

Metabarcoding Workshop

Anna Heintz-Buschart

June 2022



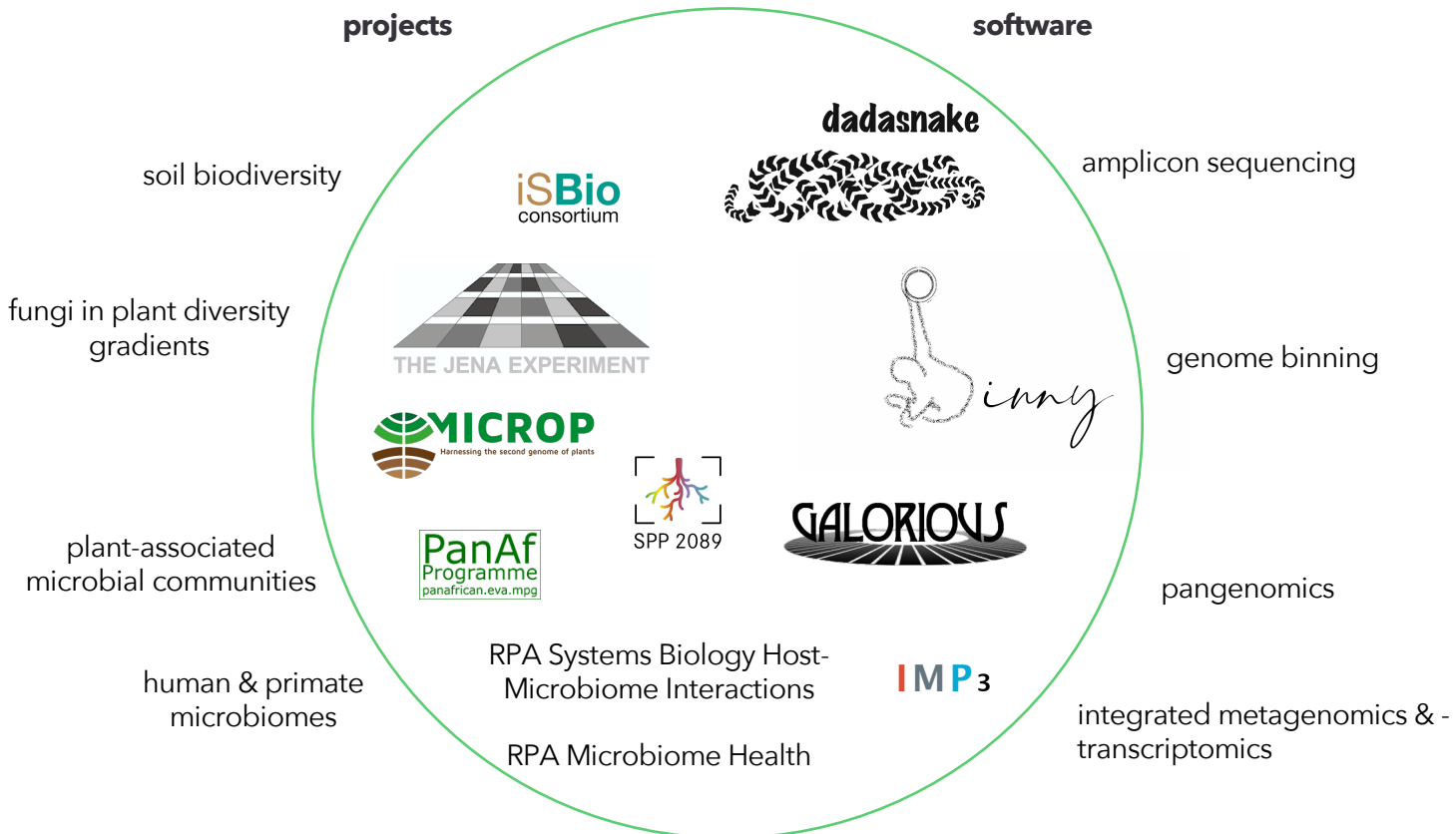
“Metabarcoding”

- 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)



2011

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



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



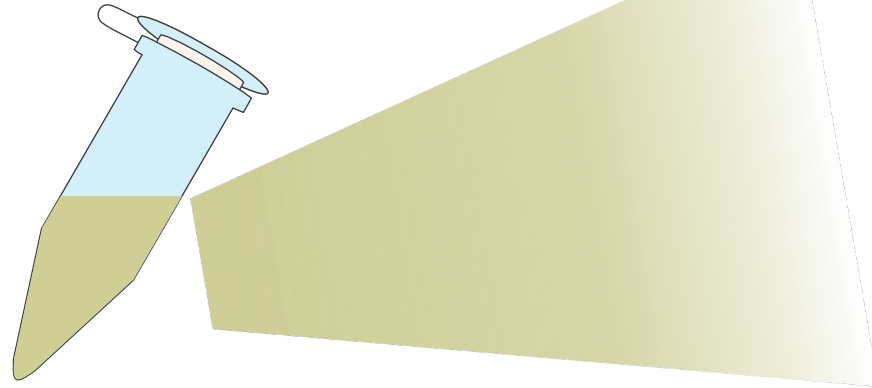
github.com/a-h-b



twitter.com/_a_h_b_

Questions

what is in my sample?

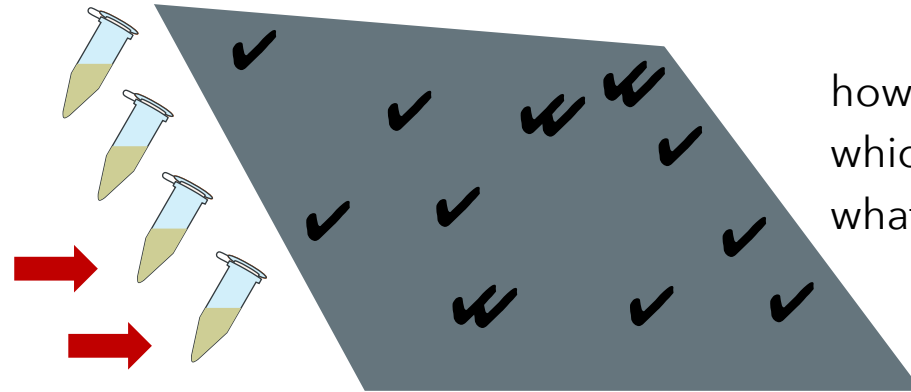
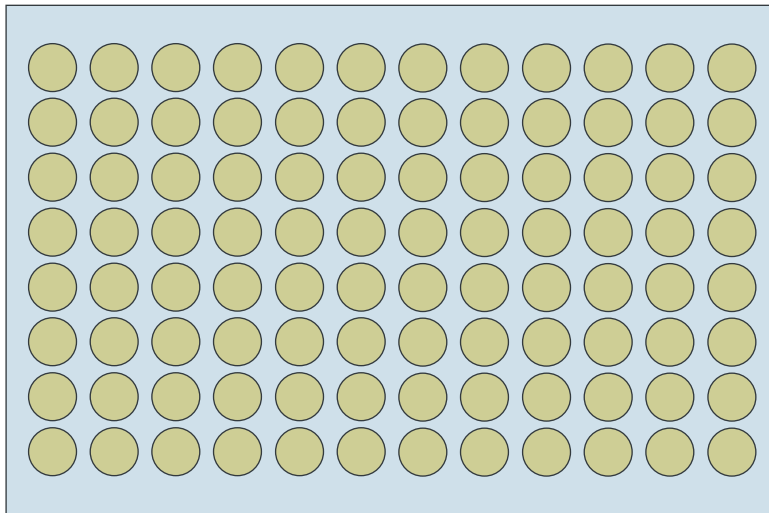


mixed community



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?

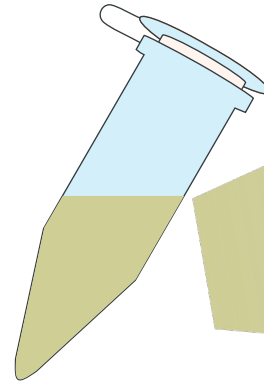
when are they there?



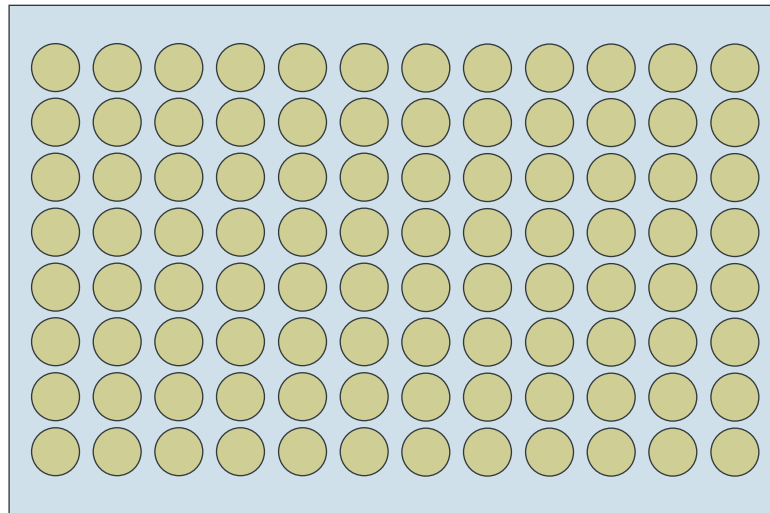
who is there often/in high numbers?
who is there with whom?

Questions

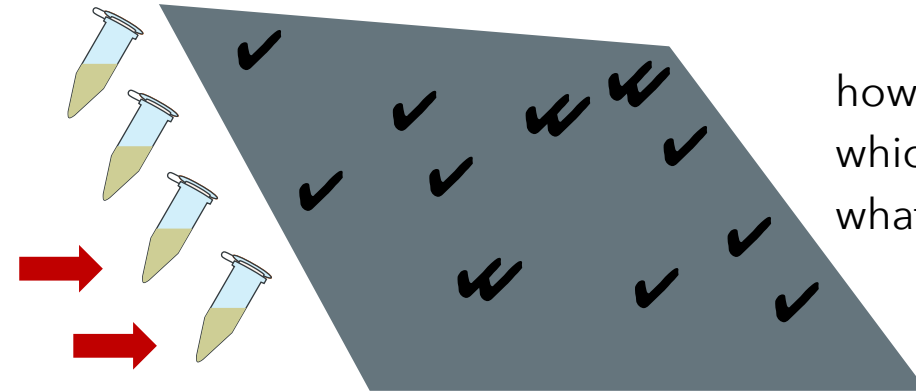
what is in my sample?



what is in my samples?



