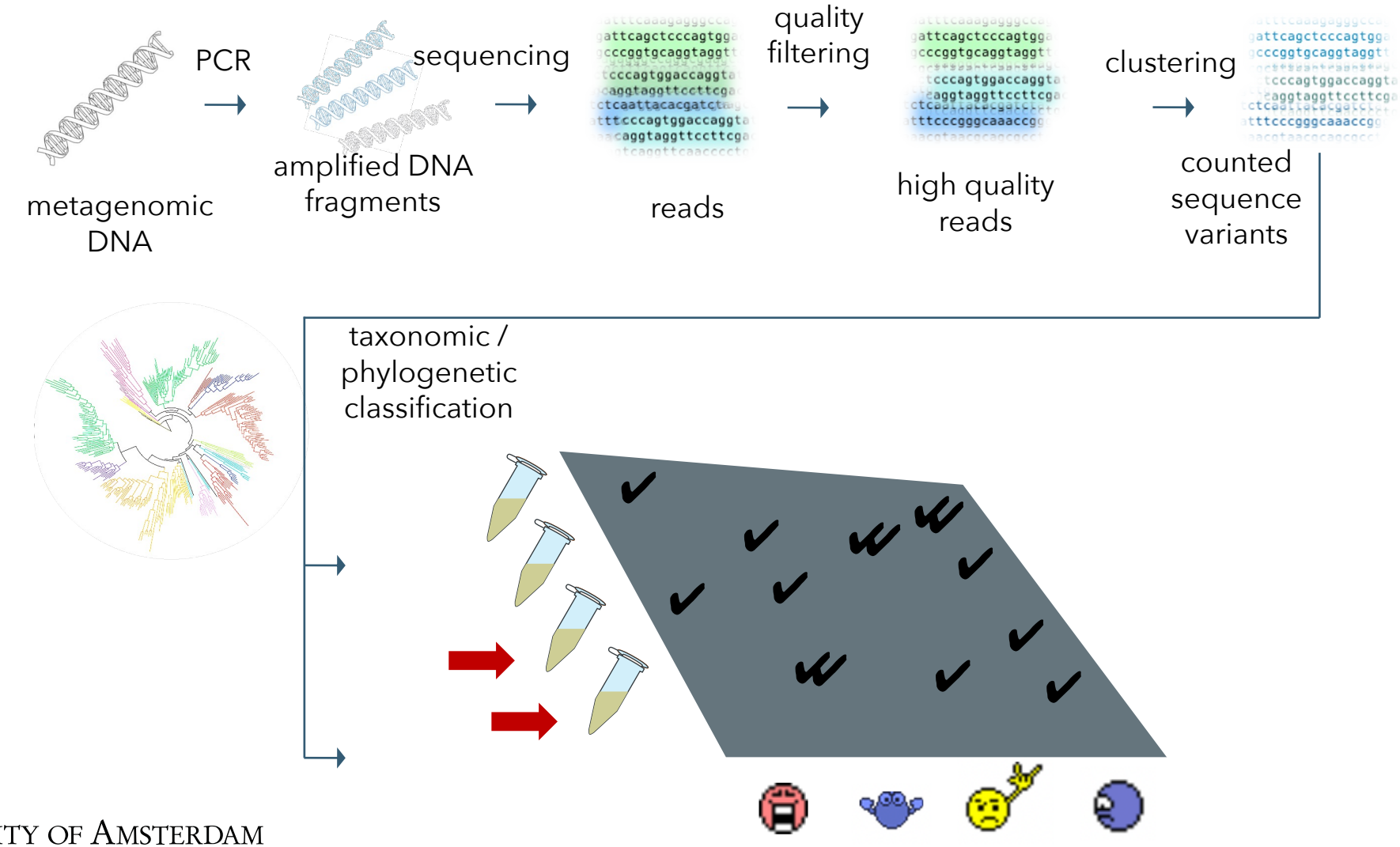


Cynthia Albracht *et al.*, unpublished

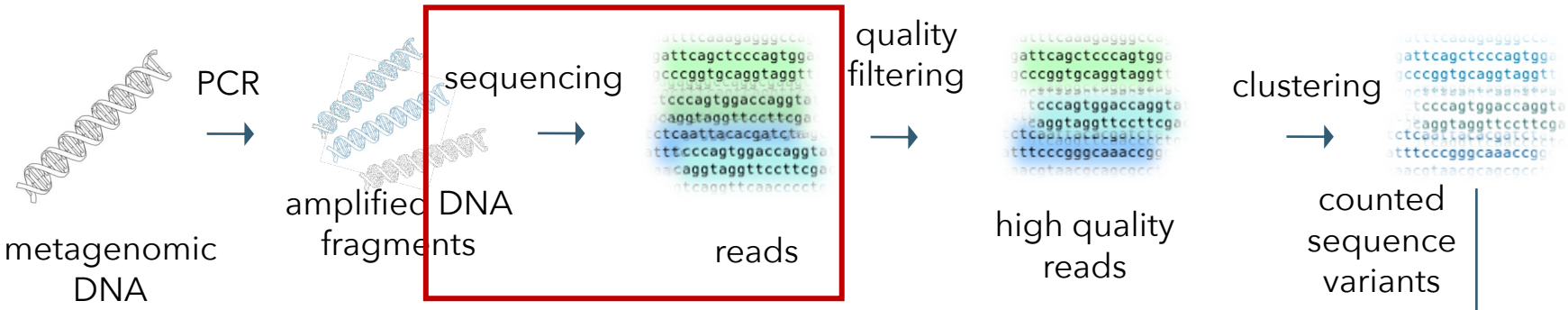
Omics paradigm



Metabarcoding workflow



Metabarcoding workflow



taxonomic / phylogenetic classification

classified sequence variants

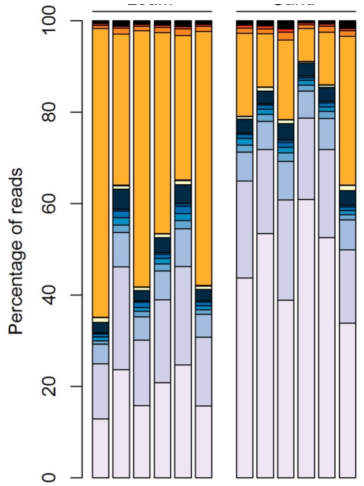
taxonomic profiles of microbial communities

alpha-diversity of microbial communities

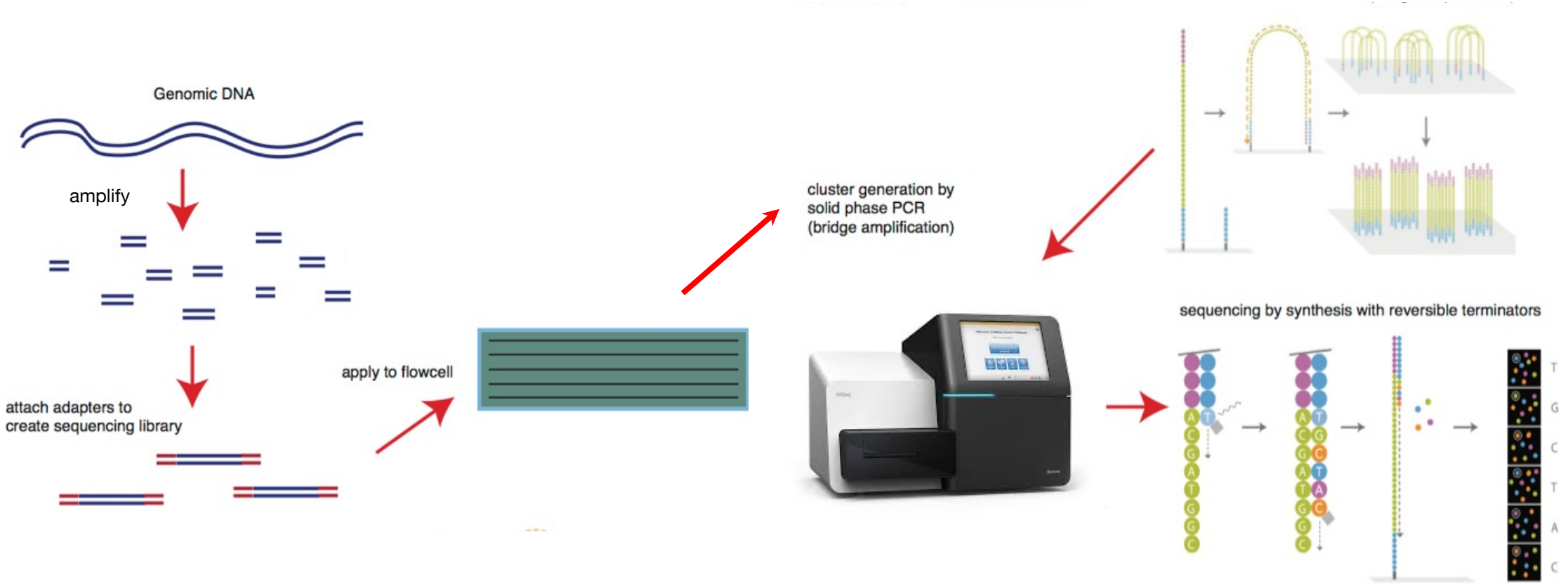
functional prediction

annotated sequence variants

statistical analyses

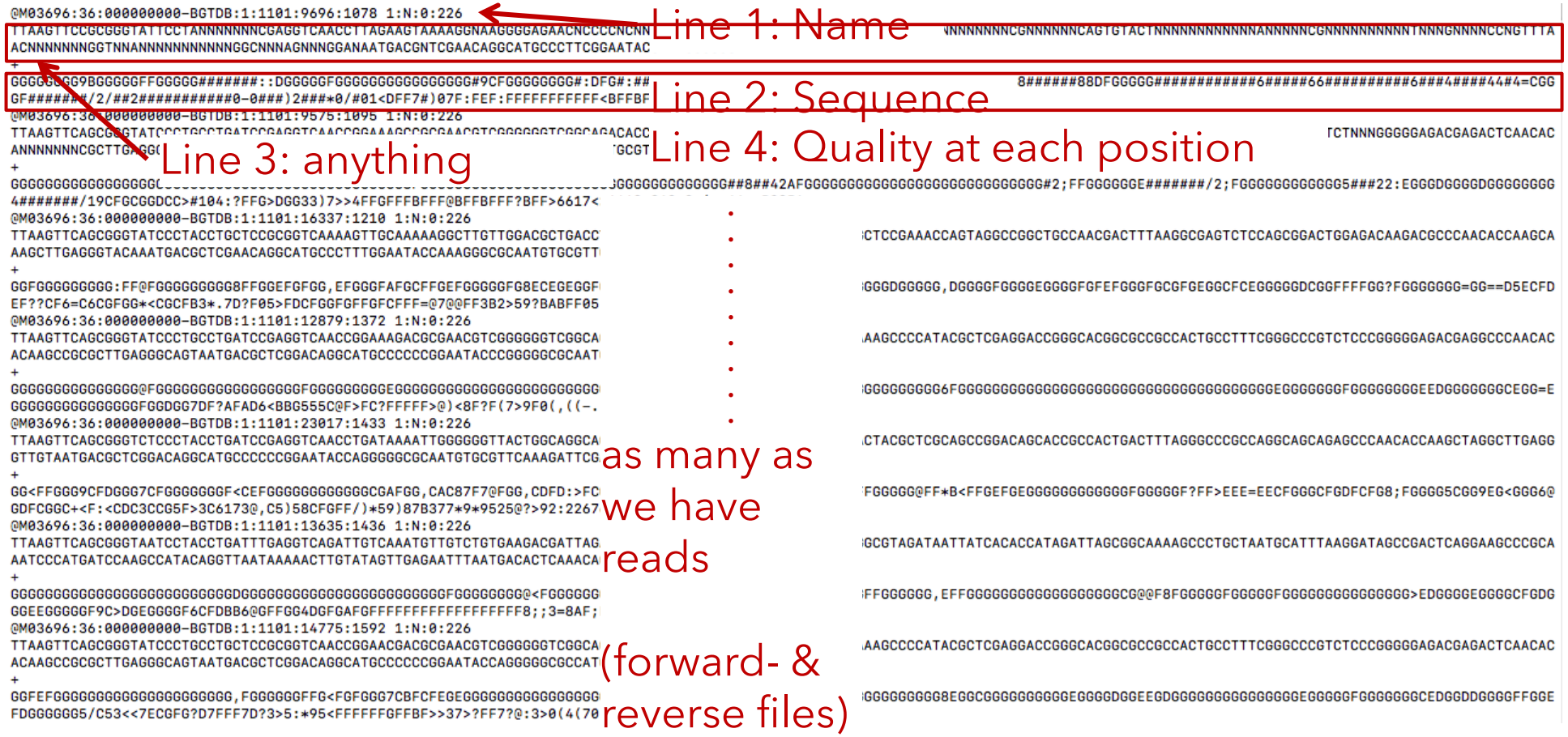


Measuring microbiomes: sequencing

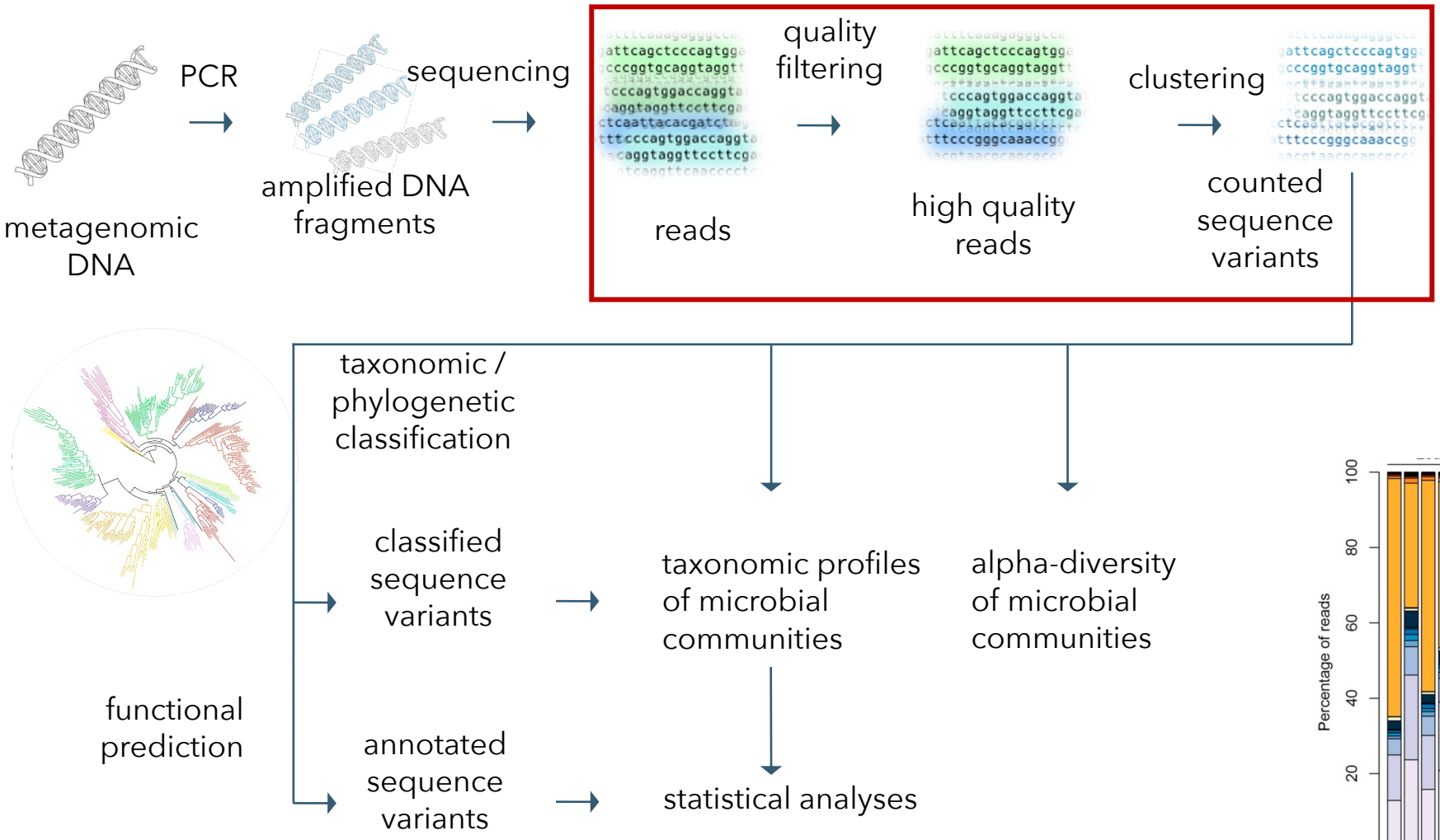


adapted from https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/

What does sequencing data look like?



Metabarcoding workflow



What's "wrong" with sequencing data?

- there are amplification errors and sequencing errors

➢sequencers recognize some sequencing errors and give quality scores

- quality can affect:
  - number of usable sequences
  - trade-off between resolution and misinterpretation of errors as real sequences

What's "wrong" with sequencing data?

- everything is measured at once
- off-target sequences

➢samples are marked ("indexed")
➢primer sequences are recognizable
➢sequences carry phylogenetic signal

- remove dubious/un-informative sequences
- interpret numbers with caution

What's "wrong" with sequencing data?

- numbers of reads per sample have no meaning
- there are no intensities/concentrations

➢reads can (must) be counted

- interpret numbers with caution
- proportions can be misleading

What's "wrong" with sequencing data?

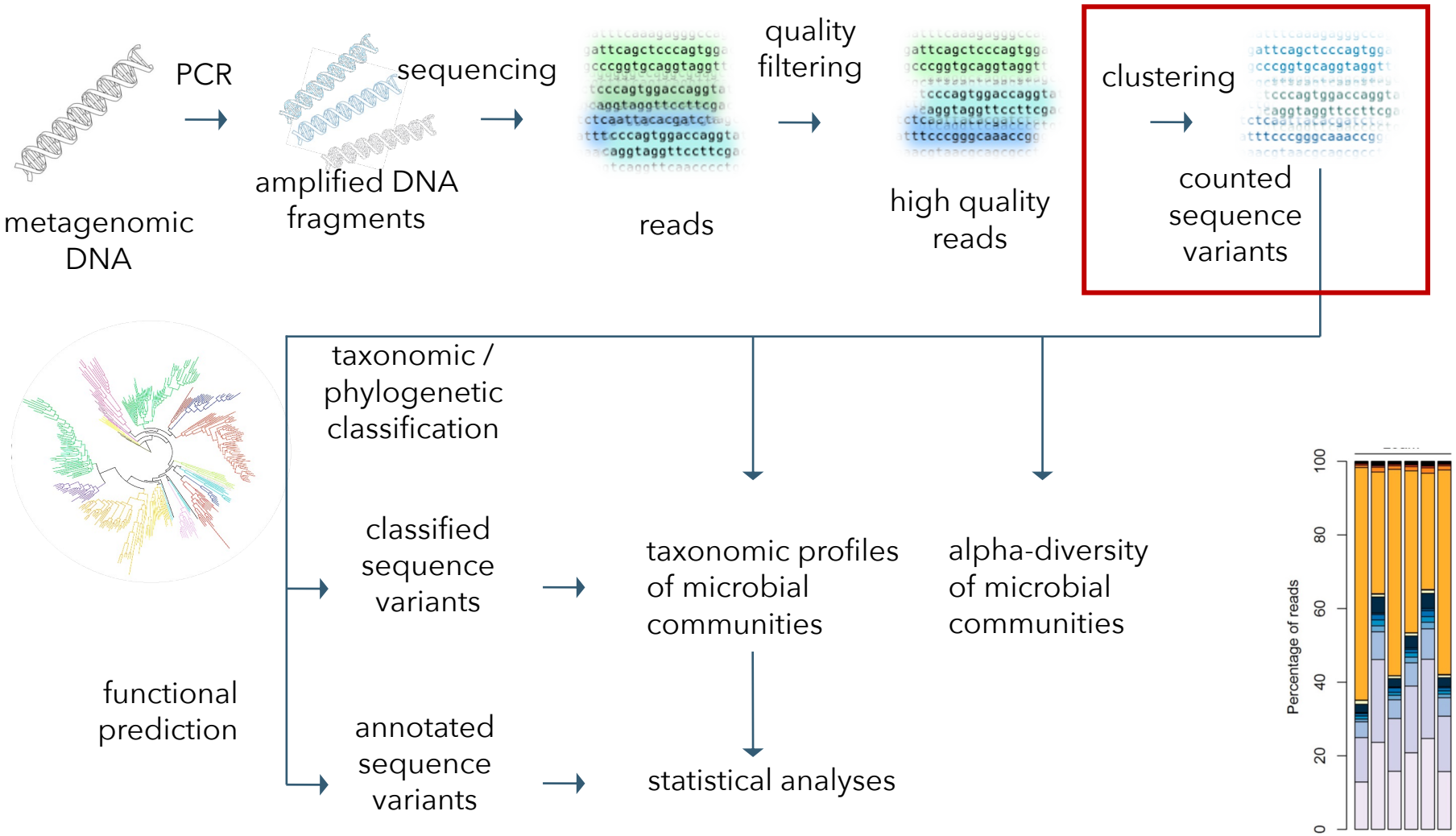
- amplicon length ≠ read length

➢overlaps are recognizable, mismatches are informative

- idiosyncrasies of formats and technologies

➢you need to know how your data was created

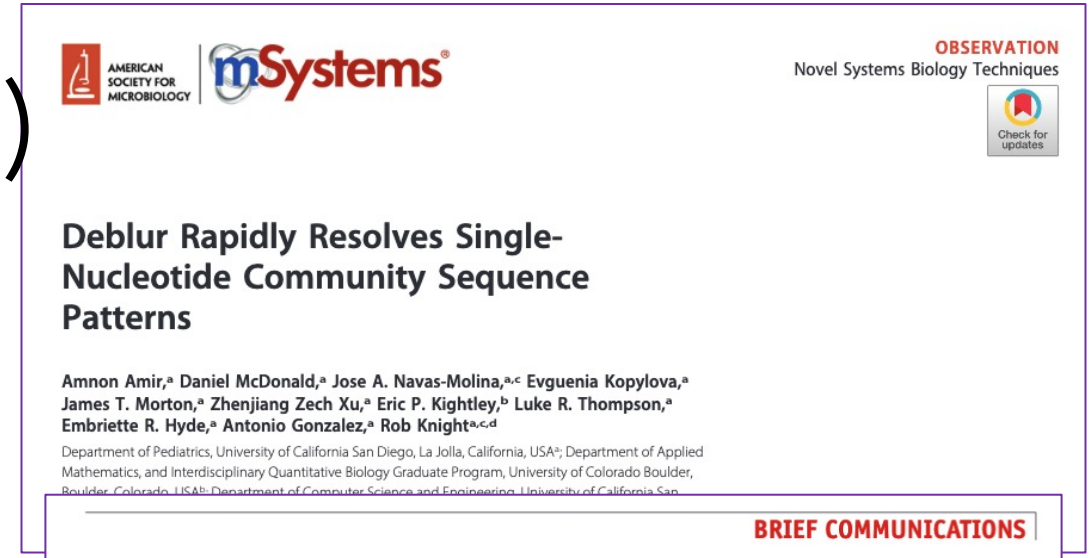
Metabarcoding workflow

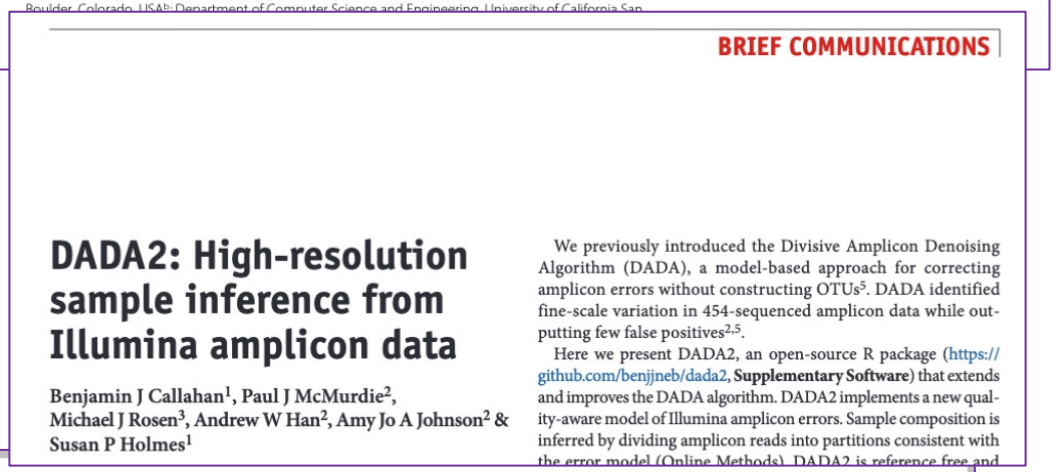


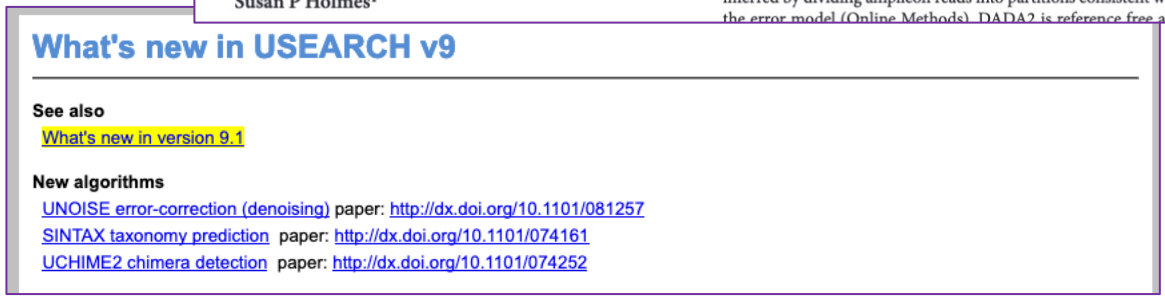
ASVs (aka. ESVs, zOTUs)

- **A**mplicon **S**equence **V**ariants
  = **E**xact **S**equence **V**ariants
  = **z**ero-radius **OTU**s

- current generation of computational tools (appeared in ~2017)
  - deblur
  - **DADA2**
  - unoise