mixed community



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



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

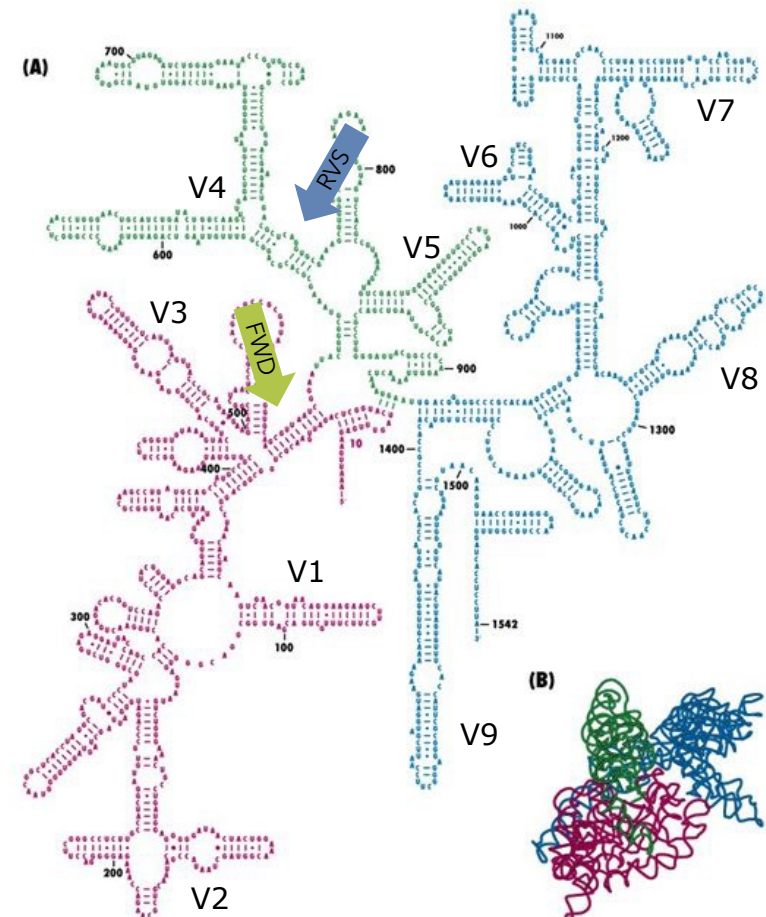
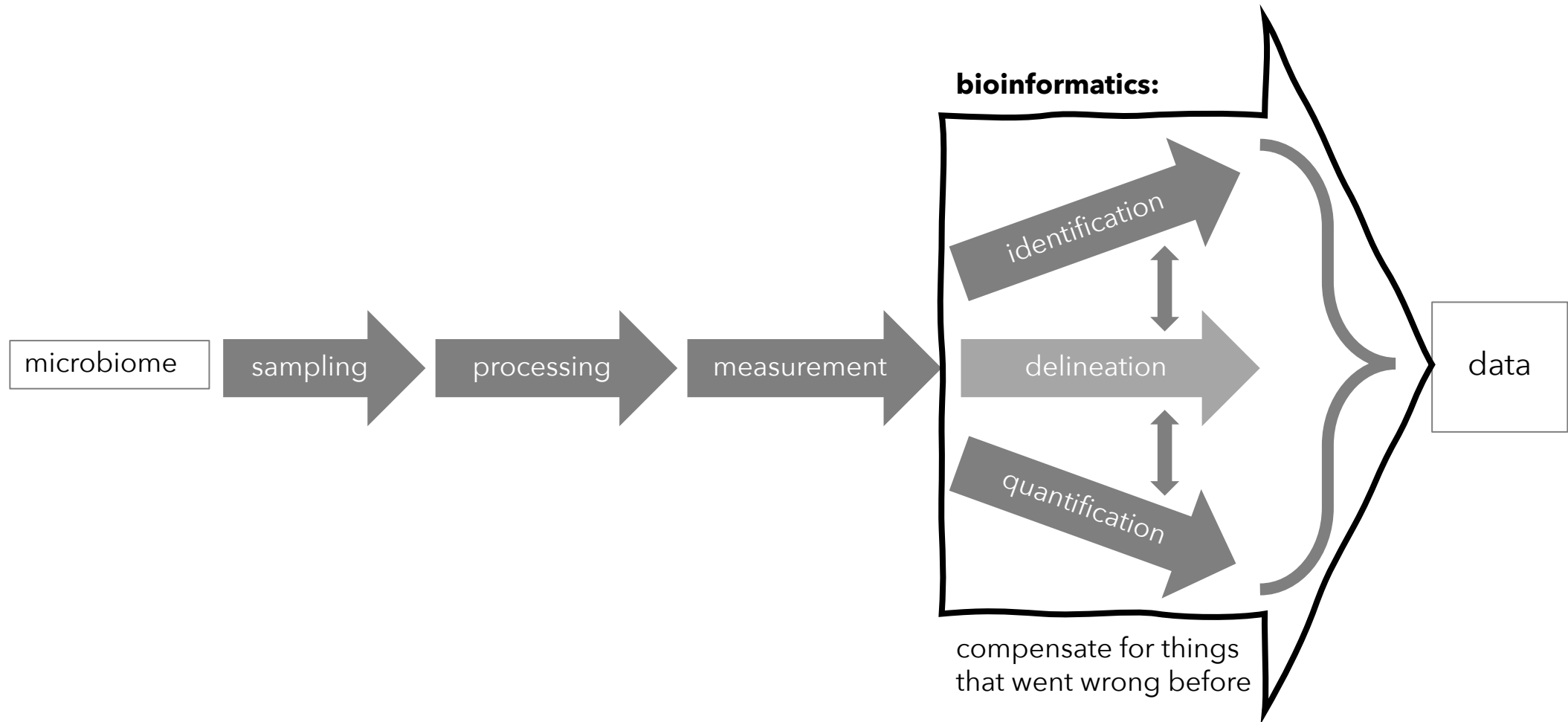


Figure 30-14
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

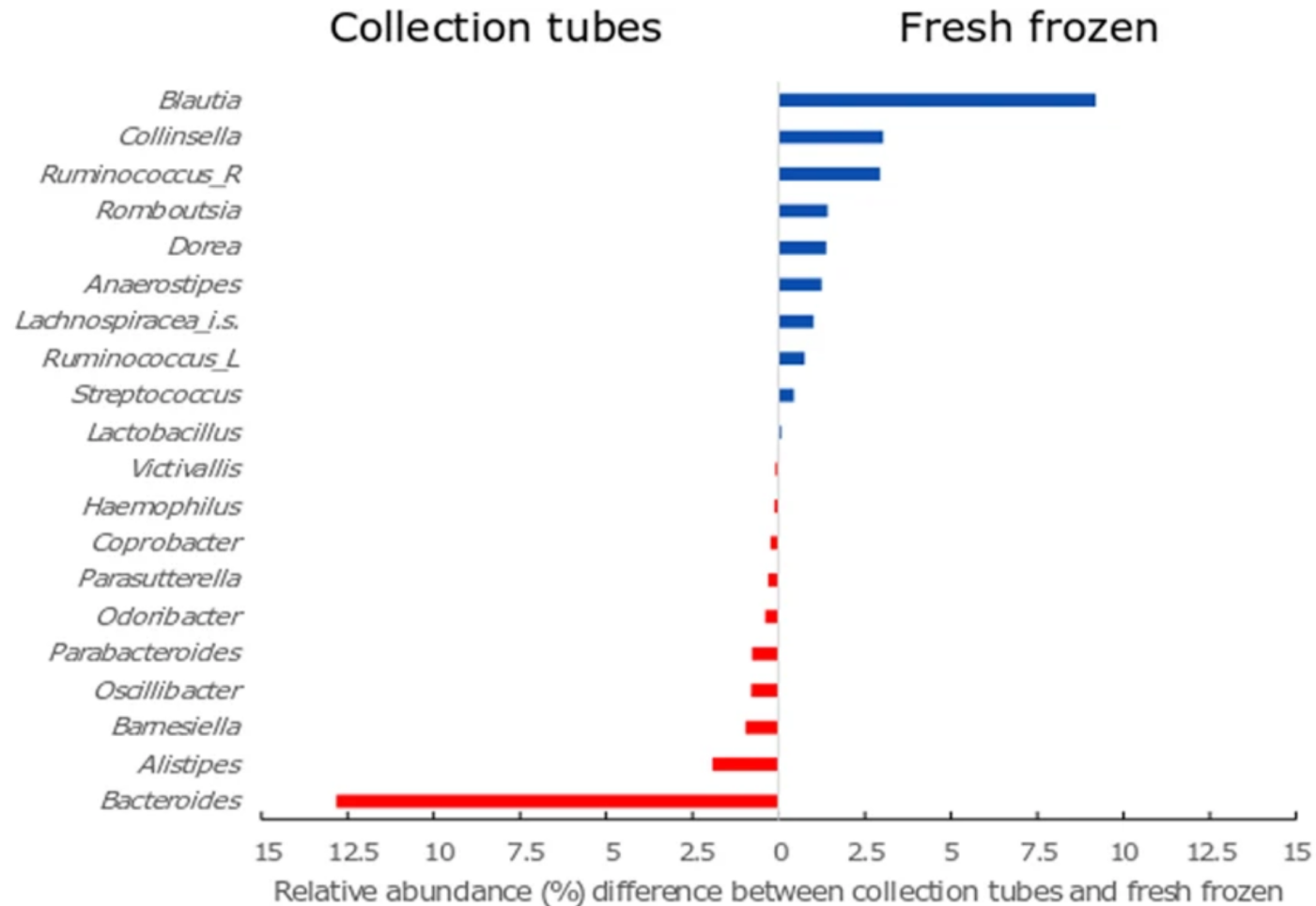
Measuring microbiomes: marker genes

- 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 (*rbcL*)
 - tRNA^{Leu} intron (*trnL*)
 - RNA polymerase (*rpoB*)

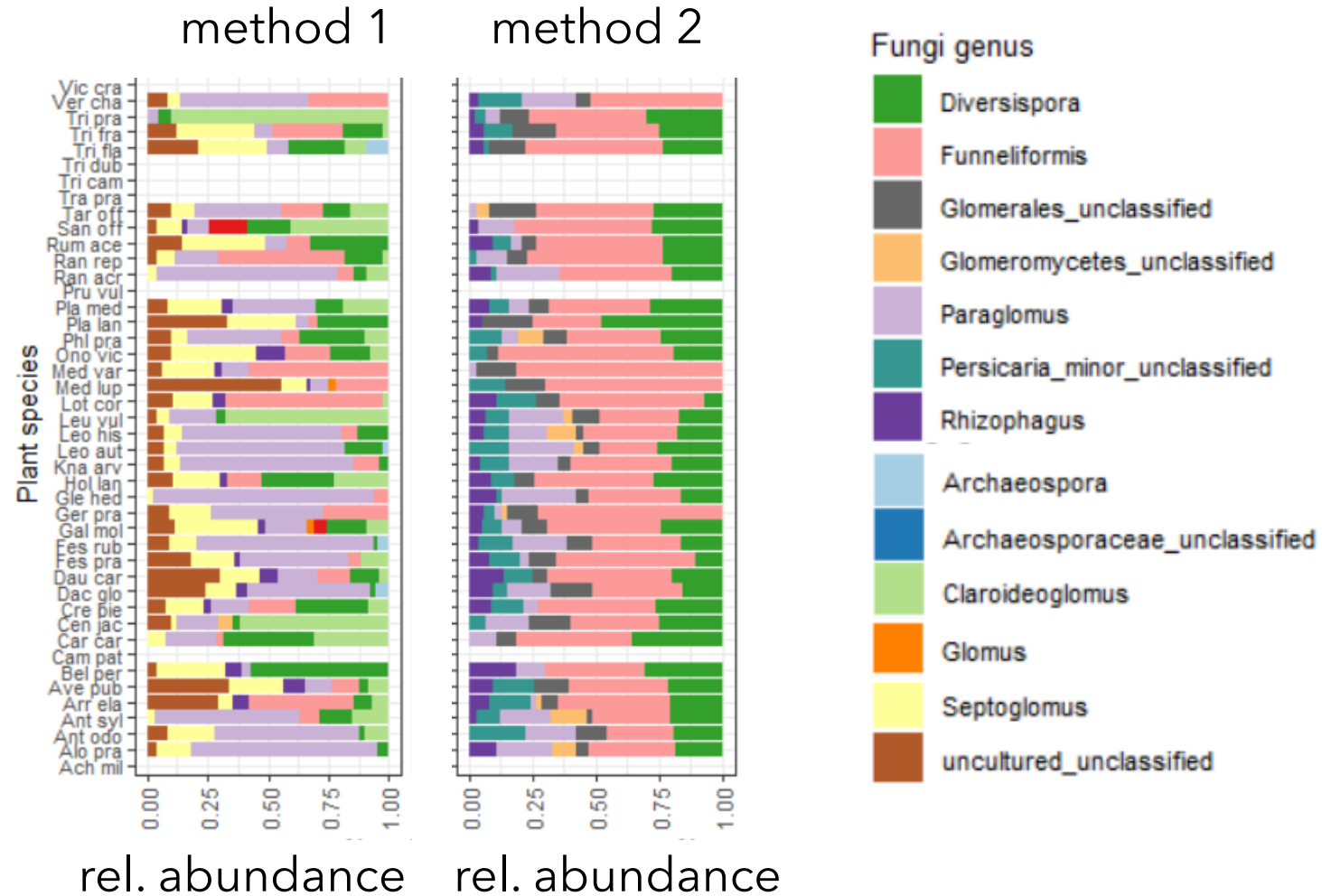
Omics paradigm



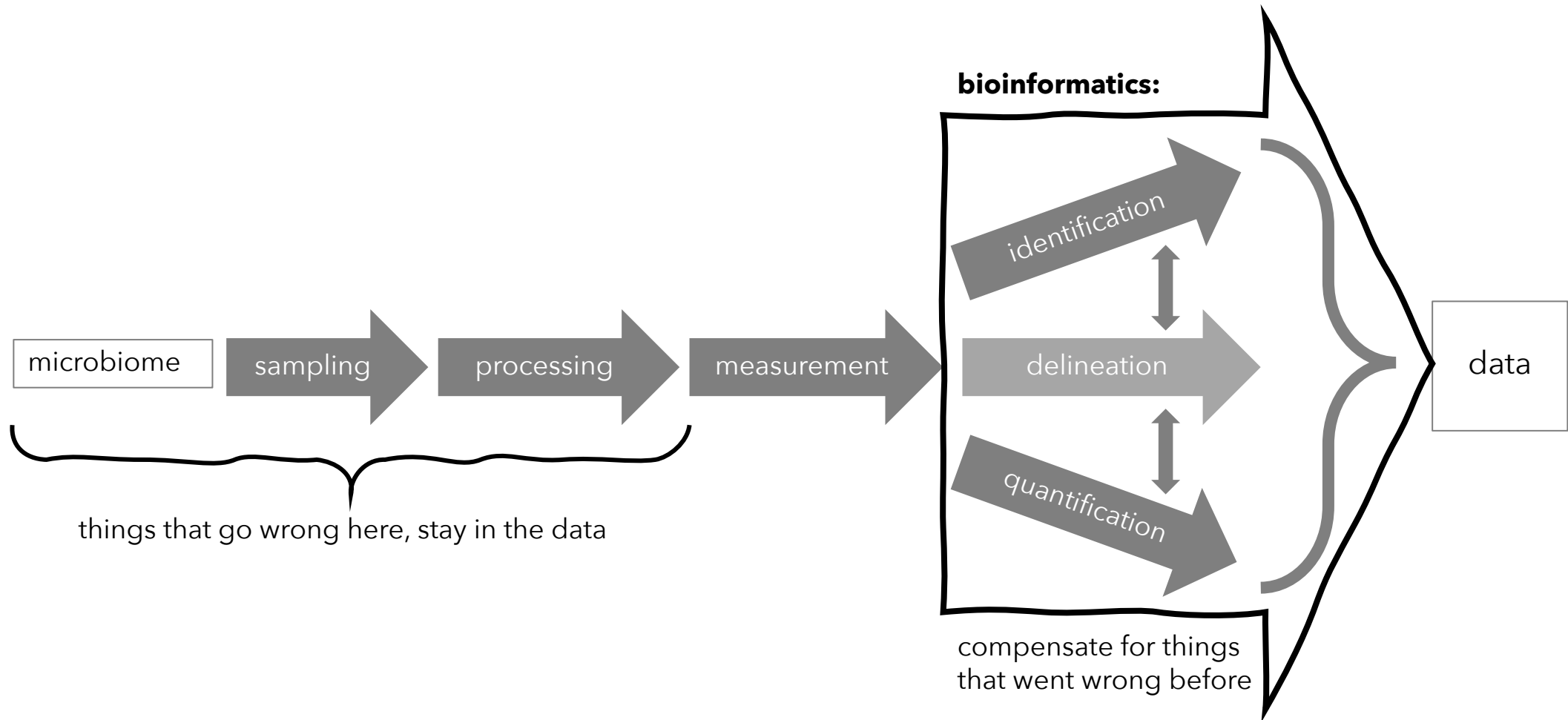
What could go wrong? Sample storage/processing