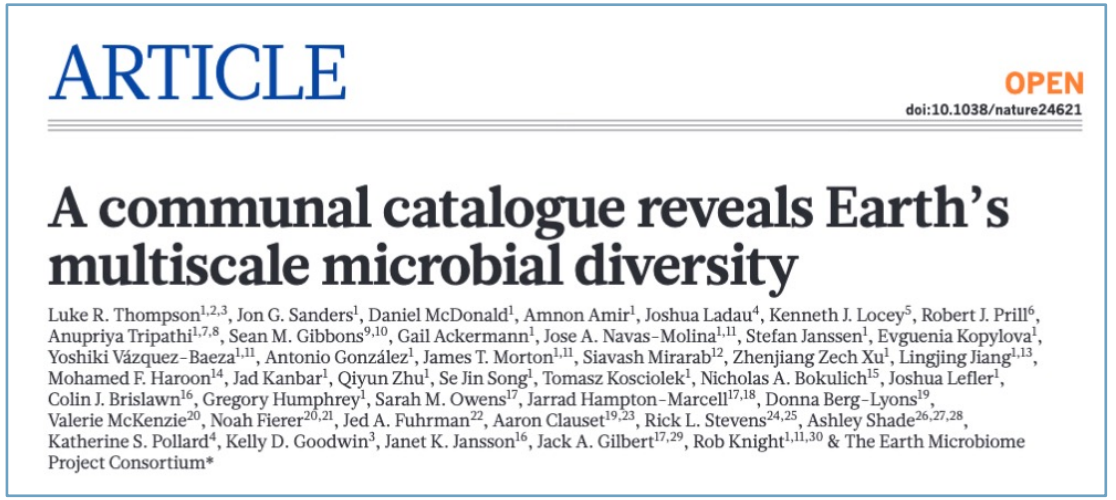




ASVs (aka. ESVs, zOTUs)

- **A**mplicon **S**equence **V**ariants
  = **E**xact **S**equence **V**ariants
  = **z**ero-radius **OTU**s

- current generation of computational tools (appeared in ~2017)
  - deblur
  - **DADA2**
  - unoise



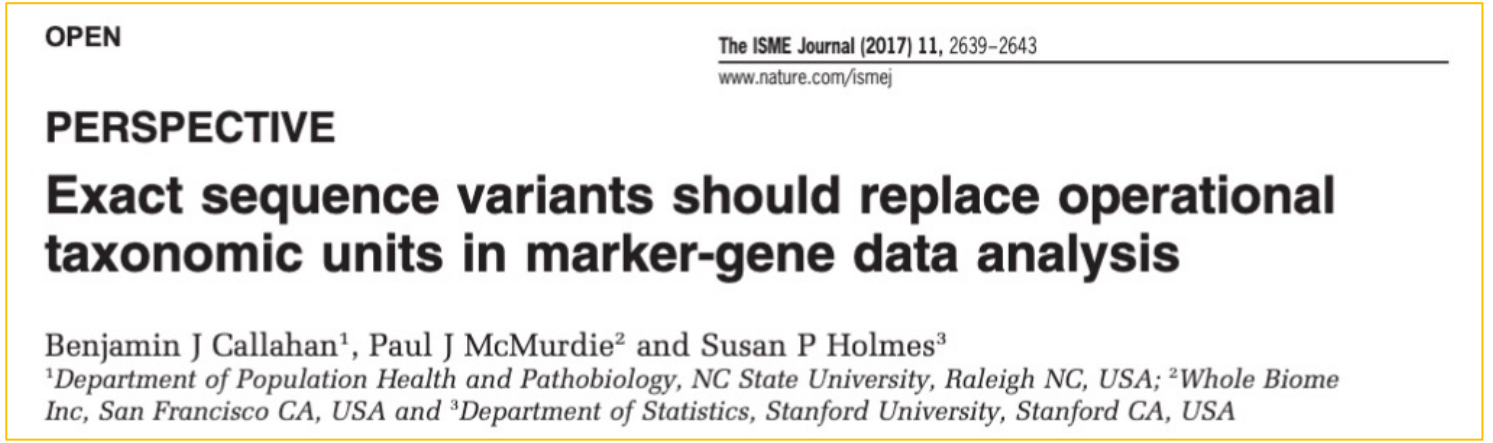
ARTICLE

OPEN

doi:10.1038/nature24621

A communal catalogue reveals Earth's multiscale microbial diversity

Luke R. Thompson[1,2,3], Jon G. Sanders[1], Daniel McDonald[1], Amnon Amir[1], Joshua Ladau[4], Kenneth J. Locey[5], Robert J. Prill[6], Anupriya Tripathi[1,7,8], Sean M. Gibbons[9,10], Gail Ackermann[1], Jose A. Navas-Molina[1,11], Stefan Janssen[1], Evguenia Kopylova[1], Yoshiki Vázquez-Baeza[1,11], Antonio González[1], James T. Morton[1,11], Siavash Mirarab[12], Zhenjiang Zech Xu[1], Lingjing Jiang[1,13], Mohamed F. Haroon[14], Jad Kanbar[1], Qiyun Zhu[1], Se Jin Song[1], Tomasz Kosciolek[1], Nicholas A. Bokulich[15], Joshua Lefler[1], Colin J. Brislawn[16], Gregory Humphrey[1], Sarah M. Owens[17], Jarrad Hampton-Marcell[17,18], Donna Berg-Lyons[19], Valerie McKenzie[20], Noah Fierer[20,21], Jed A. Fuhrman[22], Aaron Clauset[19,23], Rick L. Stevens[24,25], Ashley Shade[26,27,28], Katherine S. Pollard[4], Kelly D. Goodwin[3], Janet K. Jansson[16], Jack A. Gilbert[17,29], Rob Knight[1,11,30] & The Earth Microbiome Project Consortium*



OPEN

The ISME Journal (2017) 11, 2639–2643

www.nature.com/ismej

PERSPECTIVE

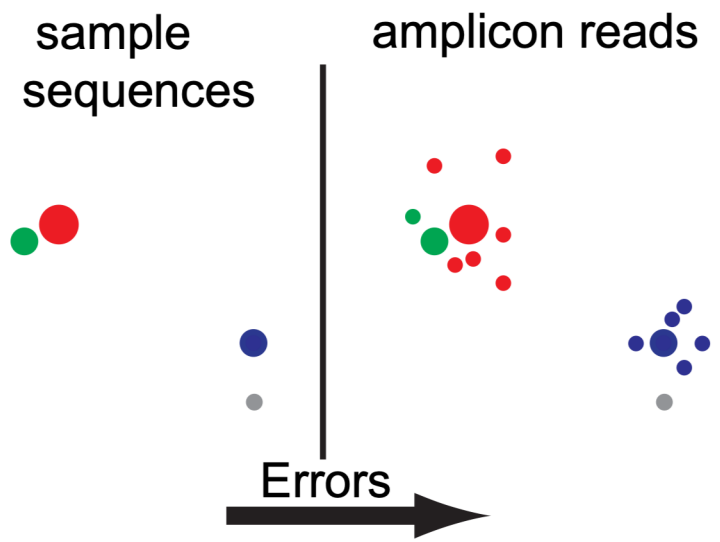
Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

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

ASVs (aka. ESVs, zOTUs)

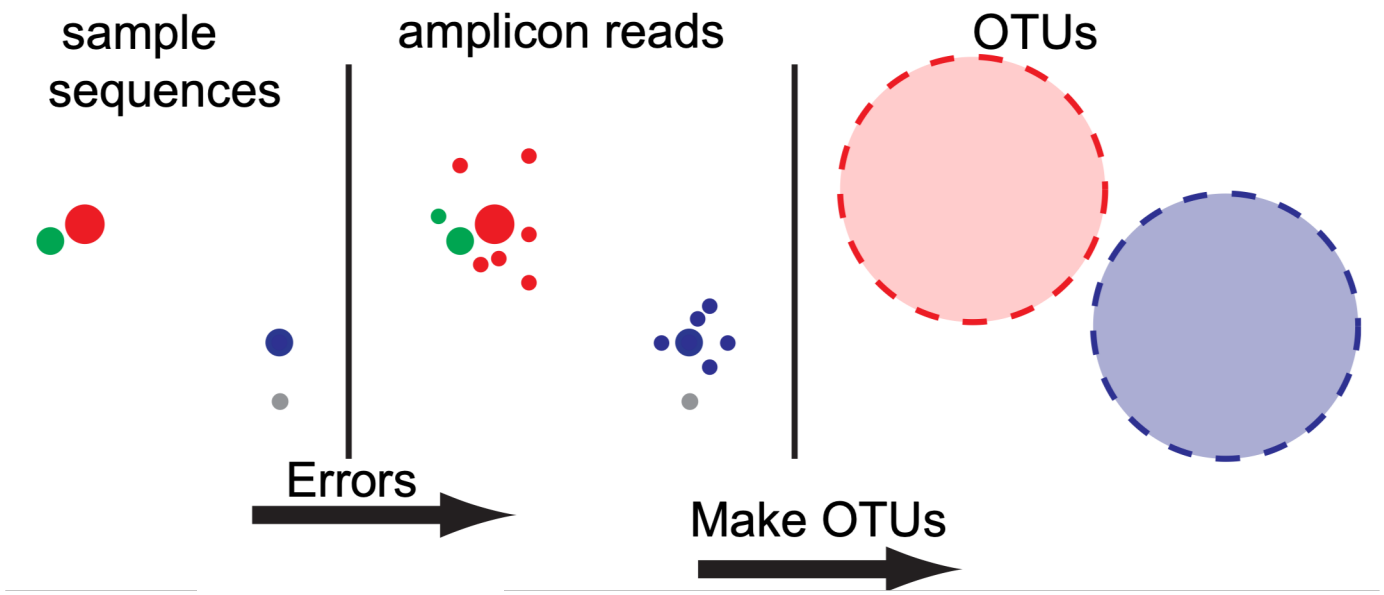
- **A**mplicon **S**equence **V**ariants
  = **E**xact **S**equence **V**ariants
  = **z**ero-radius **OTU**s



© B. Callahan (DADA2 documentation)

ASVs (aka. ESVs, zOTUs)

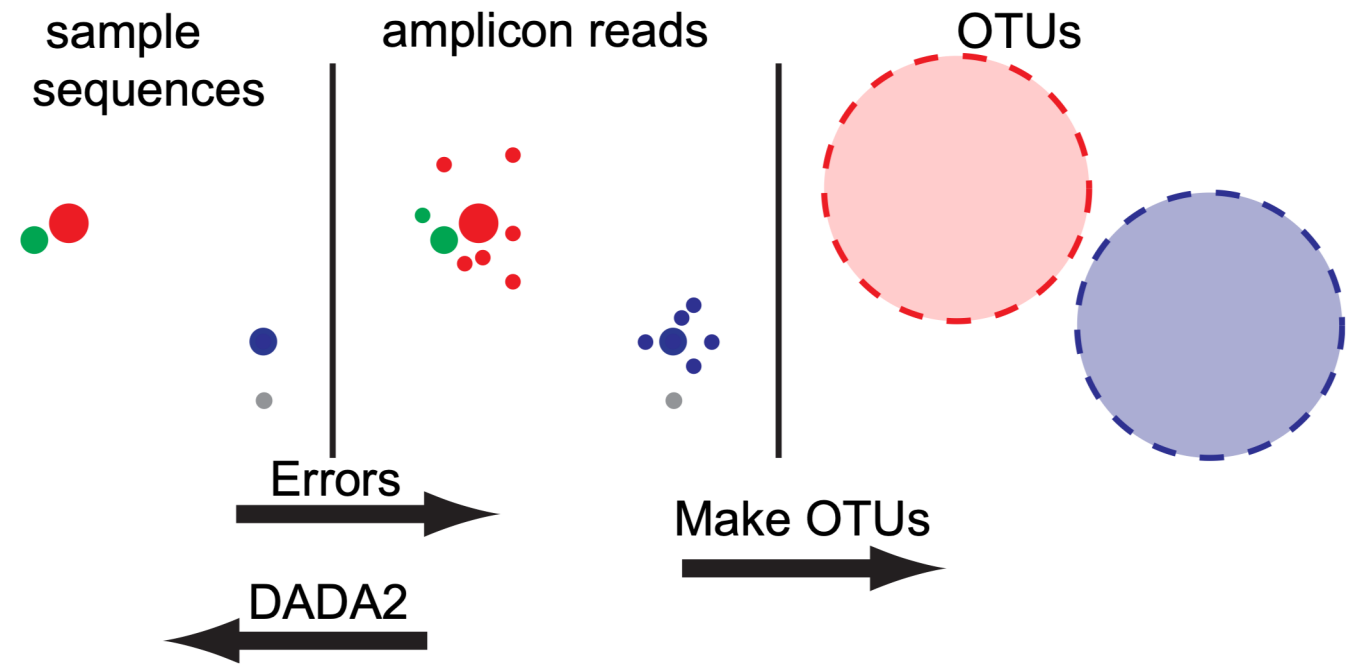
- **A**mplicon **S**equence **V**ariants
  = **E**xact **S**equence **V**ariants
  = **z**ero-radius **OTU**s



© B. Callahan (DADA2 documentation)

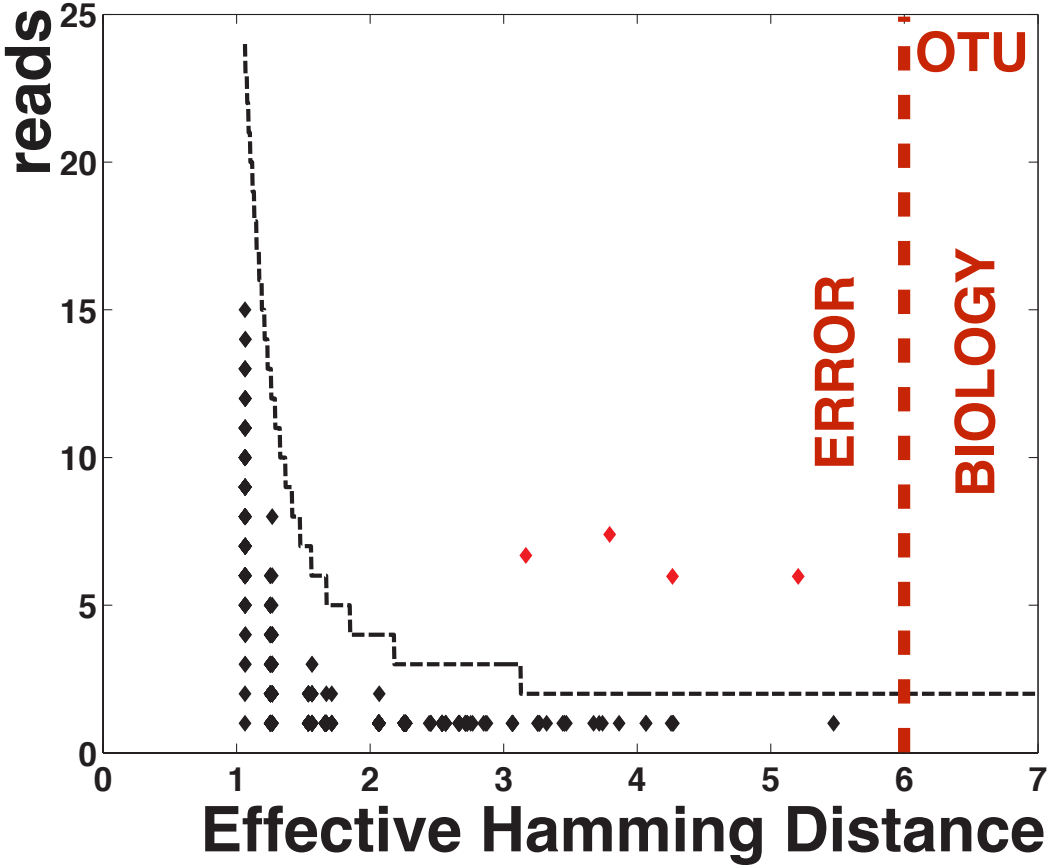
ASVs (aka. ESVs, zOTUs)

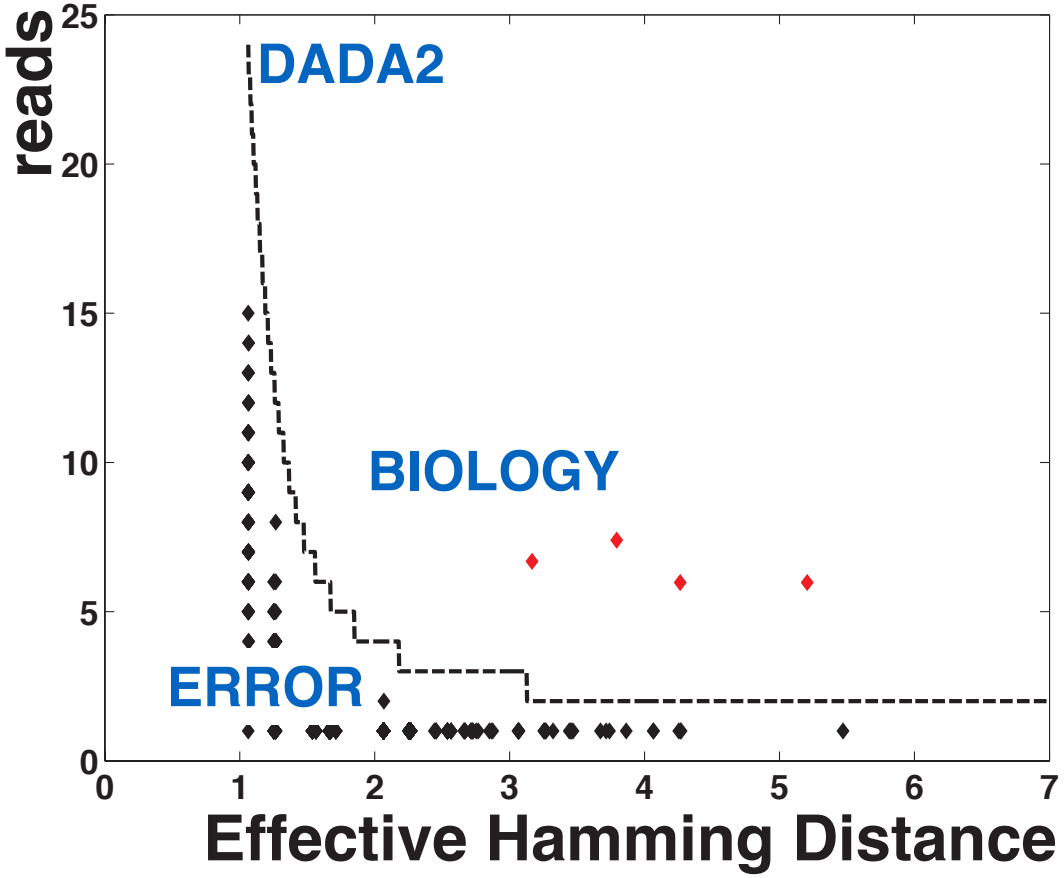
- **A**mplicon **S**equence **V**ariants
  = **E**xact **S**equence **V**ariants
  = **z**ero-radius **OTU**s



© B. Callahan (DADA2 documentation)

ASVs (aka. ESVs, zOTUs)





© B. Callahan (DADA2 documentation)

ASVs (aka. ESVs, zOTUs)

- make use of quality information

```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                    |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

Error rates depend on....

- Substitution (eg. A->C)
- Quality score (eg. Q=30)
- Batch effect (eg. run)

© B. Callahan (DADA2 documentation)

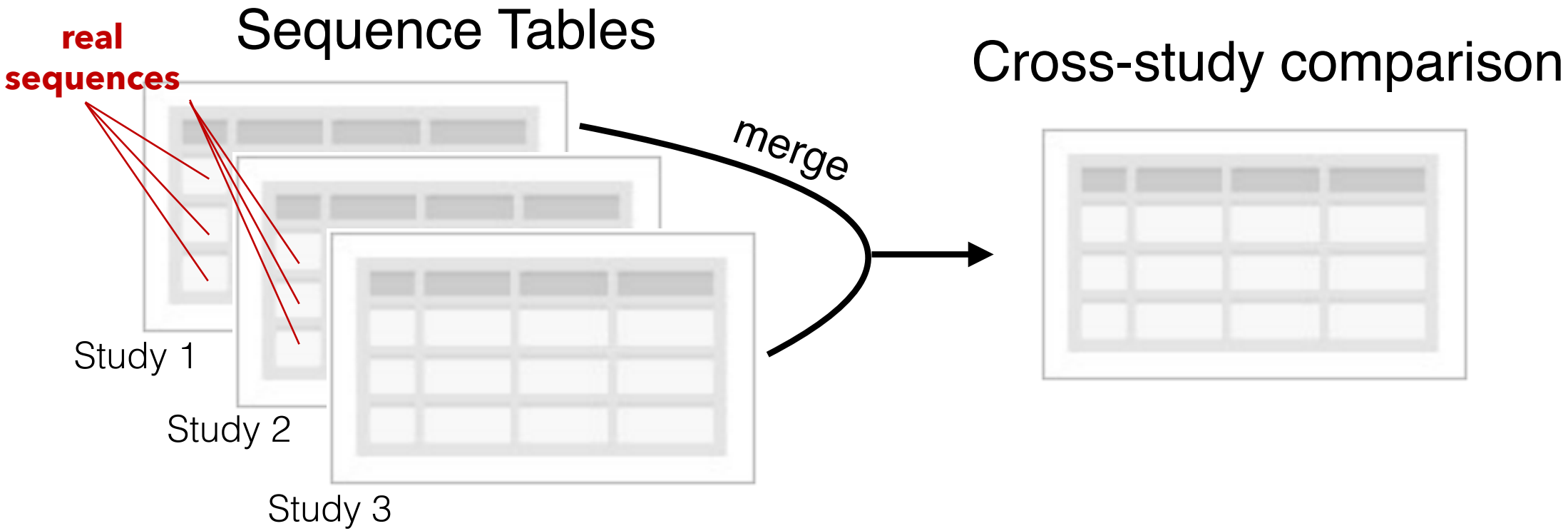
ASVs (aka. ESVs, zOTUs)

+ pipeline including denoising & filtering, chimera removal, OTU table merging…
+ R interface
+ steady maintenance
+ good documentation
+ use cases for targets other than 16S
+ settings for non-Illumina data
– not the most resource-efficient

© B. Callahan (DADA2 documentation)

ASVs (aka. ESVs, zOTUs)

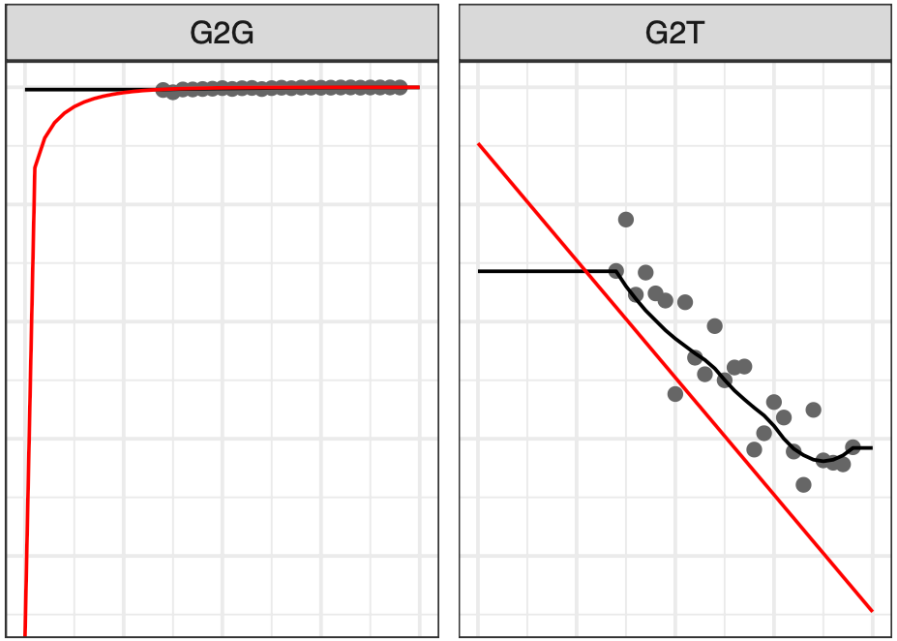
- cross-study comparability



© B. Callahan (DADA2 documentation)

ASVs (aka. ESVs, zOTUs)

- model substitutions for every run




```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                    |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

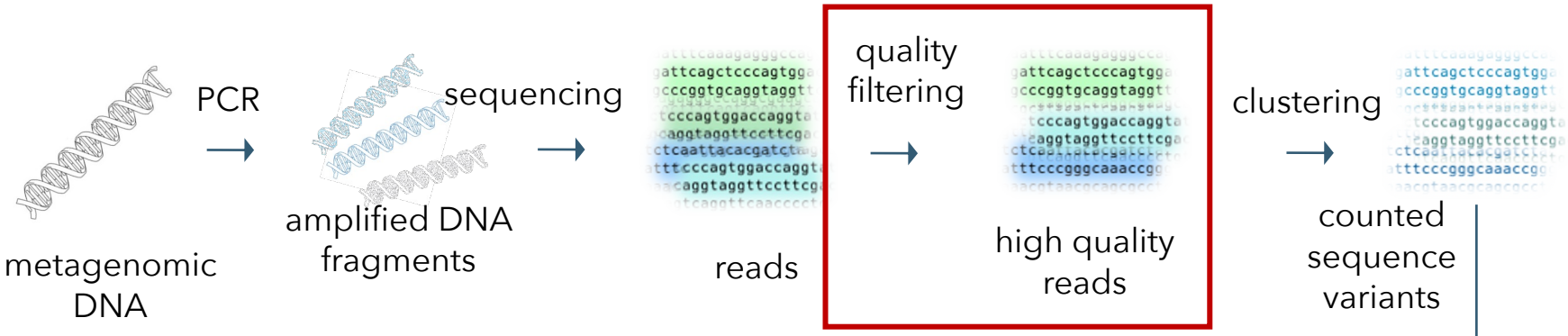
$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

Error rates depend on....