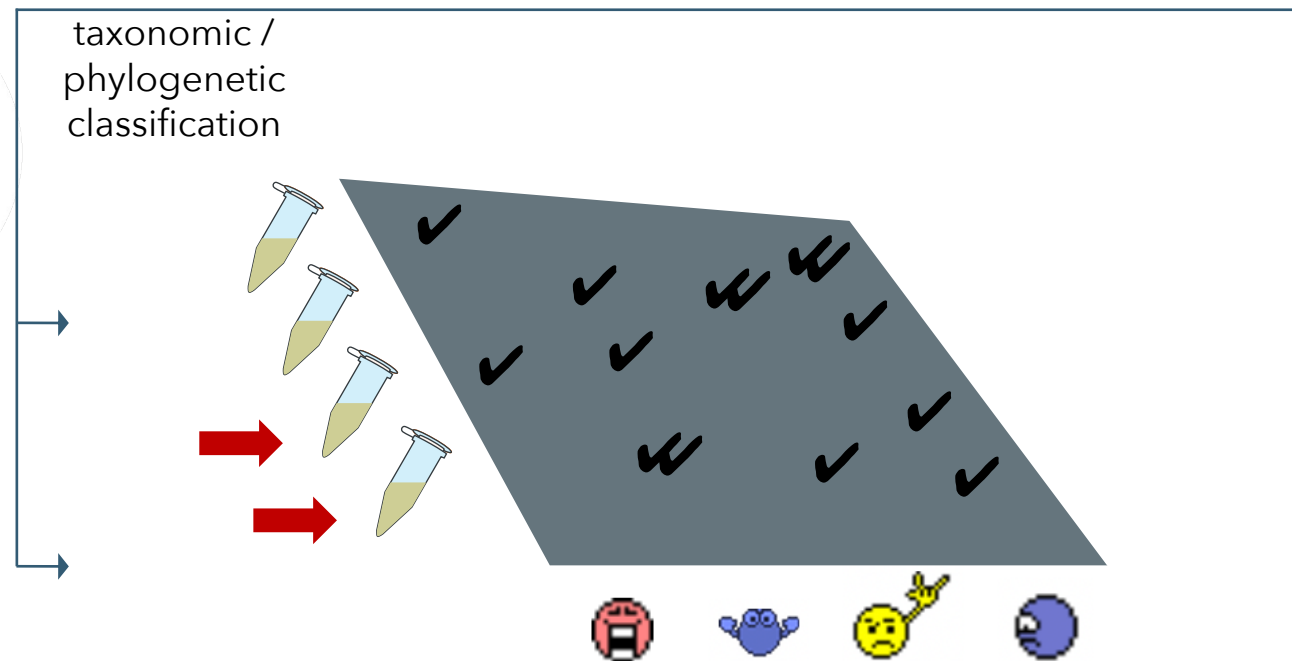
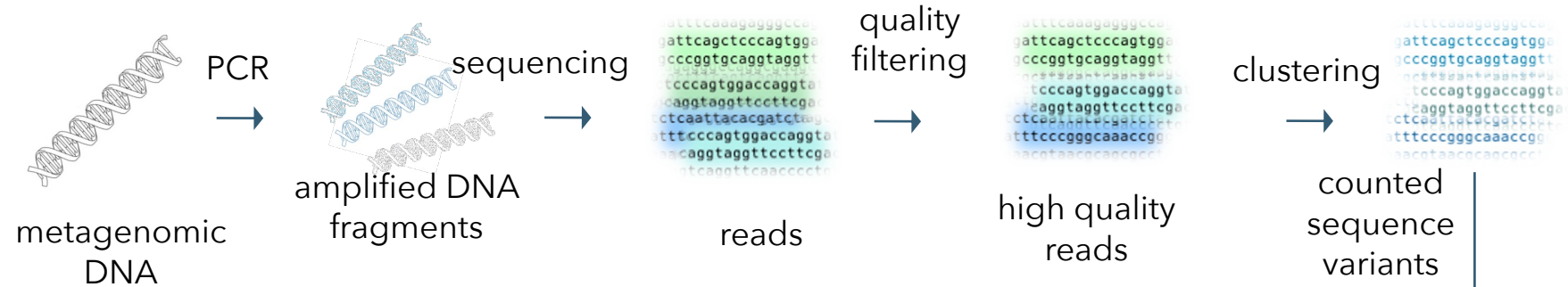
What could go wrong? Amplification