- Substitution (eg. A->C)
- Quality score (eg. Q=30)
- Batch effect (eg. run)

© B. Callahan (DADA2 documentation)

Metabarcoding workflow



taxonomic / phylogenetic classification

classified sequence variants

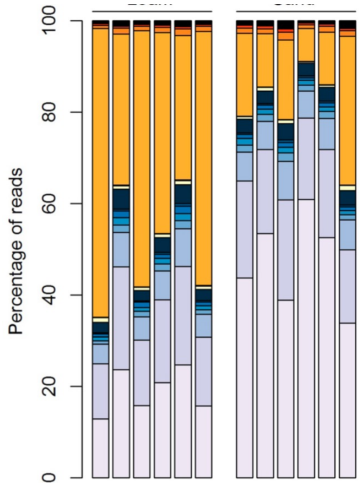
taxonomic profiles of microbial communities

alpha-diversity of microbial communities

functional prediction

annotated sequence variants

statistical analyses



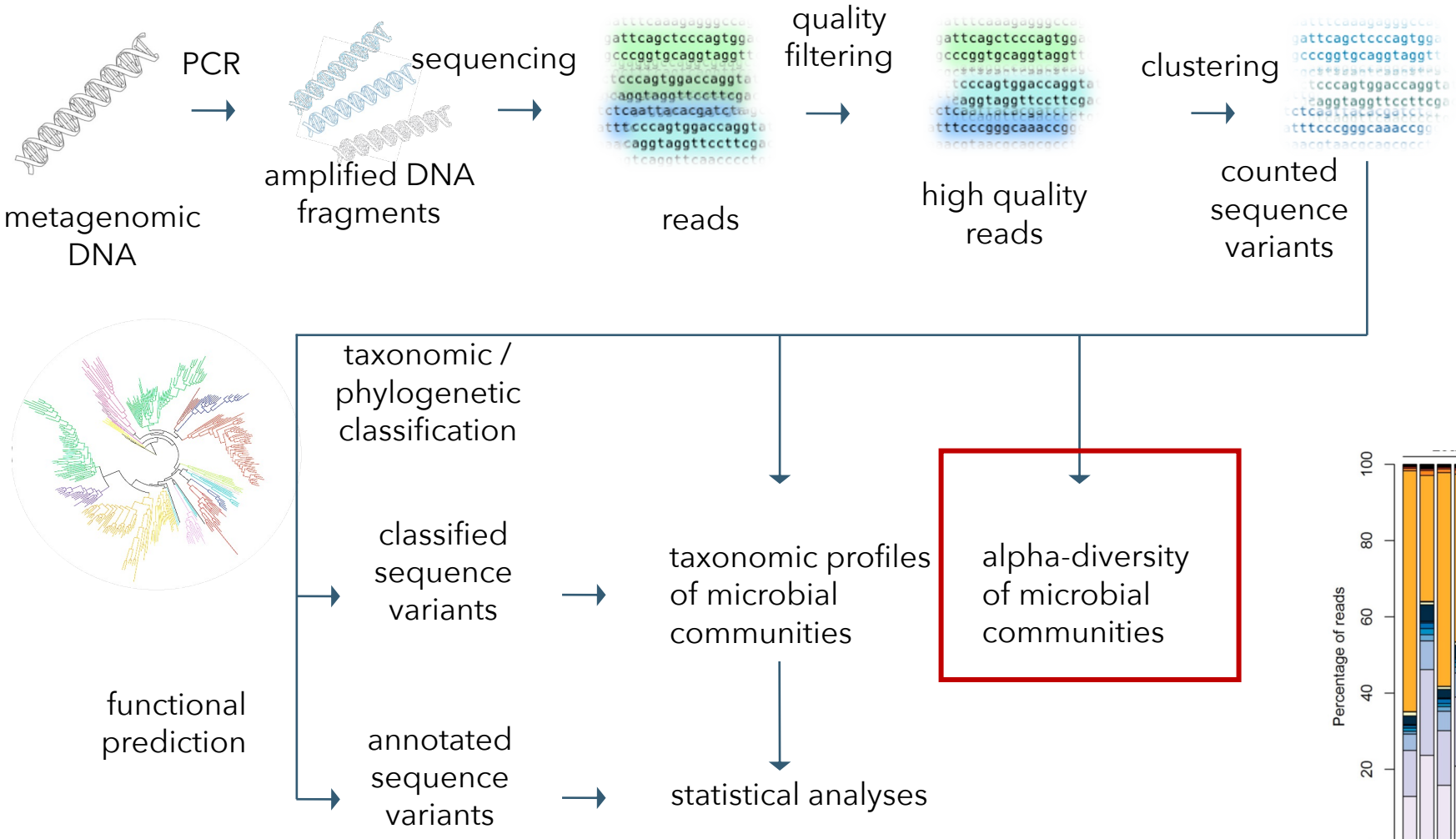
Quality filtering

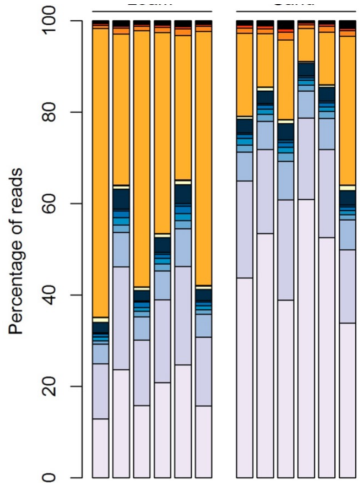
- a few positions with low quality don't hurt

- off-target sequences should be removed before modelling
  - spike-in (phiX174)
  - incomplete sequences
  - dark-cycle positions from novaseq and nextseq machines (2-color chemistry)

- note: errors from PCR are not detected by error models

Metabarcoding workflow

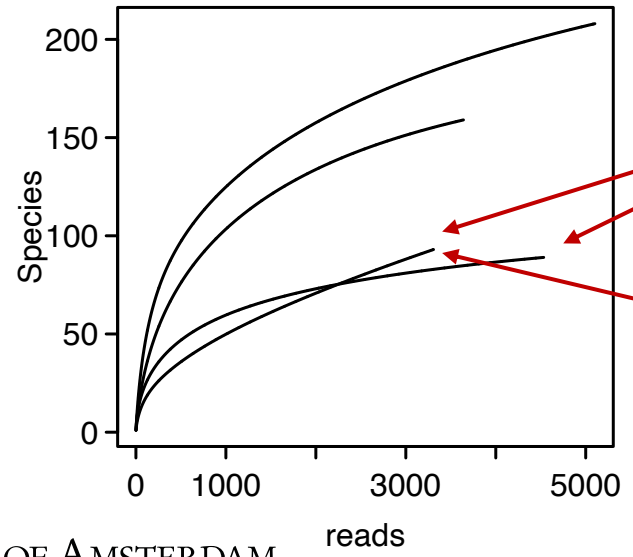




Unequal sampling depth does not reflect biology

- it's pure chance how many reads a sample gets
- representation can be unfair

without:

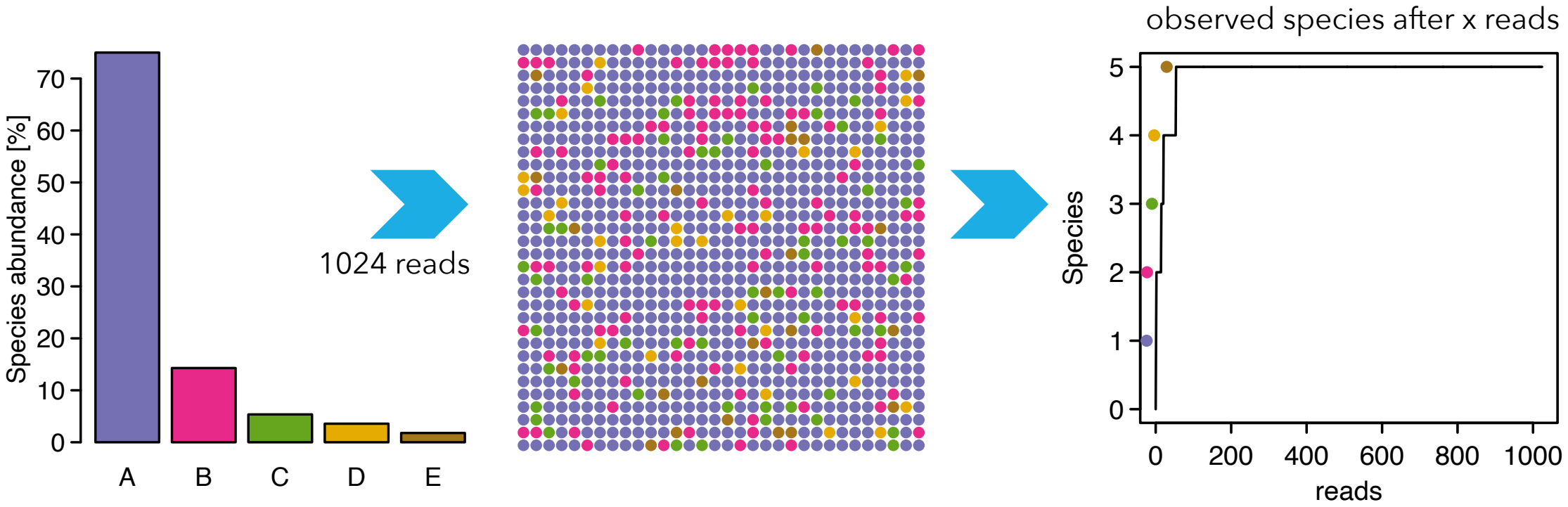


these two seem to have the same number of species

but this one would certainly outgrow the other

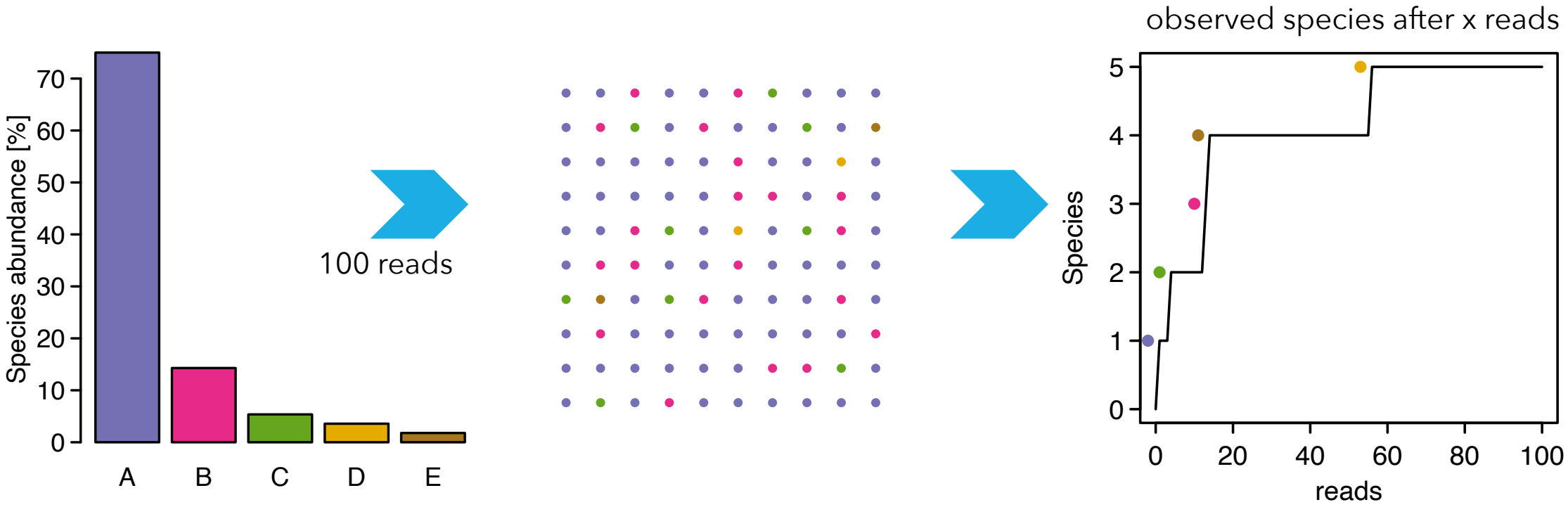
Rarefaction curves & analysis

- number of new species in increasing subsamples



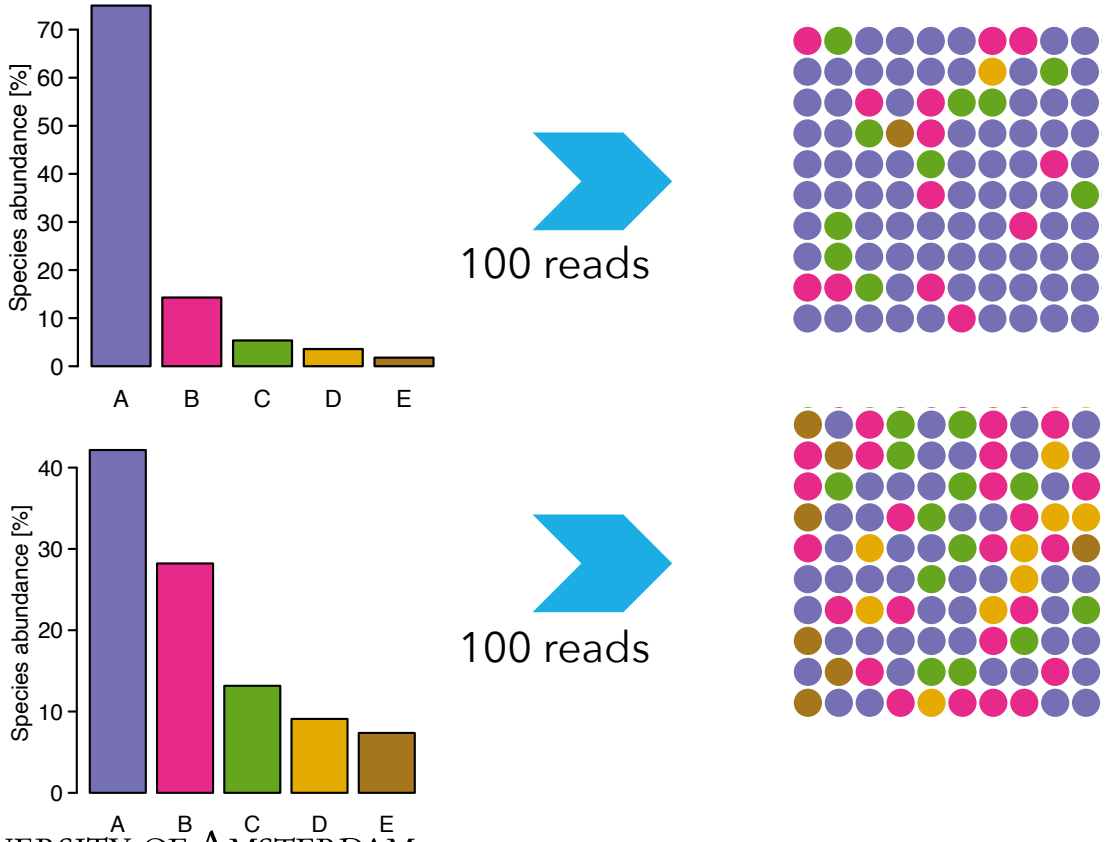
Rarefaction curves & analysis

- number of new species in increasing subsamples



Rarefaction (normalisation)

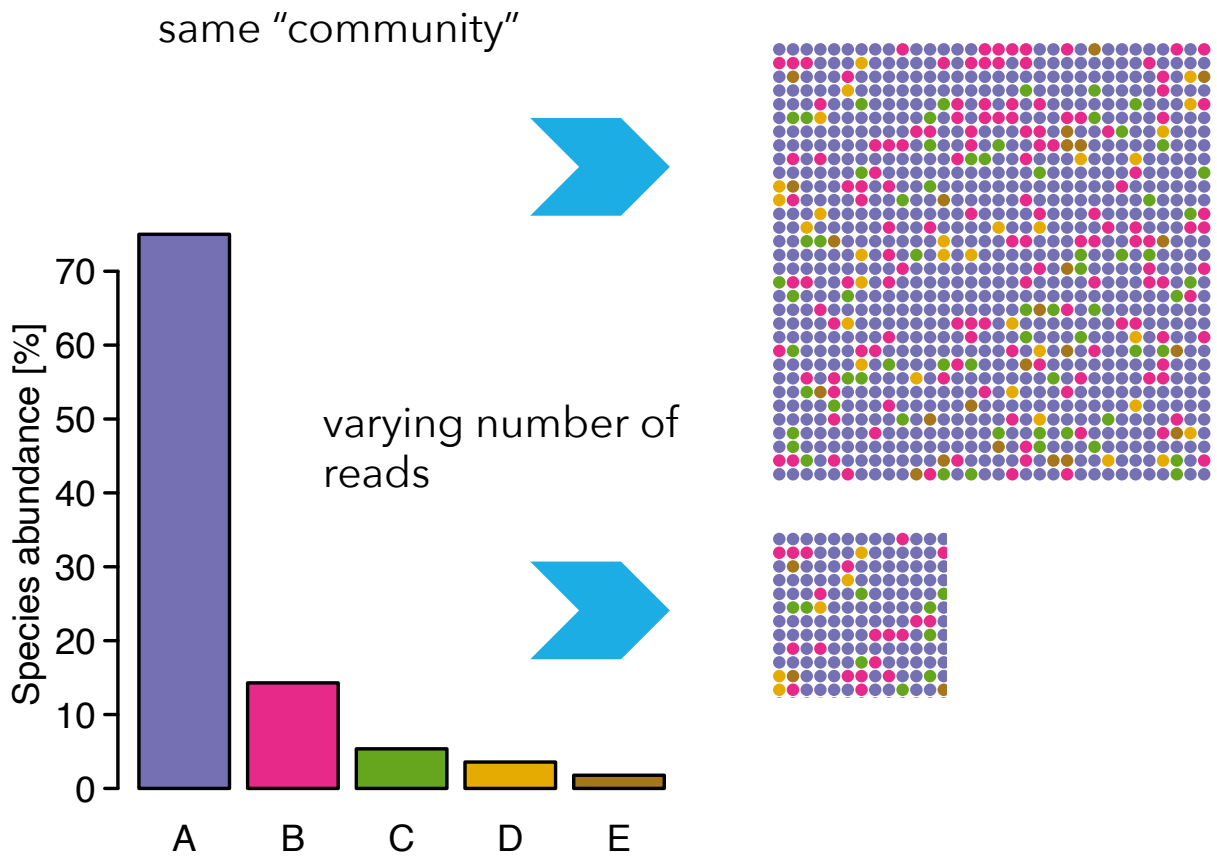
- subsample reads to keep equal numbers per sample

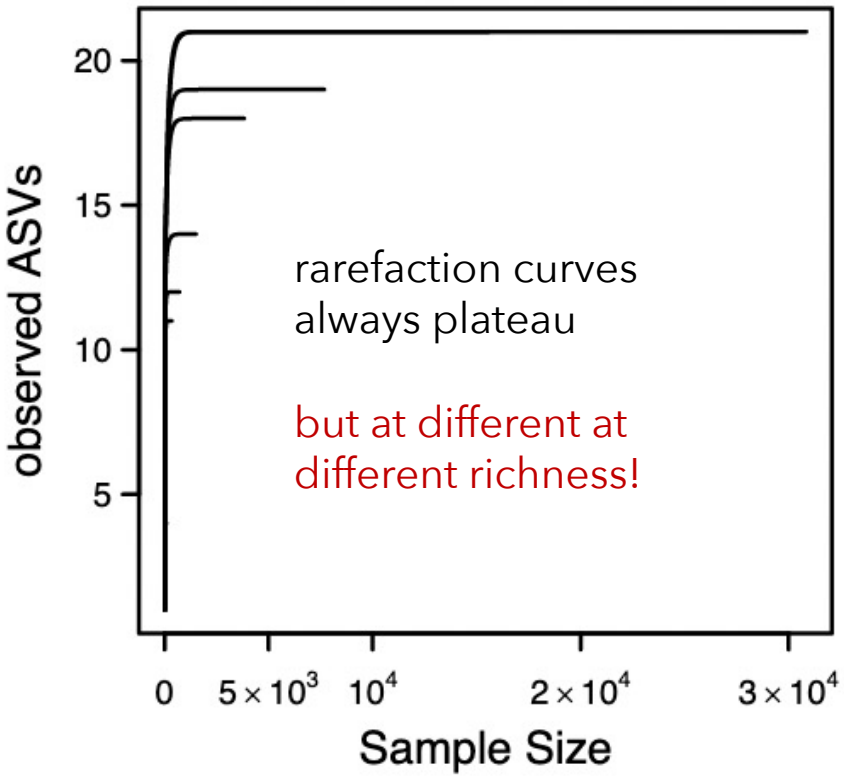


+ all samples have the same sum of reads

! most people do this after ASV generation

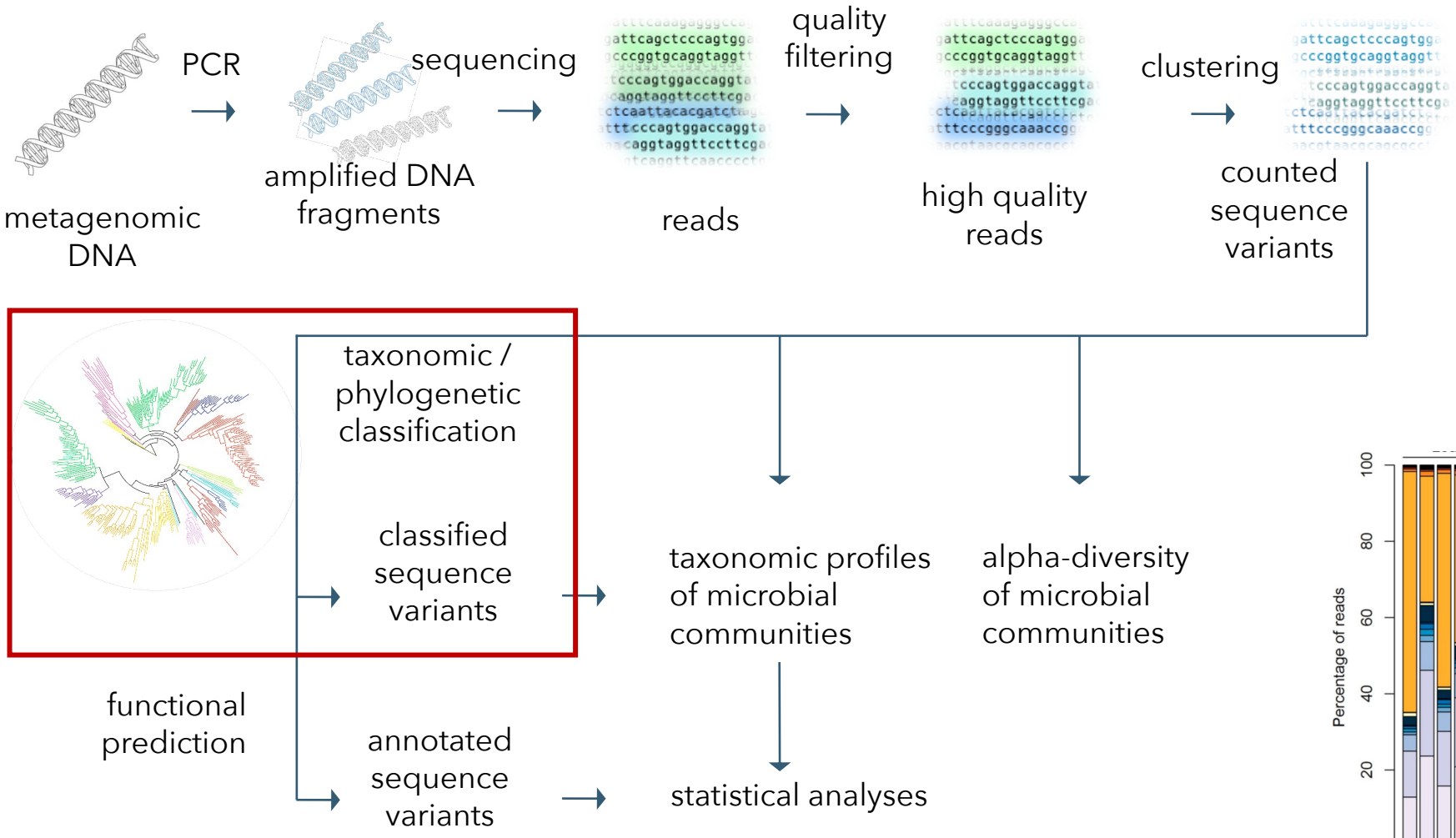
post-ASV rarefaction does not work

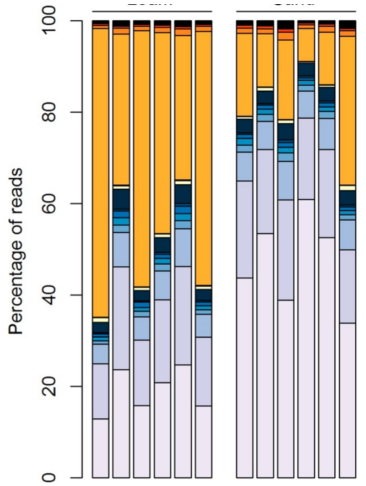




Weißbecker *et al.*, 2020, bioRxiv, doi: 10.1101/2020.05.17.095679v1

Metabarcoding workflow

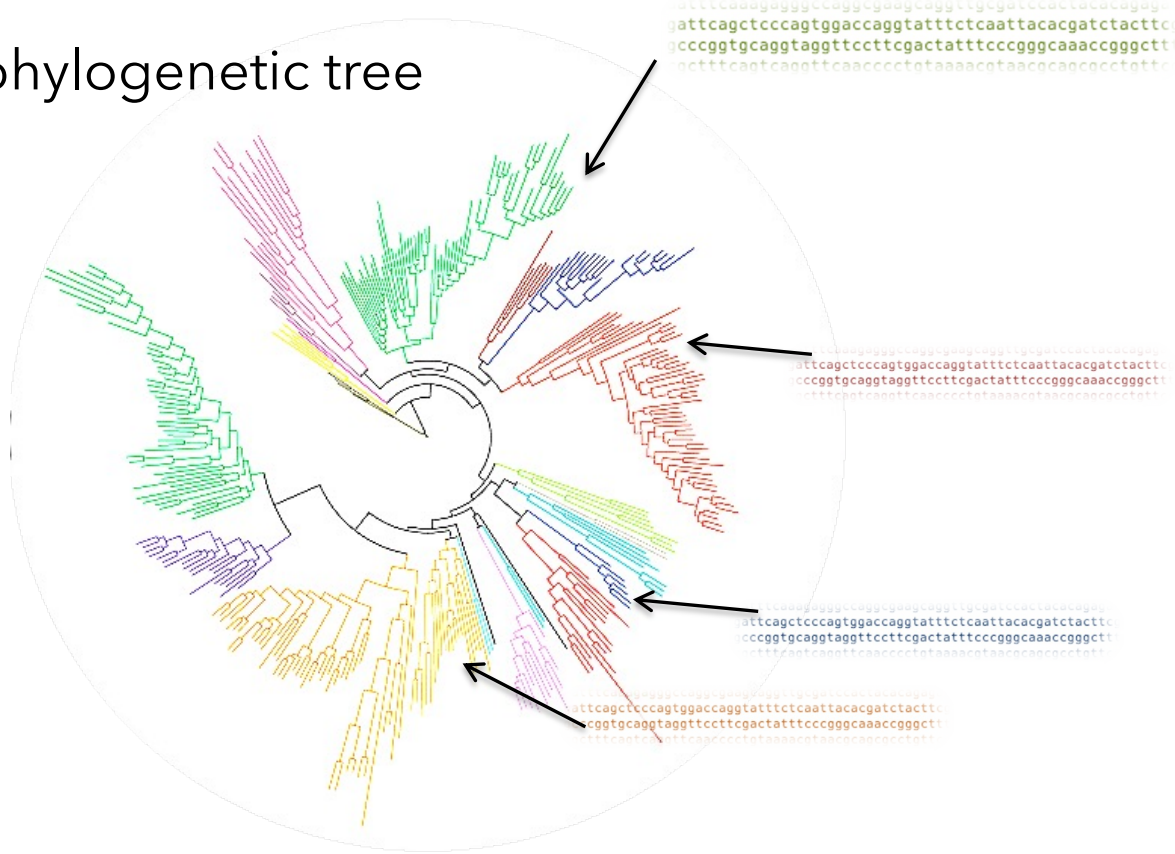




Taxonomic classification of ASVs

compare the
taxa
in the database

to your ASVs



Wu *et al*. 2009 Nature

Naïve Bayesian Classifier

P(taxon|sequence composition) = P(sequence composition|taxon) * P(taxon) / P(sequence composition)

read: the probablity of the taxon being right given the sequence data
is equal to