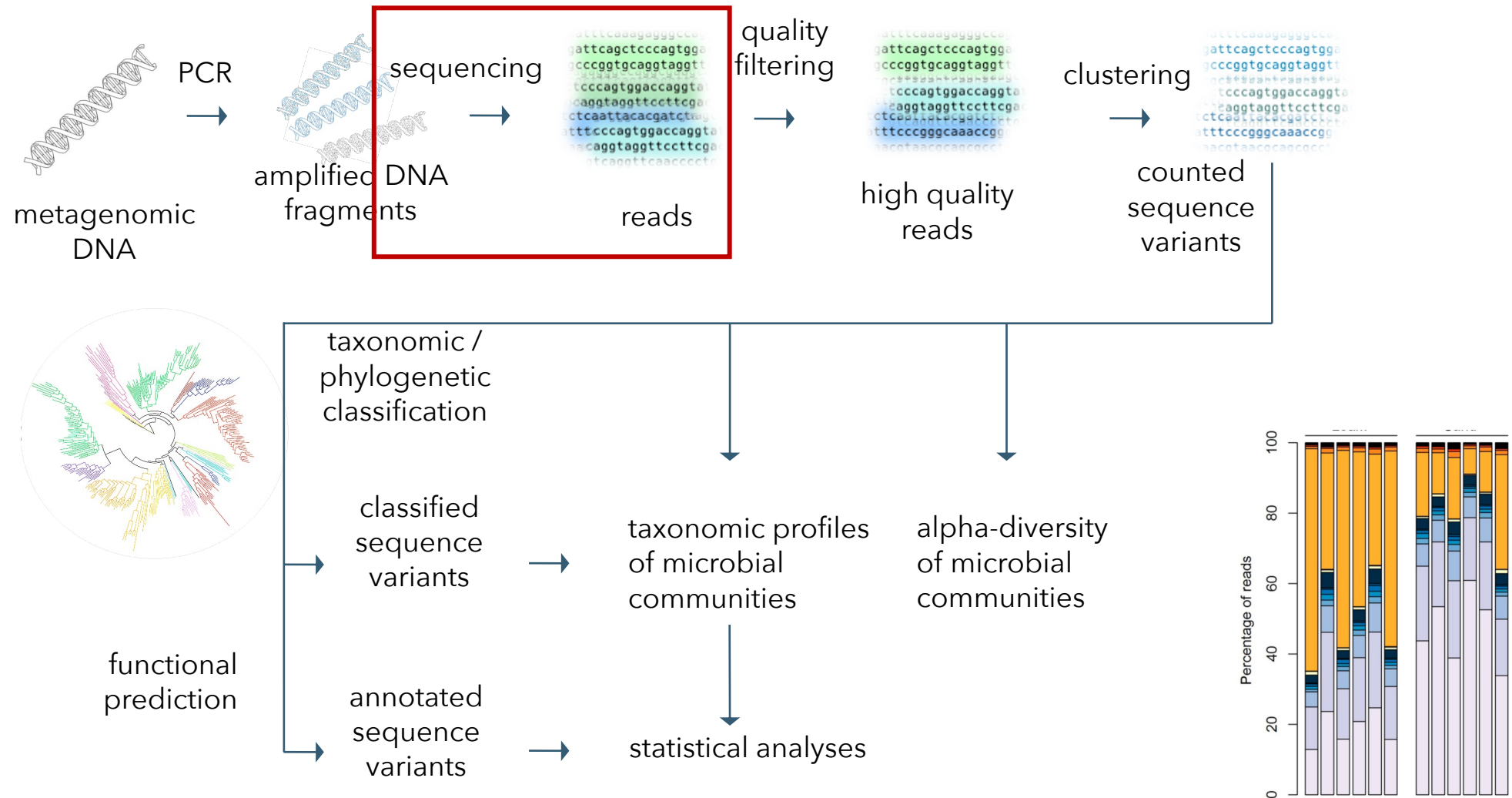
Omic paradigm



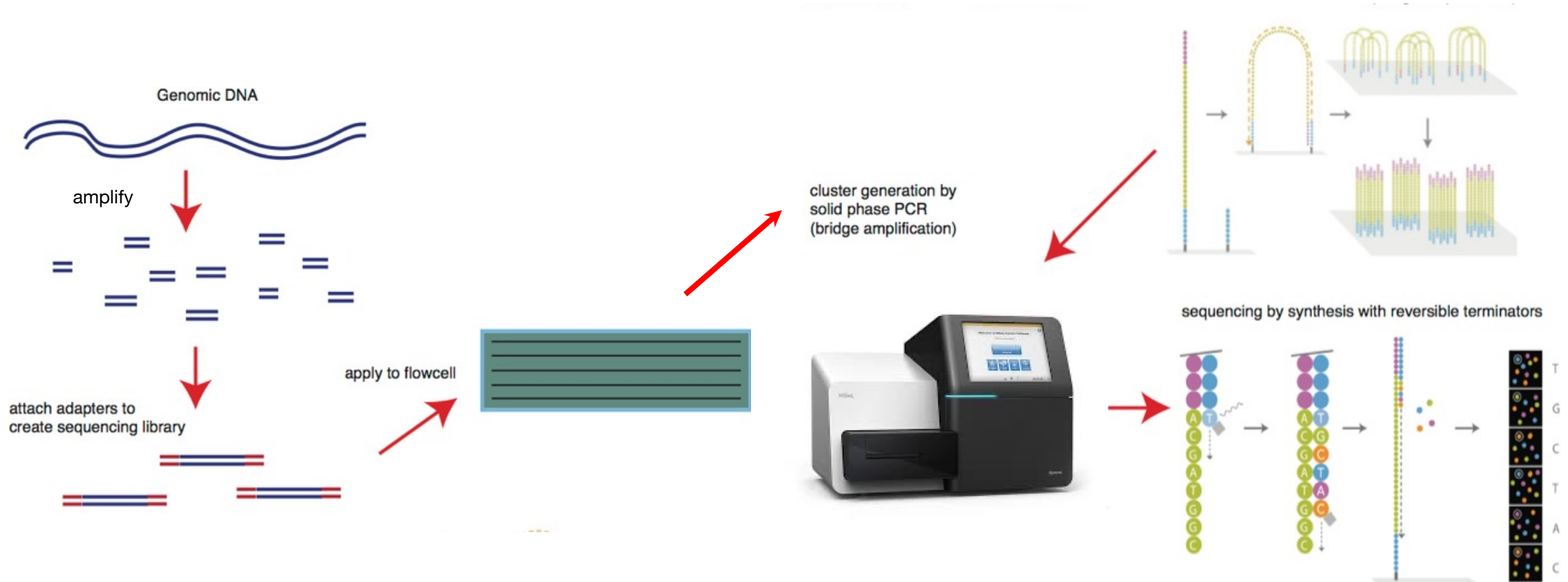
Metabarcoding workflow



Metabarcoding workflow



Measuring microbiomes: sequencing



What does sequencing data look like?

```

@M03696:36:00000000-BGTDB:1:1101:9696:1078 1:N:0:226
+
TTAAGTTCGCAGGGTATTCCTANNNNNNNCAGGTCACCTTAGAAGTAAAGGNAAGGGGAGAACNCCNCN
ACNNNNNNNGGTNNANNNNNNNNNGGCCNNNAGNNNGGANAATGACGNTCGAACAGGCATGCCCTTCGGAATAC
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GF#####/2/##2#####0-0###)2###0/#01<DFF7#)07F:FEF:FFFFFFFFFFFF<BFFBF
@M03696:36:00000000-BGTDB:1:1101:9575:1095 1:N:0:226
+
TTAAGTTCAGCGGGTATCCCTCCCTGATCCGAGGTCGAAACCGAAAGCCCGGAAACGTCGGGGGGTGGCGAGACACC
ANNNNNNNCGCTTGAGG
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
4#####/19CFGC GGDC>#104: ?FFG>DGG33)7>>4FFGFFBFFF@BFFBFFF?BFF>6617<
@M03696:36:00000000-BGTDB:1:1101:16337:1210 1:N:0:226
+
TTAAGTTCAGCGGGTATCCCTACCTGCTCCGCGGTCAAAGGTTGCAAAAAGGCTTGTGGACGCTGACC
AAGCTTGAGGGTACAAATGACGCTCGAACAGGCATGCCCTTTGGAATACCAAGGGCGCAATGTGCGT
+
GGFGGGGGGGGG:FF@FGGGGGGGGGFFGGGFFGGG, EFGGGFAGCFFFGFGGGGGGFBEGEGEGGF
EF??CF6=C6CGFGG*<CGCFB3*.7D?F05>FDCFGGFGGFGGF@7@FF3B2>59?BABFF05
@M03696:36:00000000-BGTDB:1:1101:12879:1372 1:N:0:226
+
TTAAGTTCAGCGGGTATCCCTGCTGATCCGAGGTCAACCGGAAAGACGCGAACGTCGGGGGGTGGCA
ACAAGCCGCCTTGAGGGCAGTAATGACGCTCGGACAGGCATGCCCCCGGAATACCCGGGGGGCAAT
+
GGGGGGGGGGGGGGG@FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@M03696:36:00000000-BGTDB:1:1101:23017:1433 1:N:0:226
+
TTAAGTTCAGCGGGTCTCCCTACCTGATCCGAGGTCAACCTGATAAAAATGGGGGGTTACTGGCAGGCA
GTTGTAATGACGCTCGGACAGGCATGCCCCCGGAATACAGGGGGGGCAATGTGCGTTCAAAATTCG
+
GG<FFGGG9CFDGG7CFGGGGGGG<CEFGGGGGGGGGGGGGGAGGG, CAC87F7@FGG, CDFD:>FC
GDFCGGC+<F: <CDC3CCG5F>3C61730, C5)58CFGF/(*)59)87B377*9*95250?>92:2267
@M03696:36:00000000-BGTDB:1:1101:13635:1436 1:N:0:226
+
TTAAGTTCAGCGGGTAACTCCTGATTTGAGGTGAGGTTGTAAGTTGTTGTGTTGTTGTTGTTGTTGTTGTT
AATCCATGATCCAAGCCATCAGGTTAATAAAAACTTGATAGTTGAGAATTTAATGACACTCAAACA
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGEEGGGGG9C>DGEggggF6CFDBB6@GFFGG4DGFAGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@M03696:36:00000000-BGTDB:1:1101:14775:1592 1:N:0:226
+
TTAAGTTCAGCGGGTATCCCTGCTGCTCCGCGTCAACCGAACGACGCAAGTCGGGGGGTCCGGCA
ACAAGCCGCCTTGAGGGCAGTAATGACGCTCGGACAGGCATGCCCCCGGAATACCAAGGGGGGGCCAT
+
GGFEFGGGGGGGGGGGGGGGGGGGGGGG, FGGGGGGFFG<FGGGGG7C87CFEGEGGGGGGGGGGGGGGGGG
FDGGGGGG/53<<7ECGF?D7FFF7D?3>5:*95<FFFFFFFFFBF>>37>?FF7?@:3>0(4(70

```

Line 1: Name

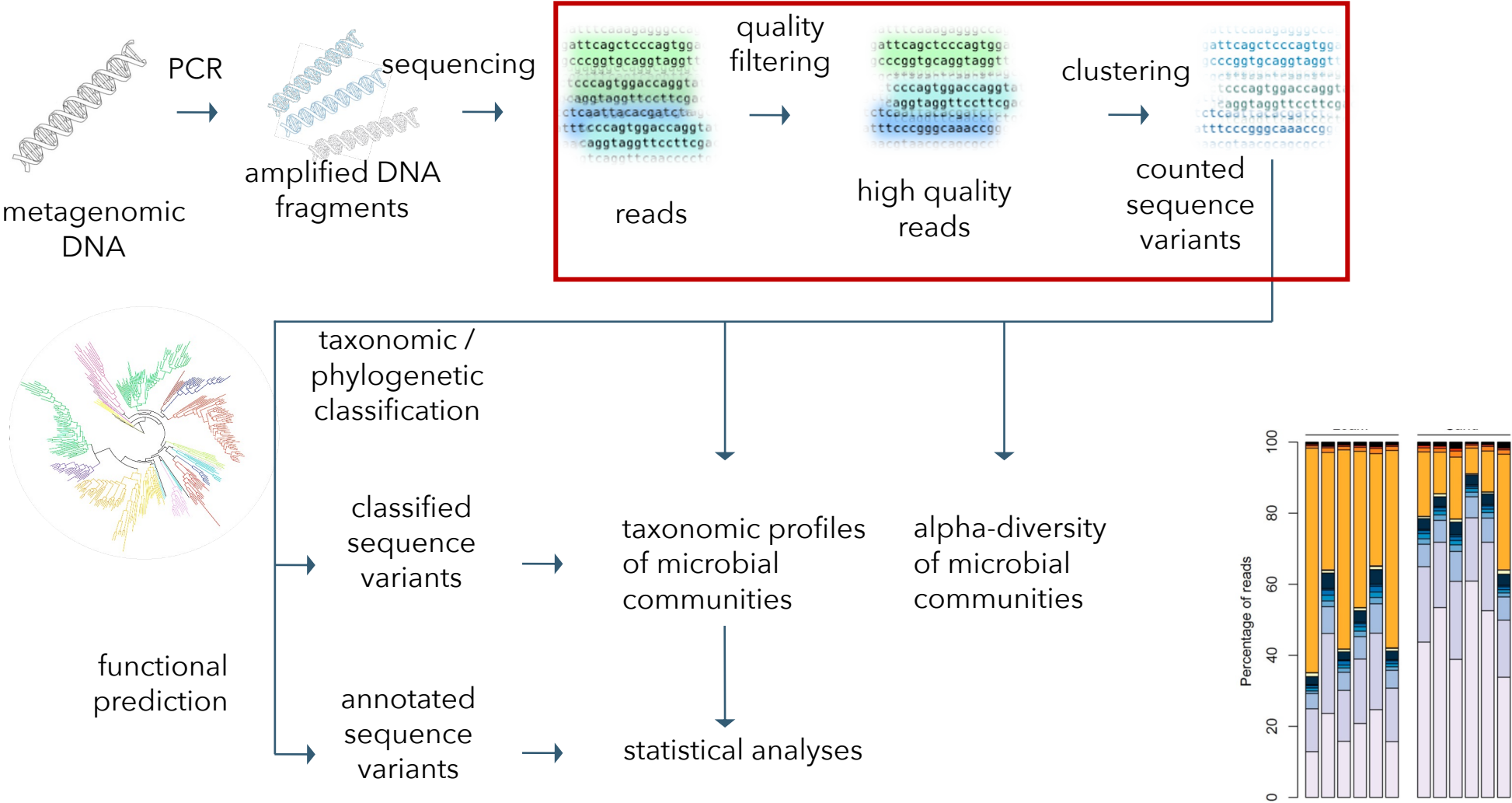
Line 2: Sequence

Line 3: anything

Line 4: Quality at each position

as many as we have reads (forward- & reverse files)

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

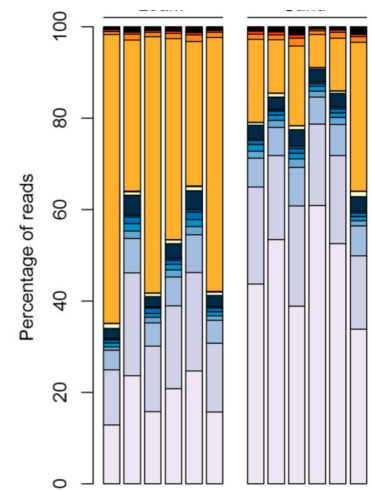
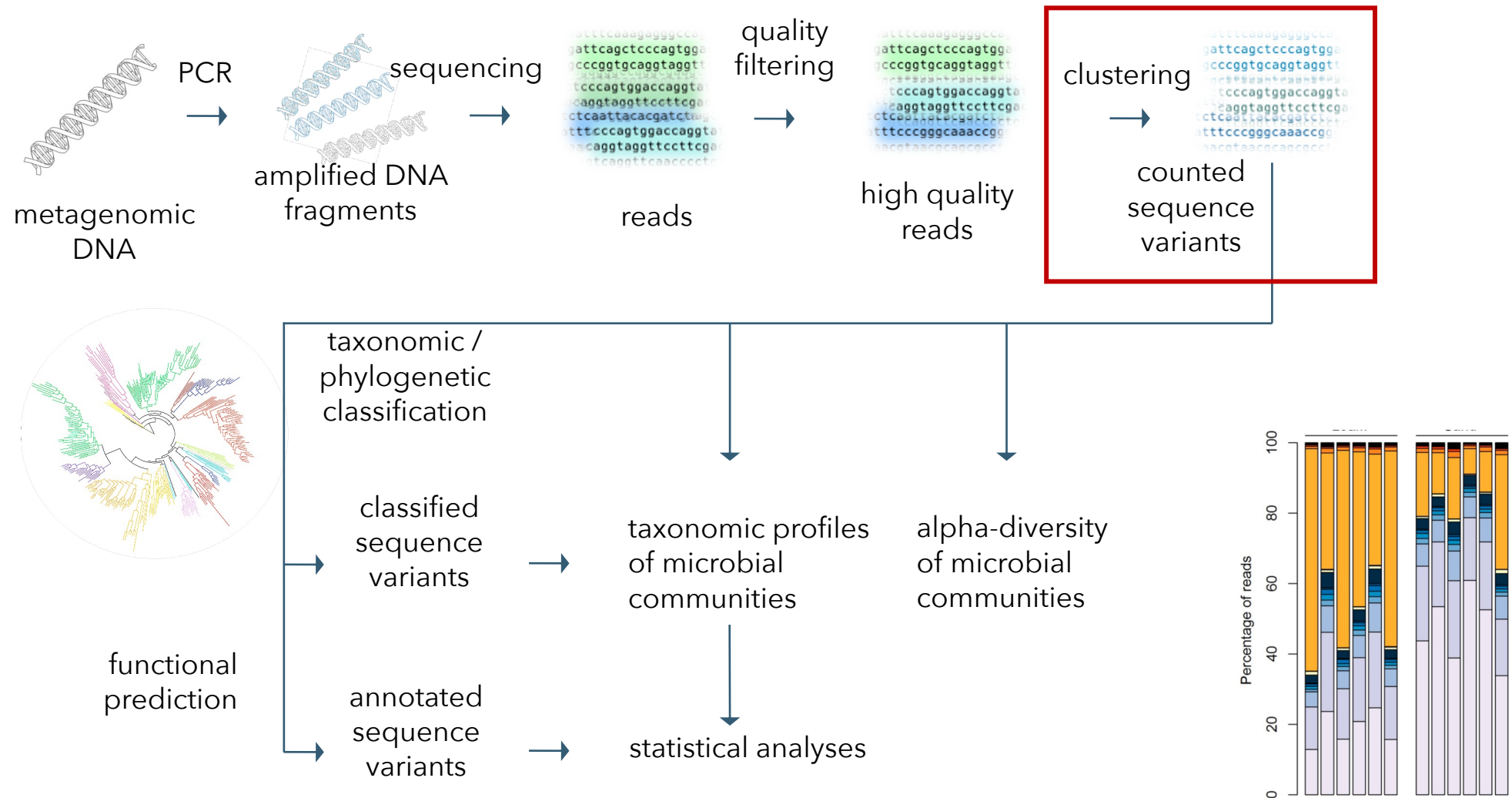
What's "wrong" with sequencing data?

- 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 \neq 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 **O**TUs

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



Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns

Amnon Amir,^a Daniel McDonald,^a Jose A. Navas-Molina,^{a,c} Evguenia Kopylova,^a James T. Morton,^a Zhenjiang Zech Xu,^a Eric P. Kightley,^b Luke R. Thompson,^a Embriette R. Hyde,^a Antonio Gonzalez,^a Rob Knight^{a,c,d}

Department of Pediatrics, University of California San Diego, La Jolla, California, USA^a; Department of Applied Mathematics, and Interdisciplinary Quantitative Biology Graduate Program, University of Colorado Boulder, Boulder, Colorado, USA^b; Department of Computer Science and Engineering, University of California, San

BRIEF COMMUNICATIONS

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan¹, Paul J McMurdie², Michael J Rosen³, Andrew W Han², Amy Jo A Johnson² & Susan P Holmes¹

We previously introduced the Divisive Amplicon Denoising Algorithm (DADA), a model-based approach for correcting amplicon errors without constructing OTUs⁵. DADA identified fine-scale variation in 454-sequenced amplicon data while outputting few false positives^{2,5}.

Here we present DADA2, an open-source R package (<https://github.com/benjjneb/dada2>, **Supplementary Software**) that extends and improves the DADA algorithm. DADA2 implements a new quality-aware model of Illumina amplicon errors. Sample composition is inferred by dividing amplicon reads into partitions consistent with the error model (Online Methods). DADA2 is reference-free and

What's new in USEARCH v9

See also

[What's new in version 9.1](#)

New algorithms

[UNOISE error-correction \(denoising\)](#) paper: <http://dx.doi.org/10.1101/081257>

[SINTAX taxonomy prediction](#) paper: <http://dx.doi.org/10.1101/074161>

[UCHIME2 chimera detection](#) paper: <http://dx.doi.org/10.1101/074252>

ASVs (aka. ESVs, zOTUs)

- **A**mplicon **S**equence **V**ariants
= **E**xact **S**equence **V**ariants
= **z**ero-radius **O**TUs

- 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. Sanders¹, Daniel McDonald¹, Amnon Amir¹, Joshua Ladau⁴, Kenneth J. Locey⁵, Robert J. Prill⁶, Anupriya Tripathi^{1,7,8}, Sean M. Gibbons^{9,10}, Gail Ackermann¹, Jose A. Navas-Molina^{1,11}, Stefan Janssen¹, Evguenia Kopylova¹, Yoshiki Vázquez-Baeza^{1,11}, Antonio González¹, James T. Morton^{1,11}, Siavash Mirarab¹², Zhenjiang Zech Xu¹, Lingjing Jiang^{1,13}, Mohamed F. Haroon¹⁴, Jad Kanbar¹, Qiyun Zhu¹, Se Jin Song¹, Tomasz Kosciółek¹, Nicholas A. Bokulich¹⁵, Joshua Lefler¹, Colin J. Brislawn¹⁶, Gregory Humphrey¹, Sarah M. Owens¹⁷, Jarrad Hampton-Marcell^{17,18}, Donna Berg-Lyons¹⁹, Valerie McKenzie²⁰, Noah Fierer^{20,21}, Jed A. Fuhrman²², Aaron Clauset^{19,23}, Rick L. Stevens^{24,25}, Ashley Shade^{26,27,28}, Katherine S. Pollard⁴, Kelly D. Goodwin³, Janet K. Jansson¹⁶, 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

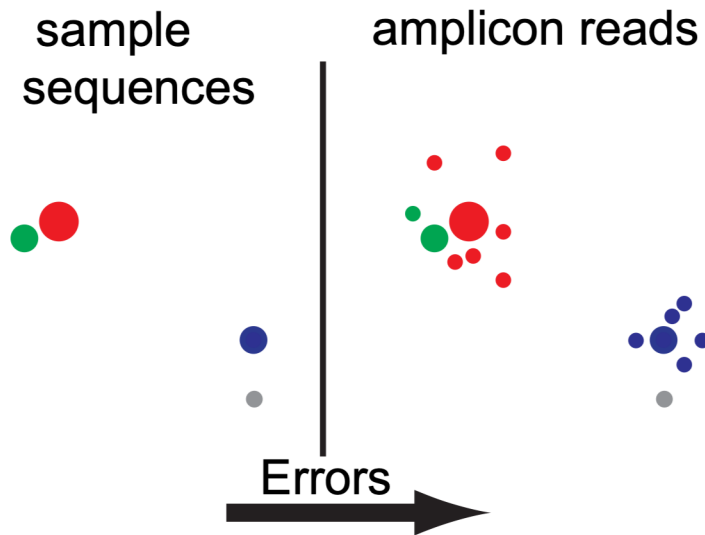
PERSPECTIVE

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

Benjamin J Callahan¹, Paul J McMurdie² and Susan P Holmes³
¹Department of Population Health and Pathobiology, NC State University, Raleigh NC, USA; ²Whole Biome Inc, San Francisco CA, USA and ³Department of Statistics, Stanford University, Stanford CA, USA

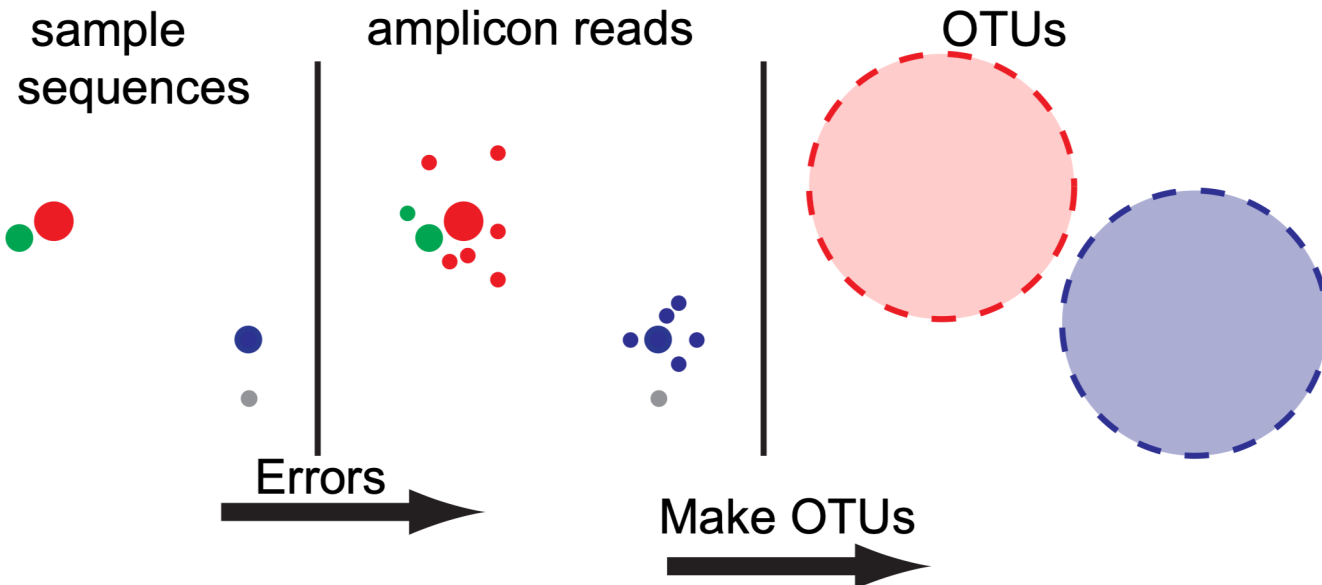
ASVs (aka. ESVs, zOTUs)

- **A**mplicon **S**equence **V**ariants
= **E**xact **S**equence **V**ariants
= **z**ero-radius **O**TUs



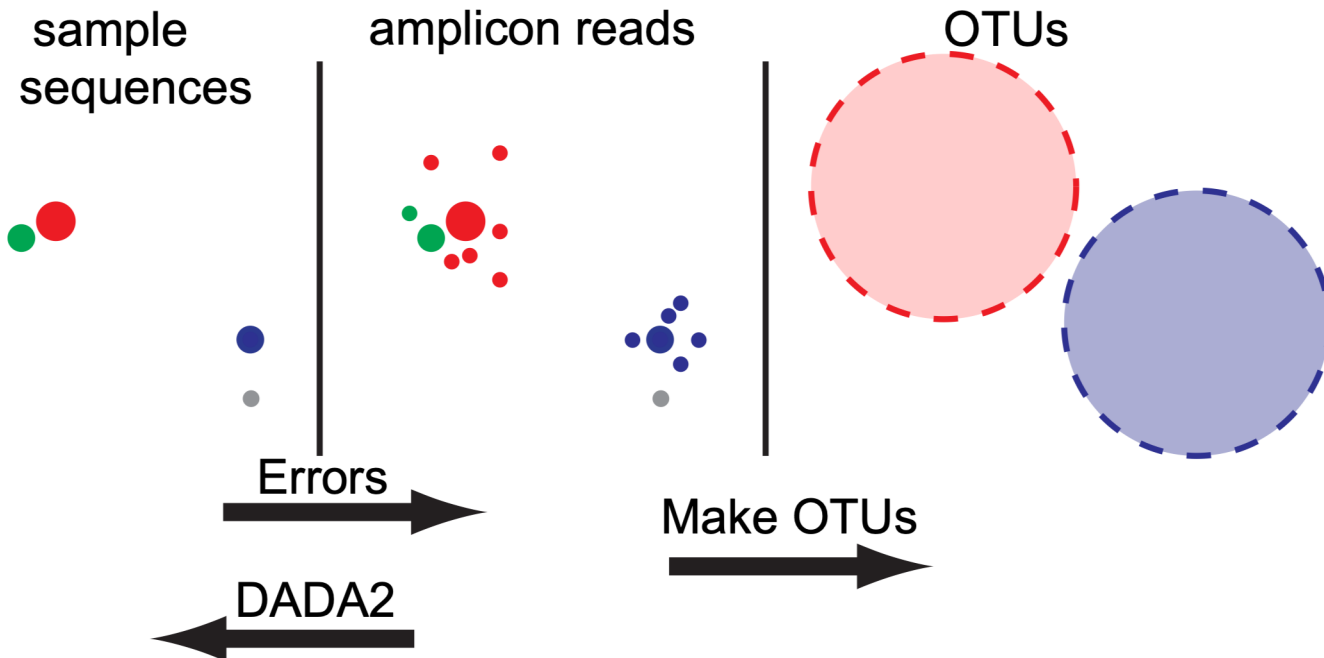
ASVs (aka. ESVs, zOTUs)

- **A**mplicon **S**equence **V**ariants
= **E**xact **S**equence **V**ariants
= **z**ero-radius **O**TUs