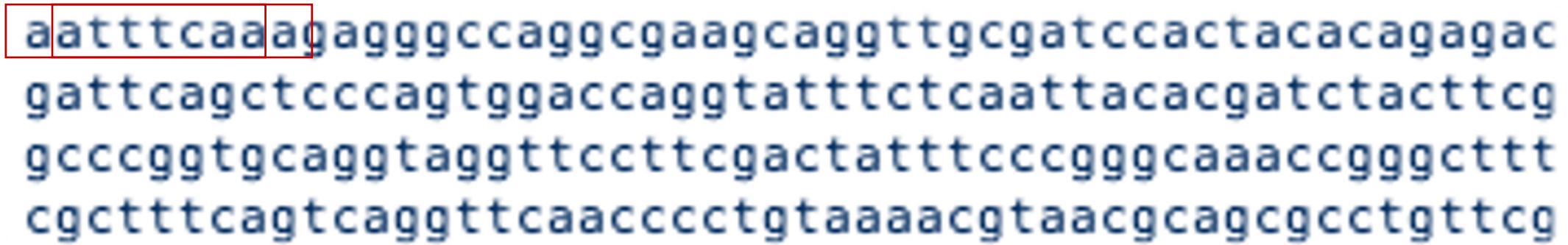
the probablity that the sequence composition is correct given the taxon

(while the probabilty of the taxon and the probability of the sequence composition being true are constant)

Wang *et al.* 2007 Applied & Environmental Microbiology

Taxonomic classification of ASVs

sequence of one ASV/taxon:



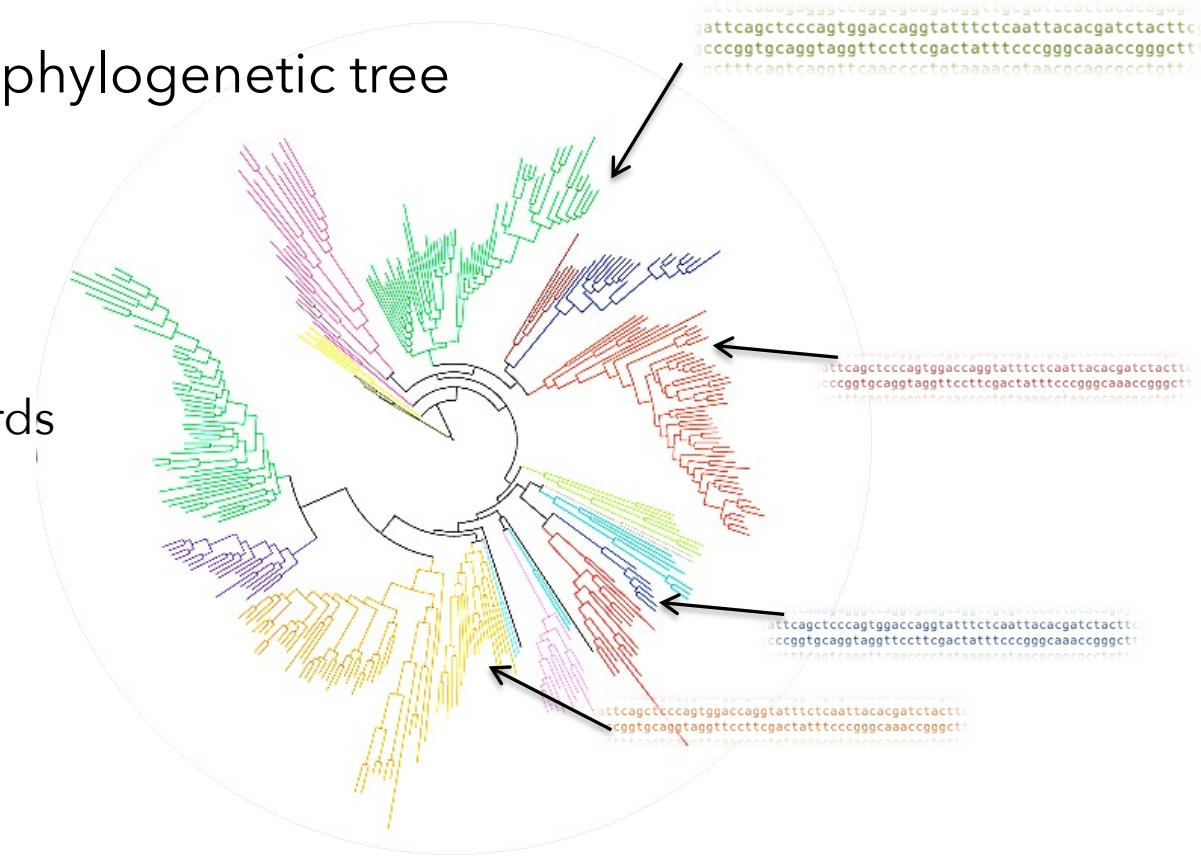

```
aatttcaa
atttcaaa
...
```

65,536 possible words

Taxonomic classification of ASVs

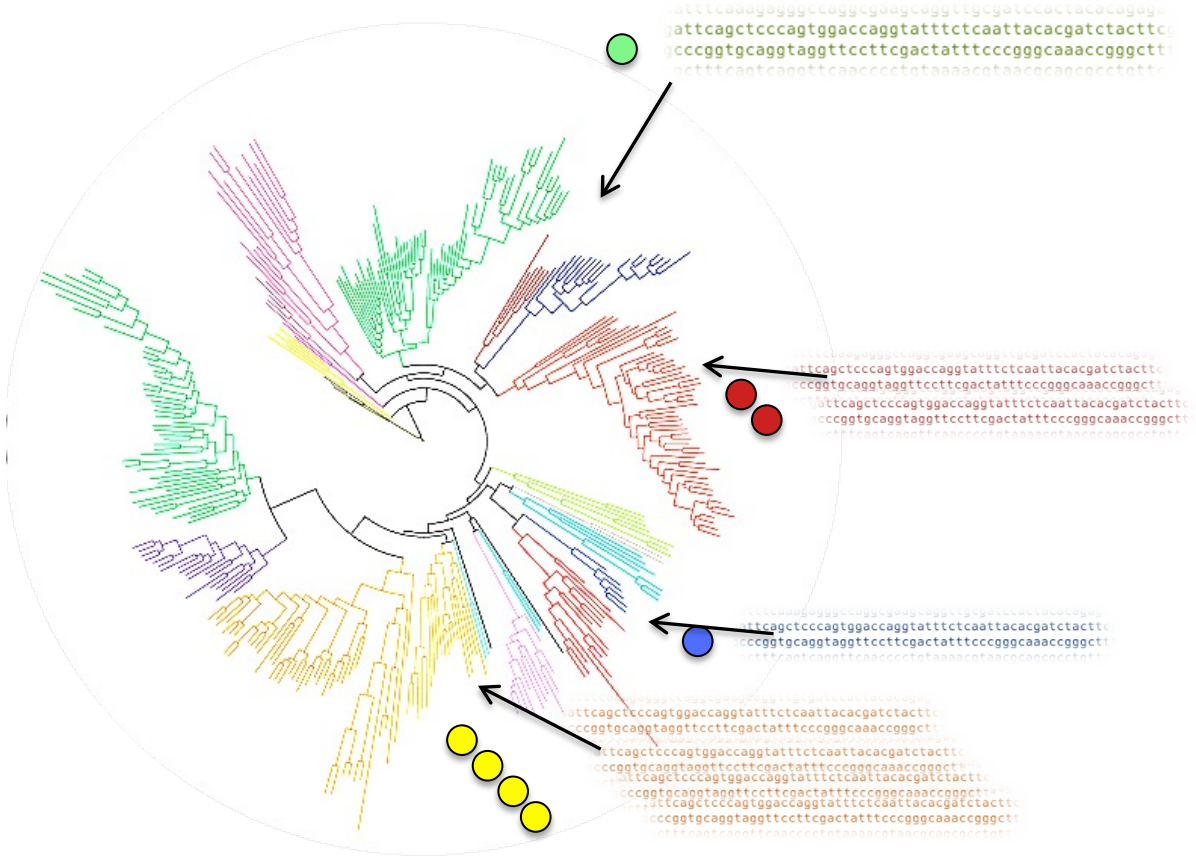
compare the
~500,000 taxa x 65,536 possible words
in the database

to the words in your ASVs



Wu *et al*. 2009 Nature

Taxonomic profiling



Wu *et al*. 2009 Nature

count the number of reads per ASV/taxon

	ASV 1	ASV 2	ASV 3	ASV 4
Sample 1	1	2	1	4
Sample 2	2	2	3	2
Sample 3	1	0	0	1
Sample …	…	…	…	…
Sample N	4	1	7	0

The problems with always counting reads out of a total

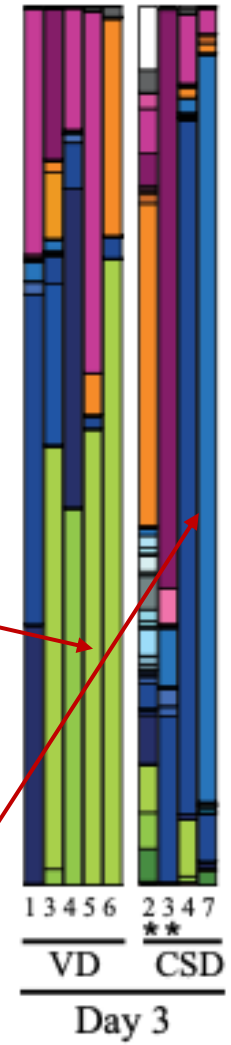
- the sum of reads is often not representative of anything we know

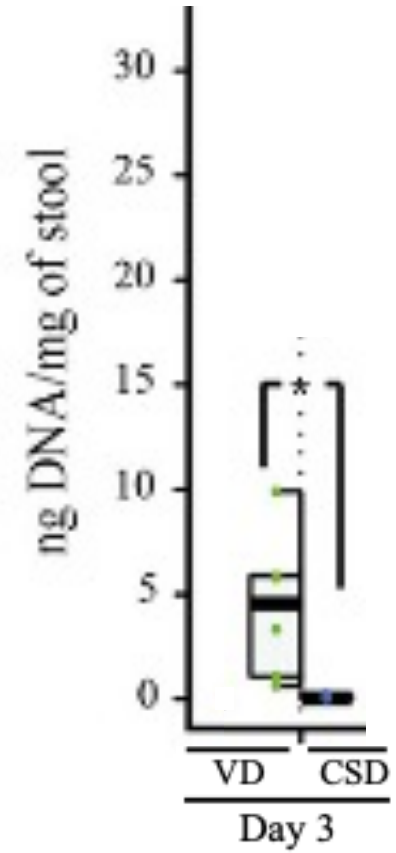
- we need to measure the total to be sure

are these missing in the other samples?

OR

are these a lot more abundant in here?