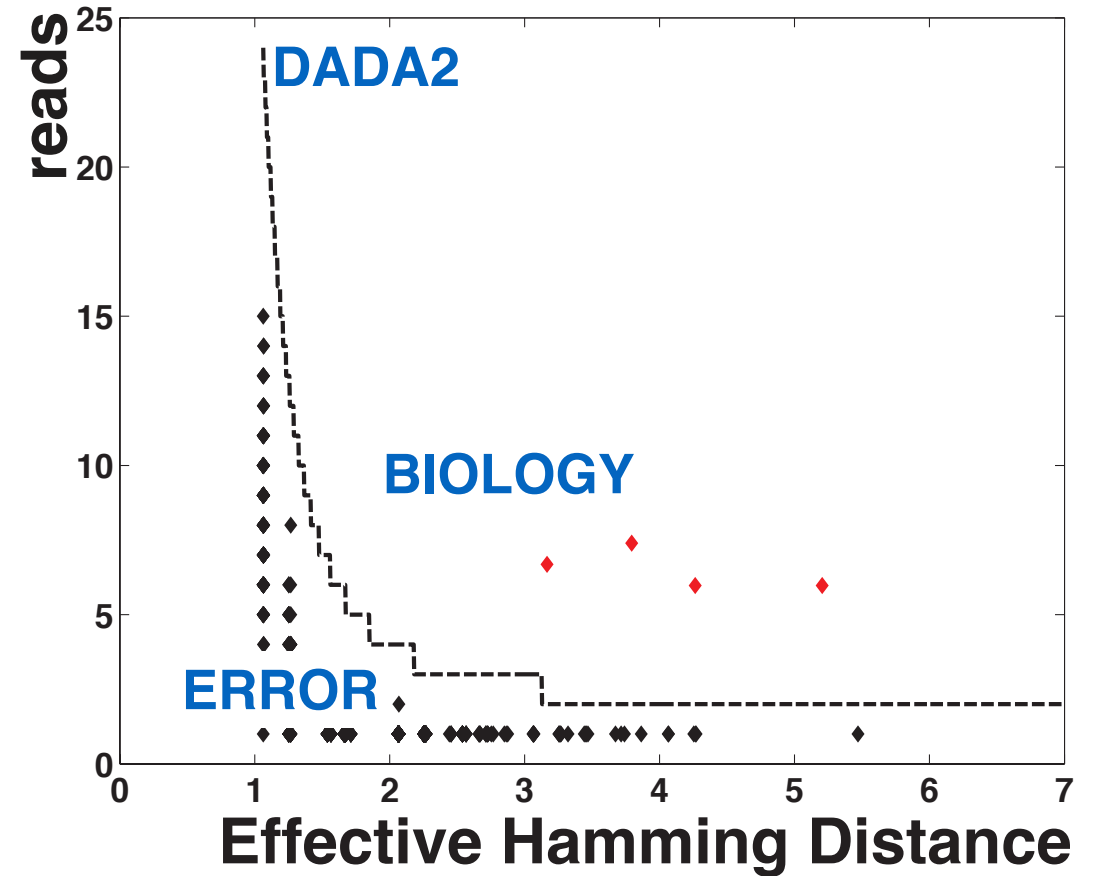
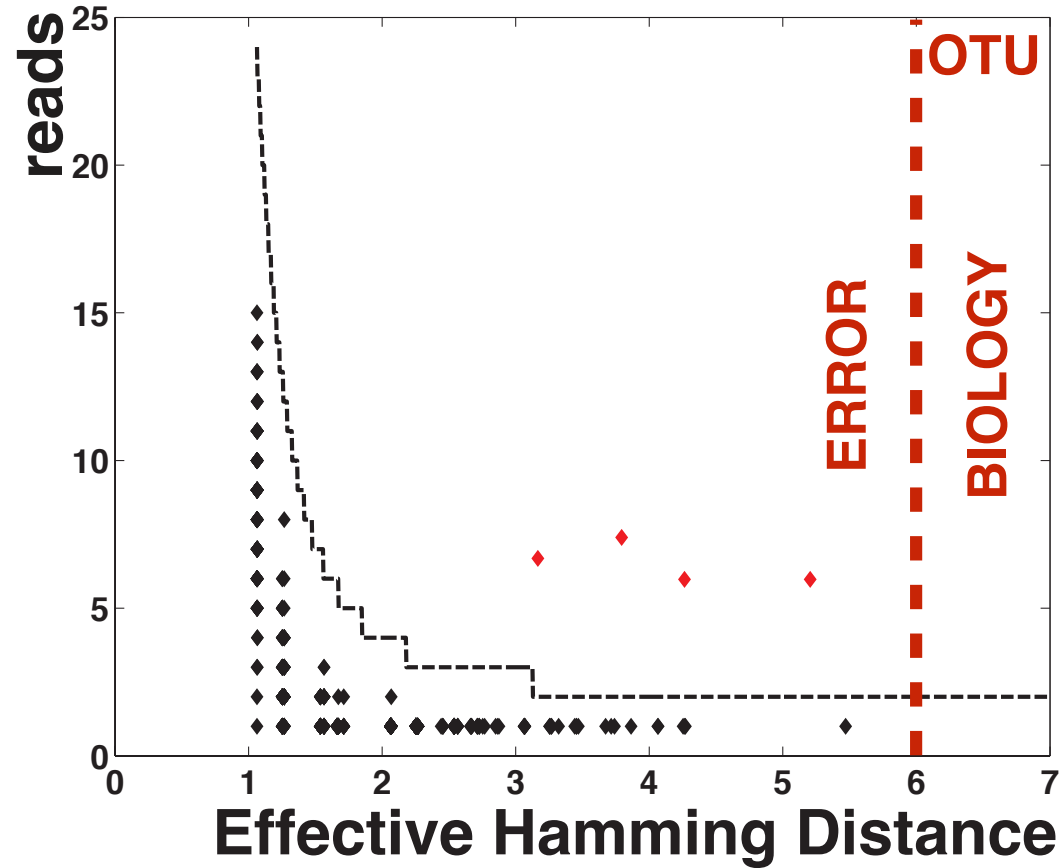


ASVs (aka. ESVs, zOTUs)

- **A**mplicon **S**equence **V**ariants
= **E**xact **S**equence **V**ariants
= **z**ero-radius **O**TUs



ASVs (aka. ESVs, zOTUs)



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)

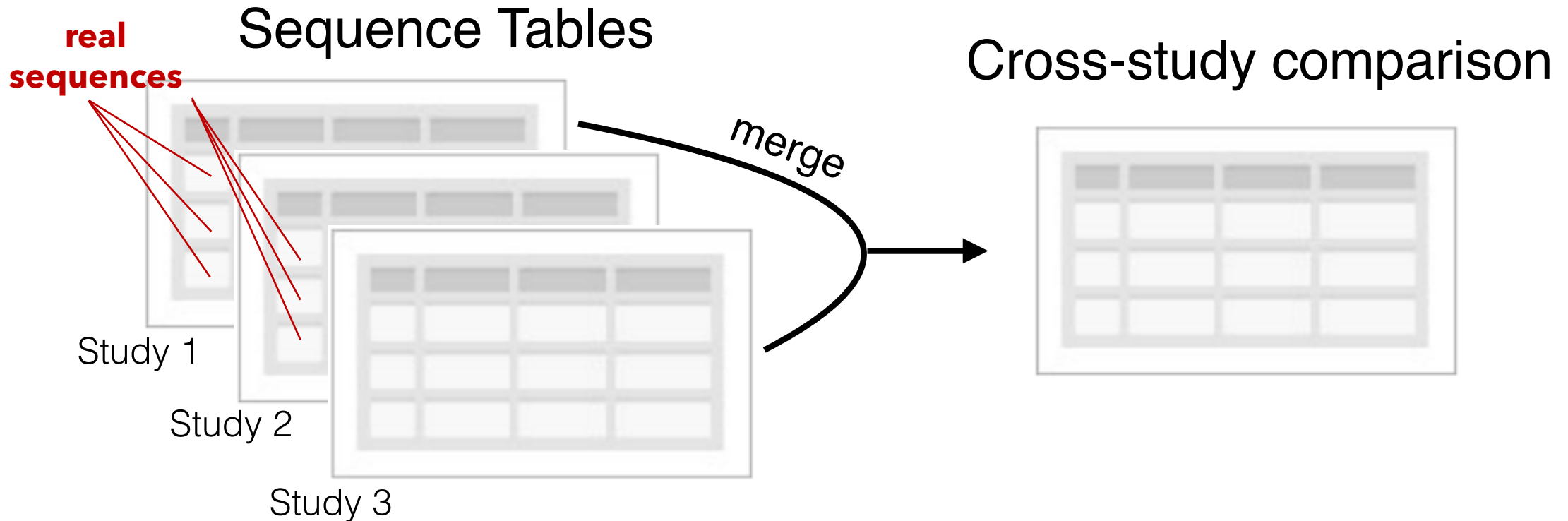
ASVs (aka. ESVs, zOTUs)

- + 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



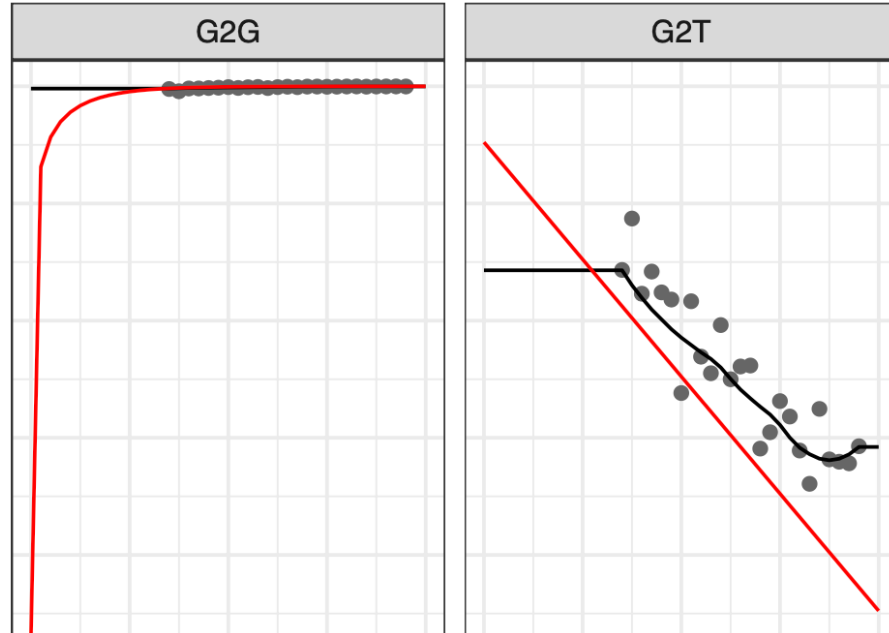
ASVs (aka. ESVs, zOTUs)

- cross-study comparability



ASVs (aka. ESVs, zOTUs)

- o model substitutions for every run



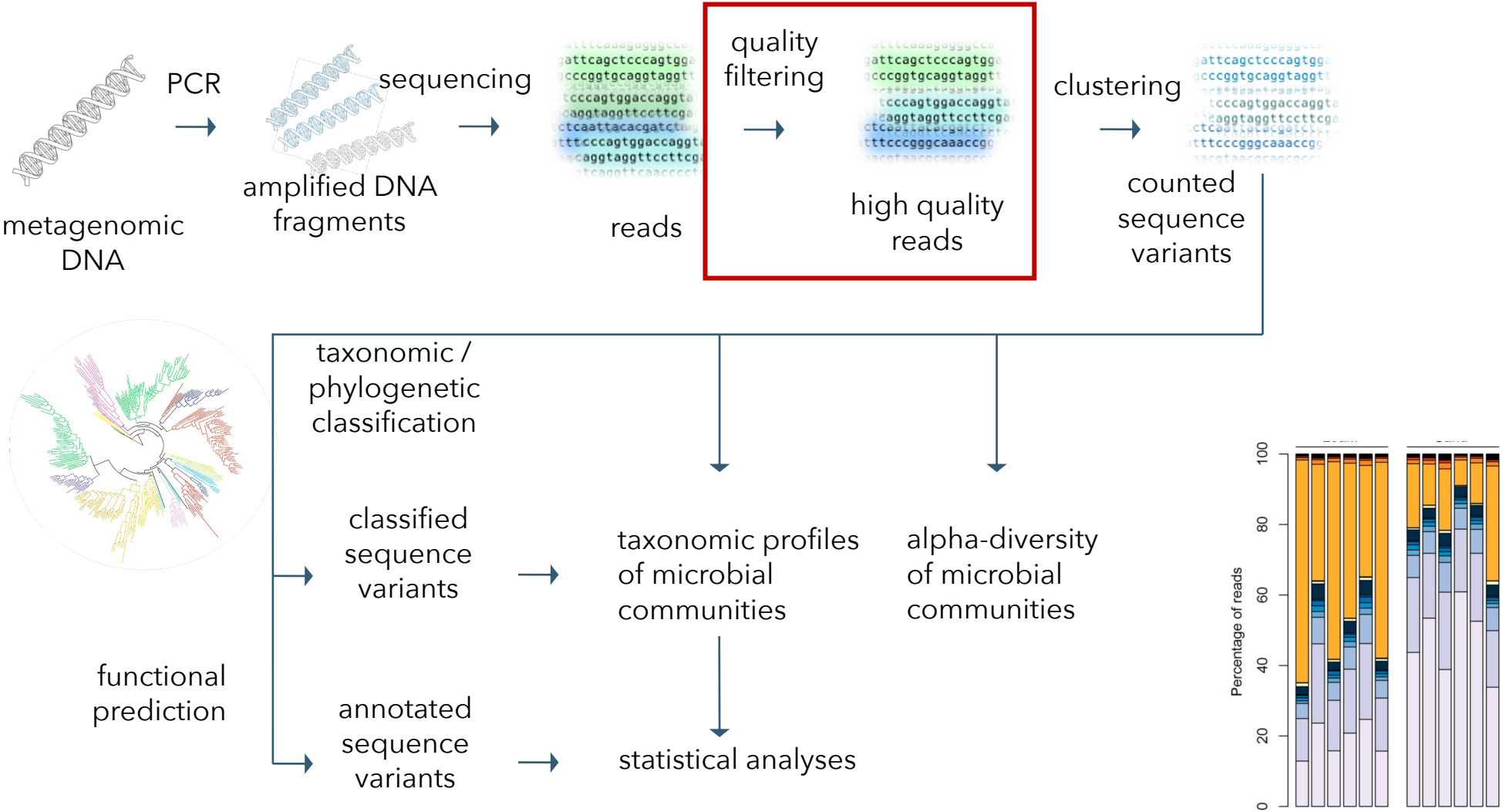
s: ATTAACGAGATTATAACCAGAGTACGAATA...
| |
r: ATCAACGAGATTATAACAAGAGTACGAATA...

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

Error rates depend on....

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

Metabarcoding workflow



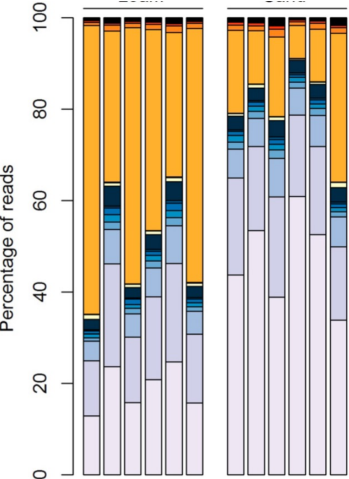
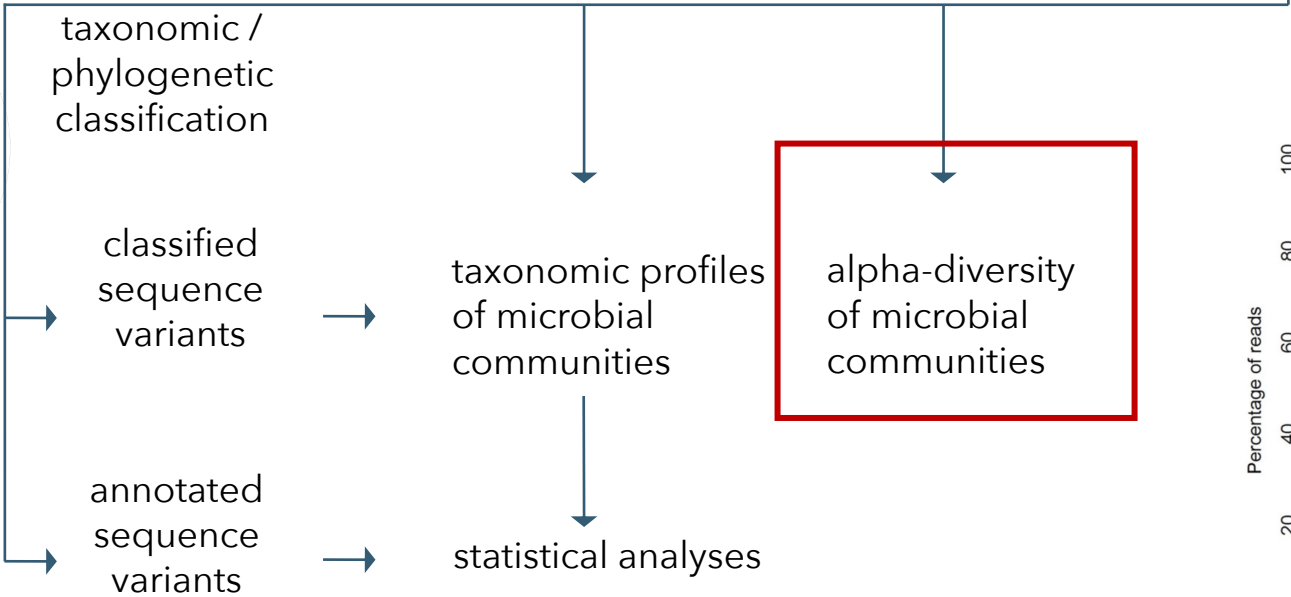
Quality filtering

- 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



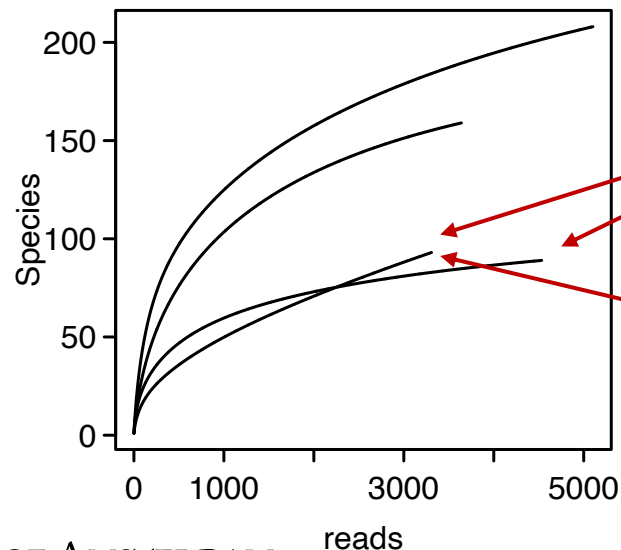
functional prediction



Unequal sampling depth does not reflect biology

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

without:

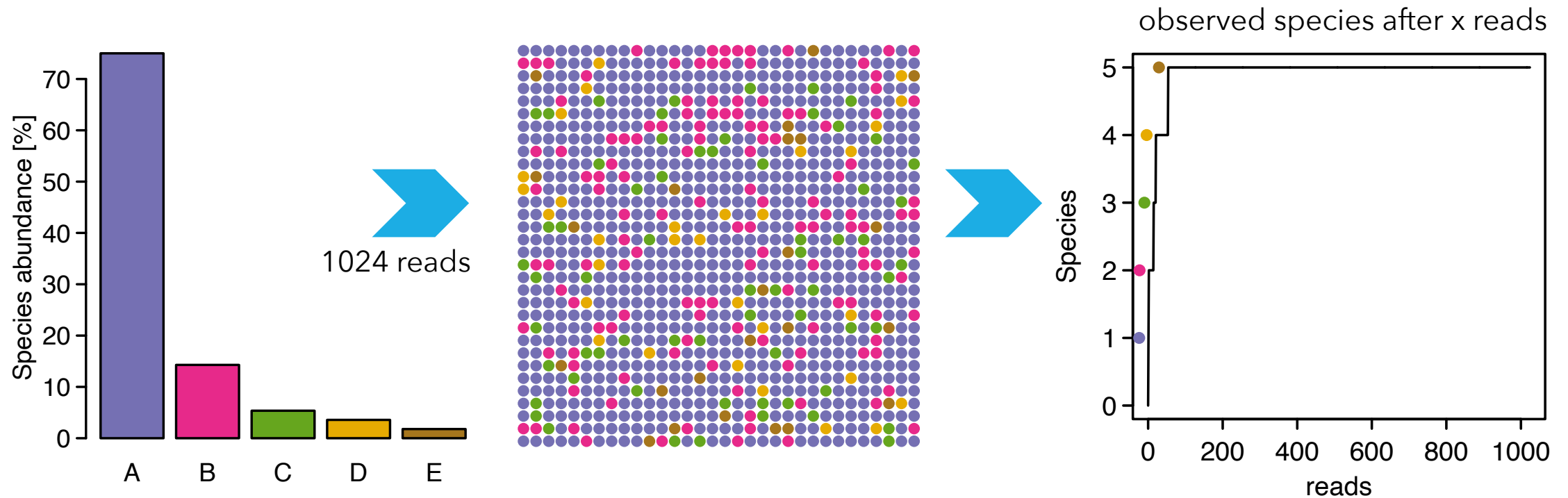


these two seem to have the same number of species

but this one would certainly outgrow the other

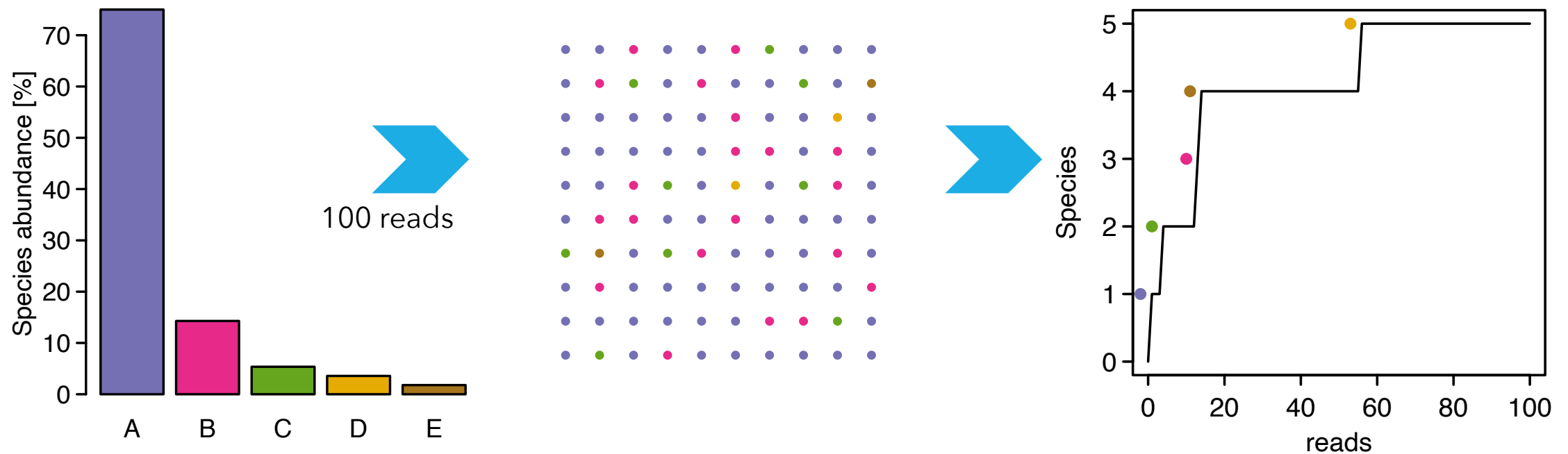
Rarefaction curves & analysis

- number of new species in increasing subsamples



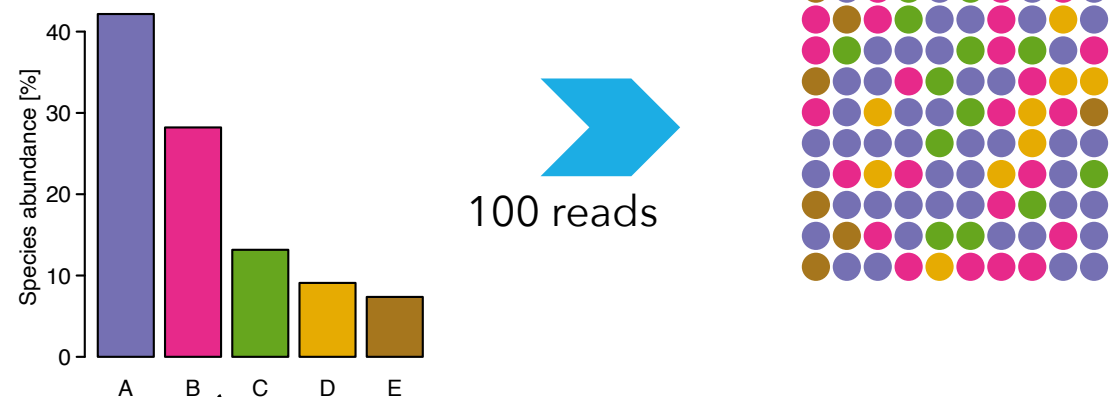
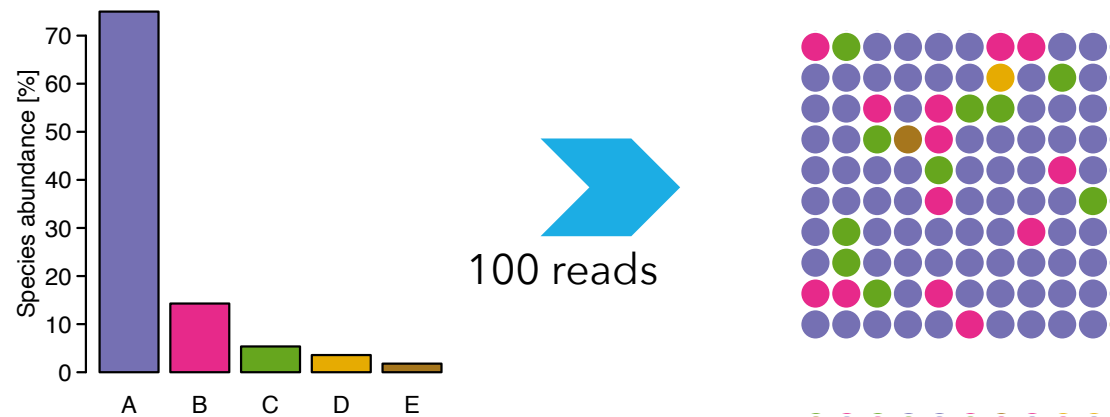
Rarefaction curves & analysis

- number of new species in increasing subsamples



Rarefaction (normalisation)