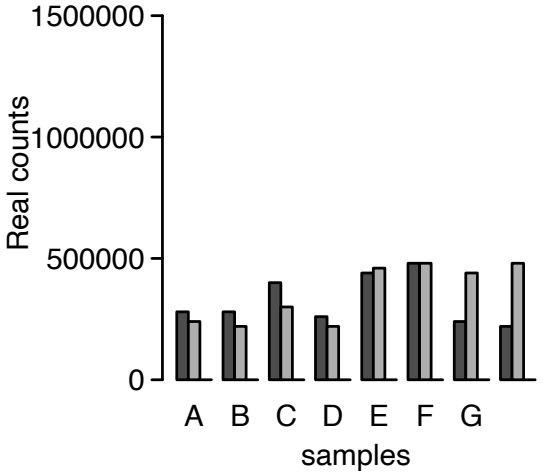


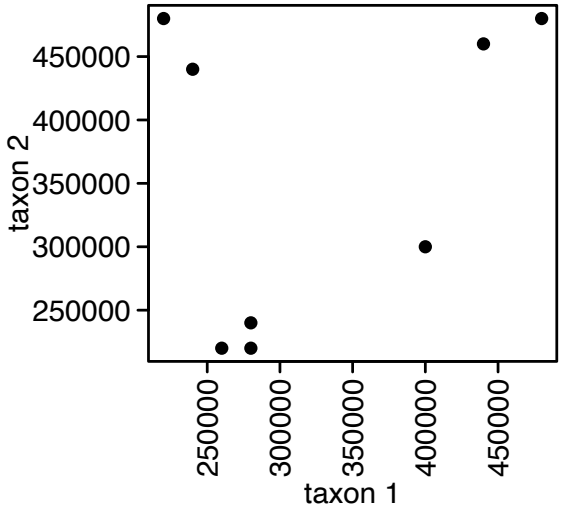


Wampach *et al.* 2017
Frontiers in Microbiol

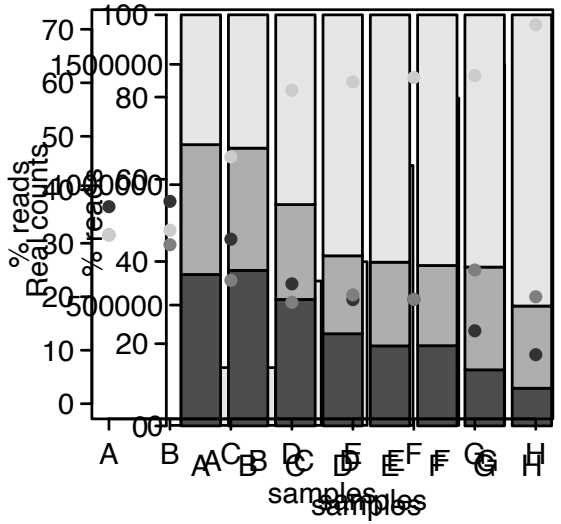
Compositionality - example

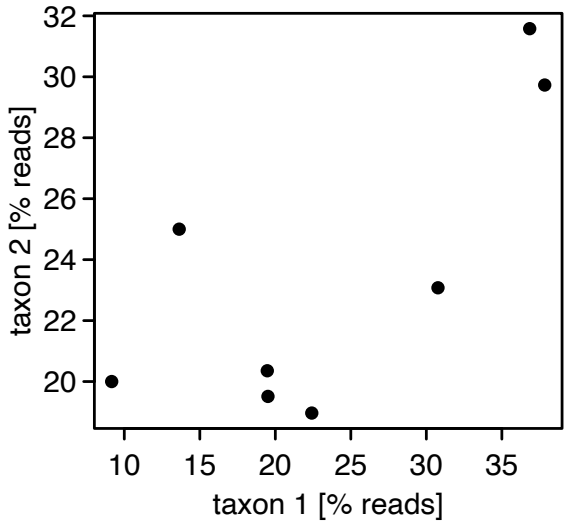
- taxa 1 and 2 have no special relationship





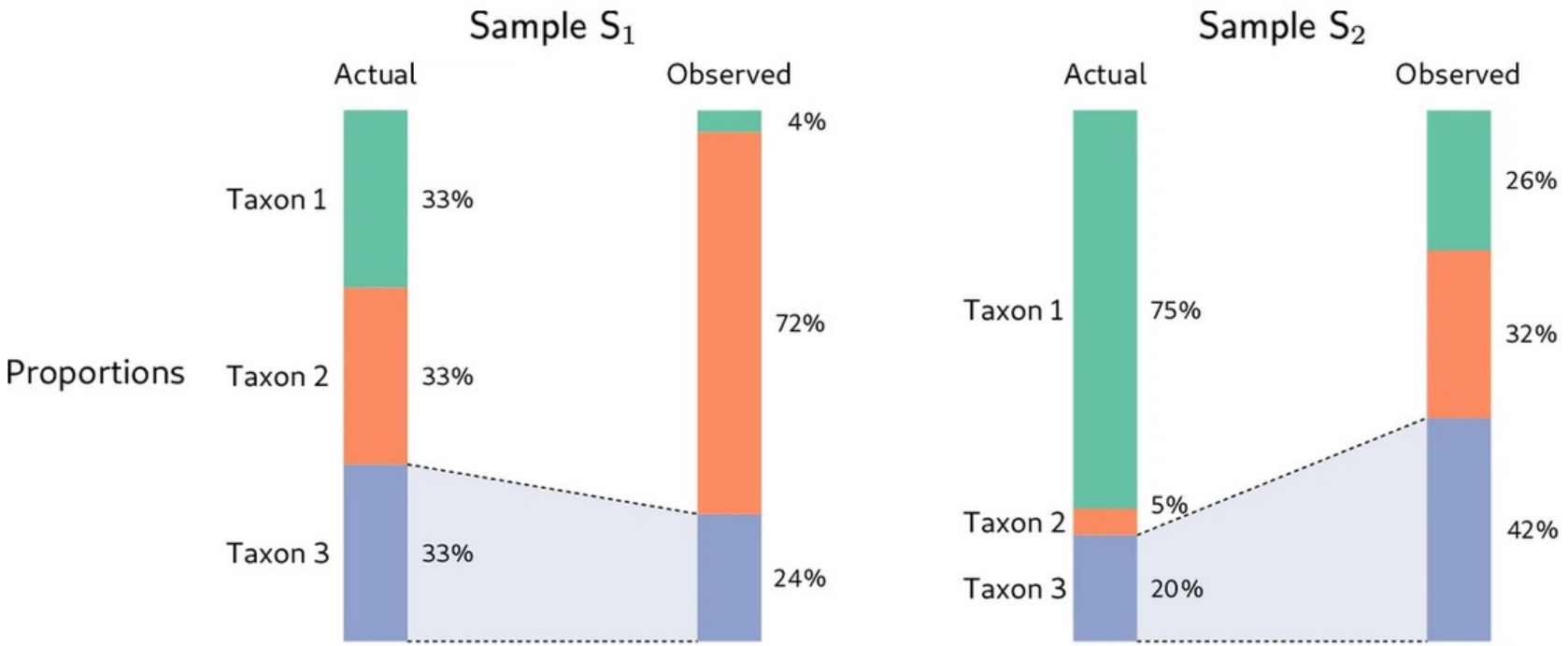
- taxon 3 **introduces a positive correlation**





Add to that: bias

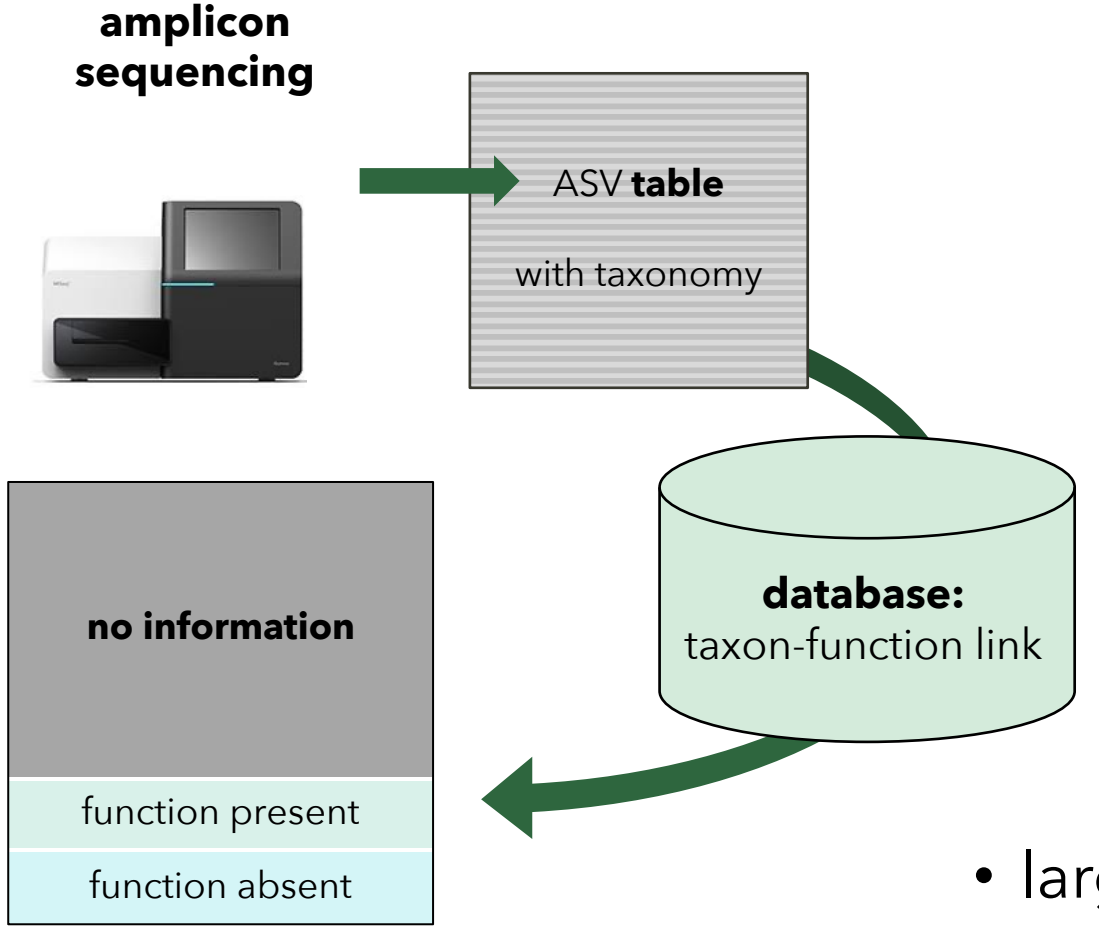
- assuming a constant bias for every ASV/taxon:

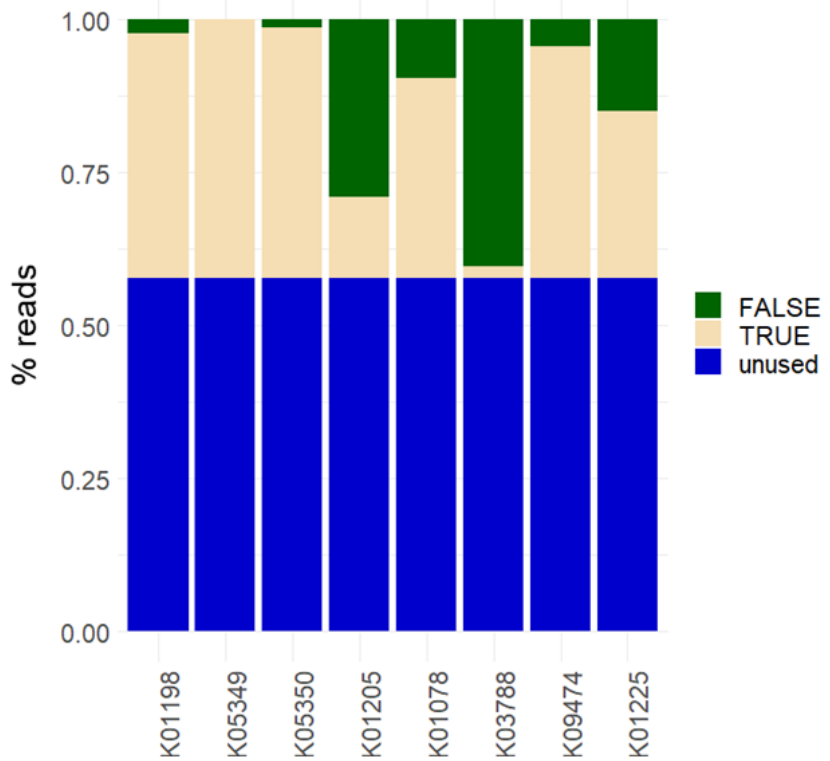


- you can transform data and correct for biases

MR McLaren, AD Willis, BJ Callahan (2019) Consistent and correctable bias in metagenomic sequencing experiments. eLife 8:e46923.

A word on functional prediction





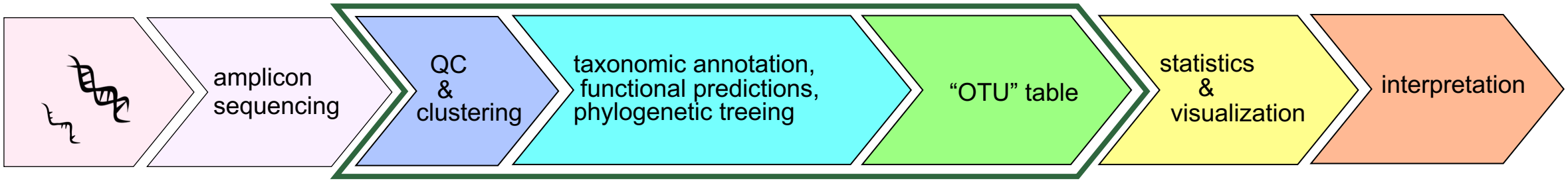
- large proportion of the community is not taken into account
- usefulness depends on the context

Breitkreuz *et al.* 2021 Microbiology Spectrum 9: e00278-21

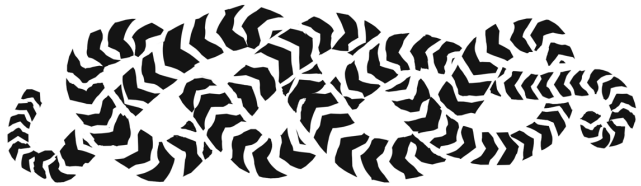
Summary

- **research questions** can ask about sample-sample or sample-(exp.)factor or taxon-taxon or taxon-(exp.)factor relationships
- **marker genes** determine target organisms and resolution – studies using different markers are incomparable
- **biases** in sample processing persist
- sequencing data **processing** needs to be **error-aware**
- sequencing data processing to ASVs dictates **pre-processing**
- use of ASVs (also OTUs) facilitates **taxonomic profiling**
- **functional predictions** need to be handled with caution
- **metabarcode-sequencing-based numbers are treacherous**

After the break….



dadasnake



Thanks for your attention!

a.u.s.heintzbuschart@uva.nl

SP C2.205

github.com/a-h-b

twitter.com/_a_h_b_

University of Amsterdam
Life Sciences