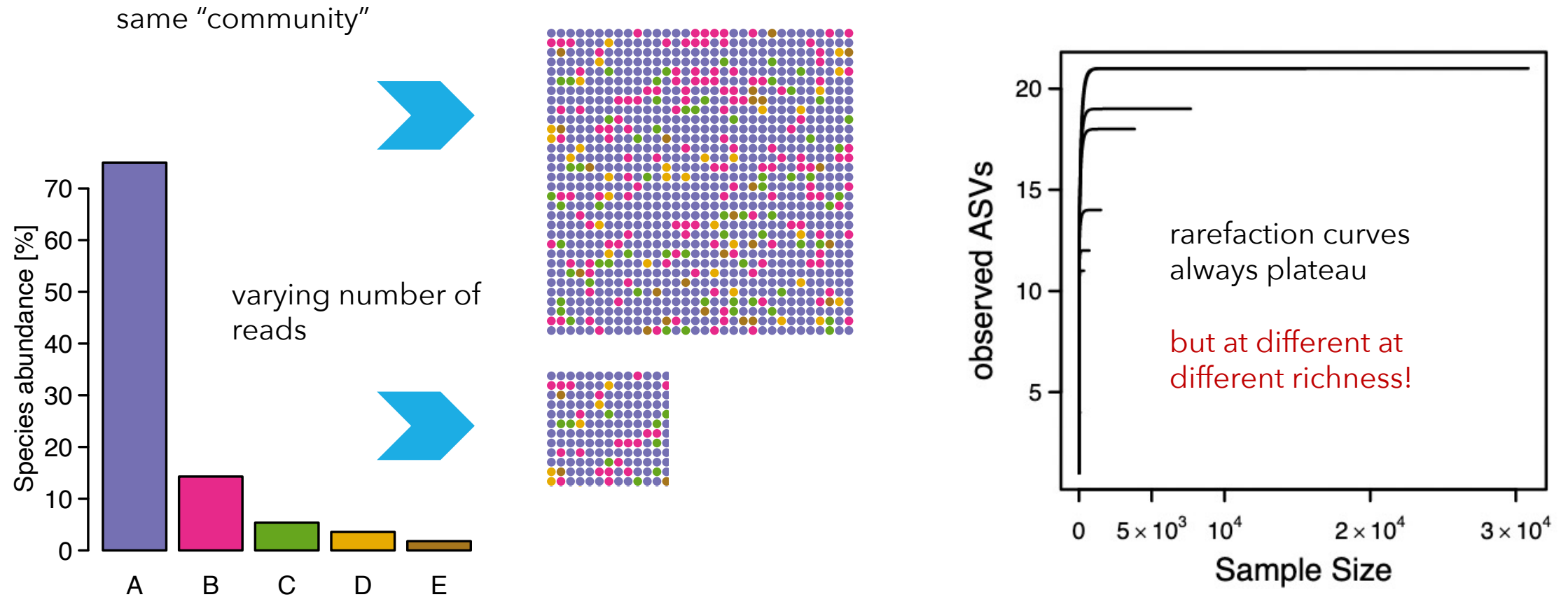
- subsample reads to keep equal numbers per sample



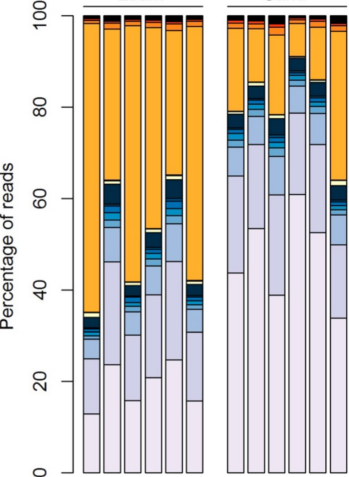
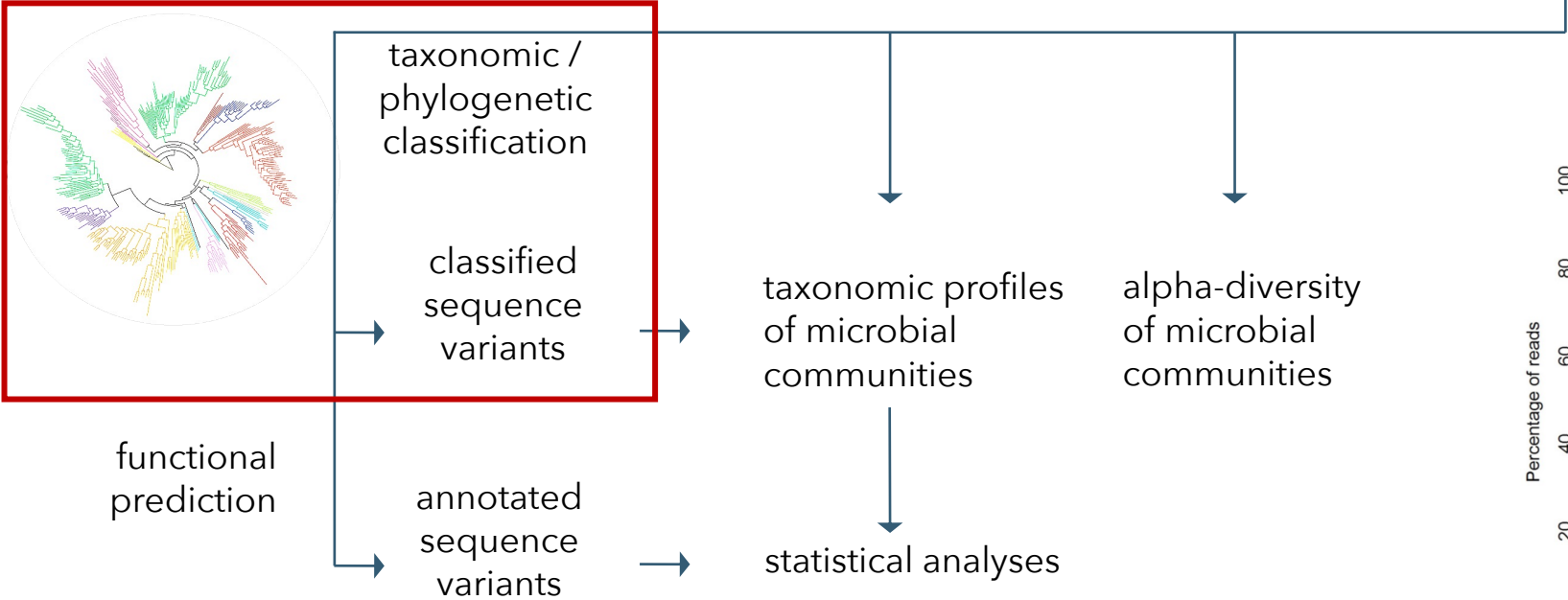
+ all samples have the same sum of reads

! most people do this after ASV generation

post-ASV rarefaction does not work

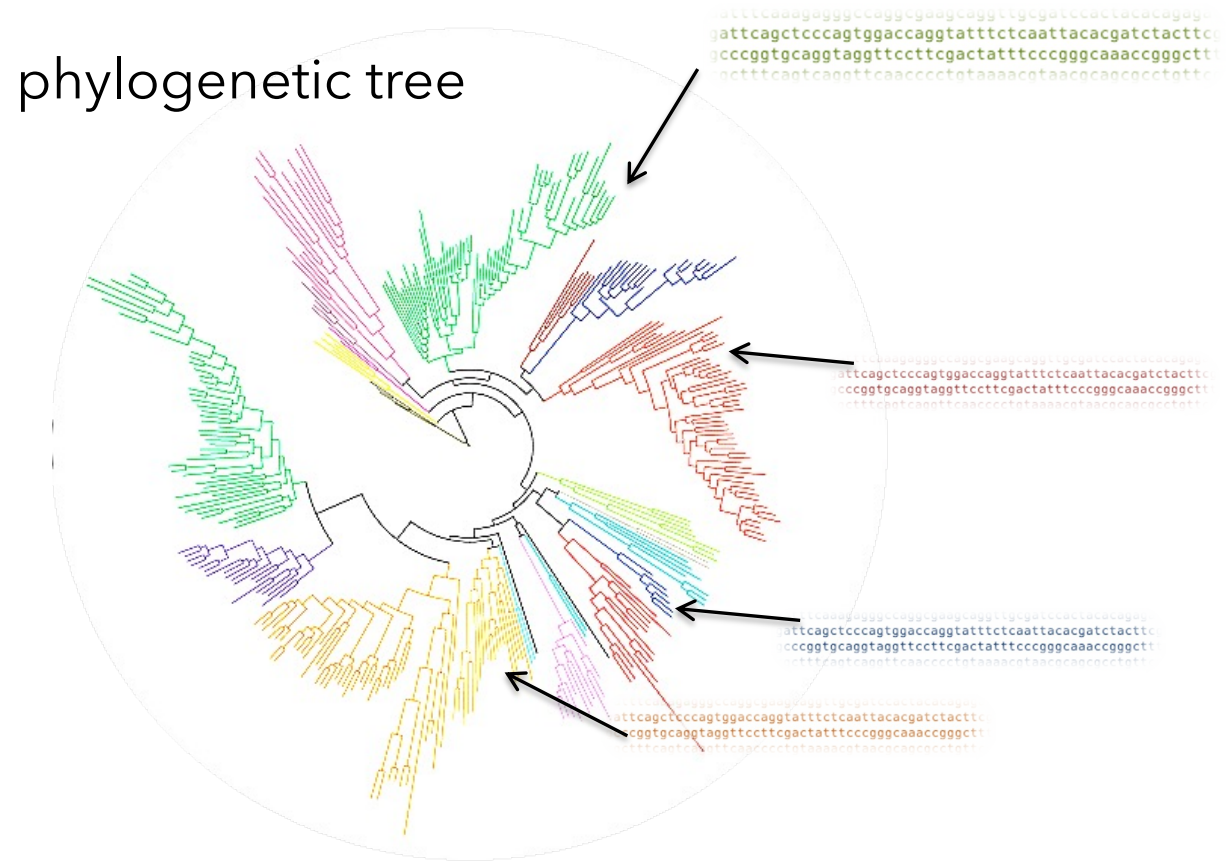


Metabarcoding workflow



Taxonomic classification of ASVs

compare the
taxa
in the database
to your ASVs



Wu et al. 2009 Nature

Naïve Bayesian Classifier

$$P(\text{taxon}|\text{sequence composition}) = P(\text{sequence composition}|\text{taxon}) * P(\text{taxon}) / P(\text{sequence composition})$$

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

the probability that the sequence composition is correct given the taxon

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

Taxonomic classification of ASVs

sequence of one ASV/taxon:

```
aatttcaaagagggccagggcgaagcaggttgcgatccactacacagagac  
gattcagctcccagtgaccaggtatttctcaattacacgatctacttcg  
gcccgggtgcaggtagggtccttcgactatttccggggcaaaccgggcttt  
cgctttcagtcaggttcaaccctgtaaaacgtaacgcagcgcctgttcg
```

aatttcaa

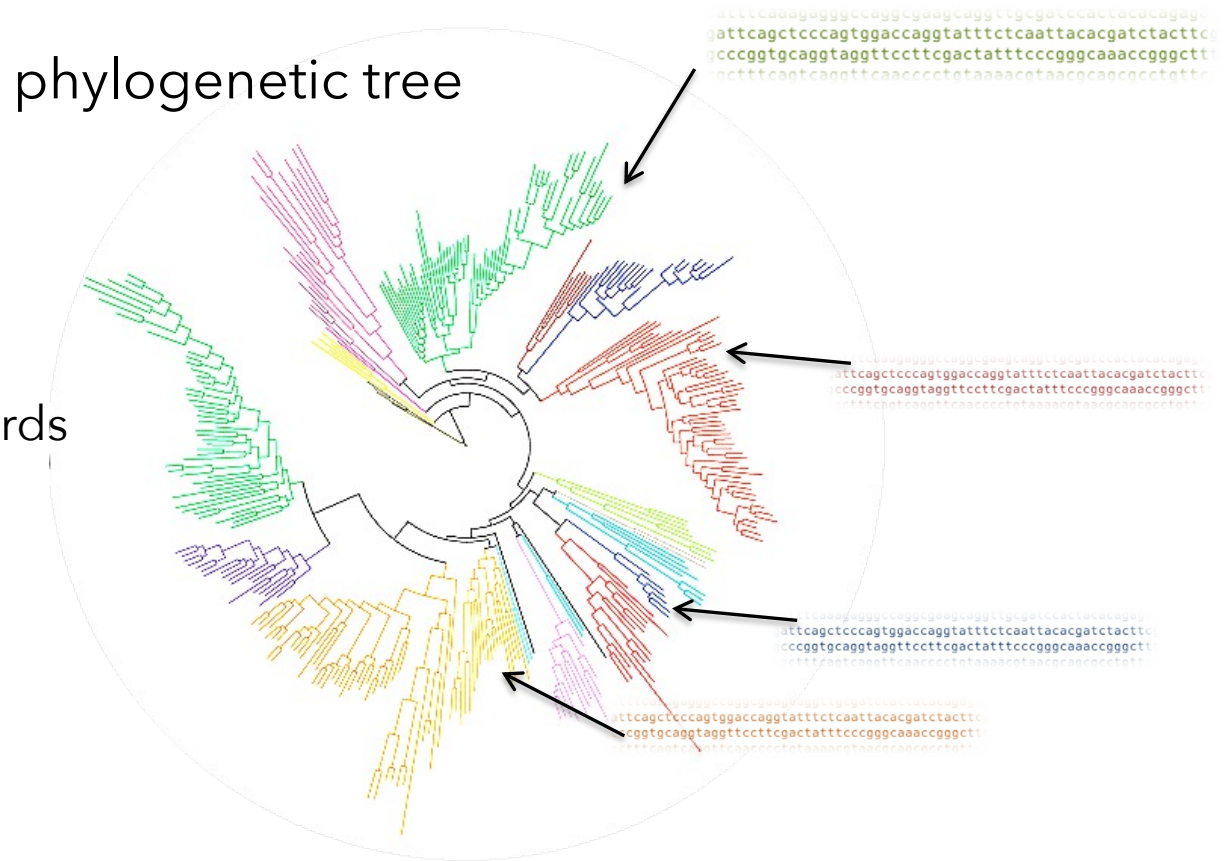
atttcaaa

...

65,536 possible words

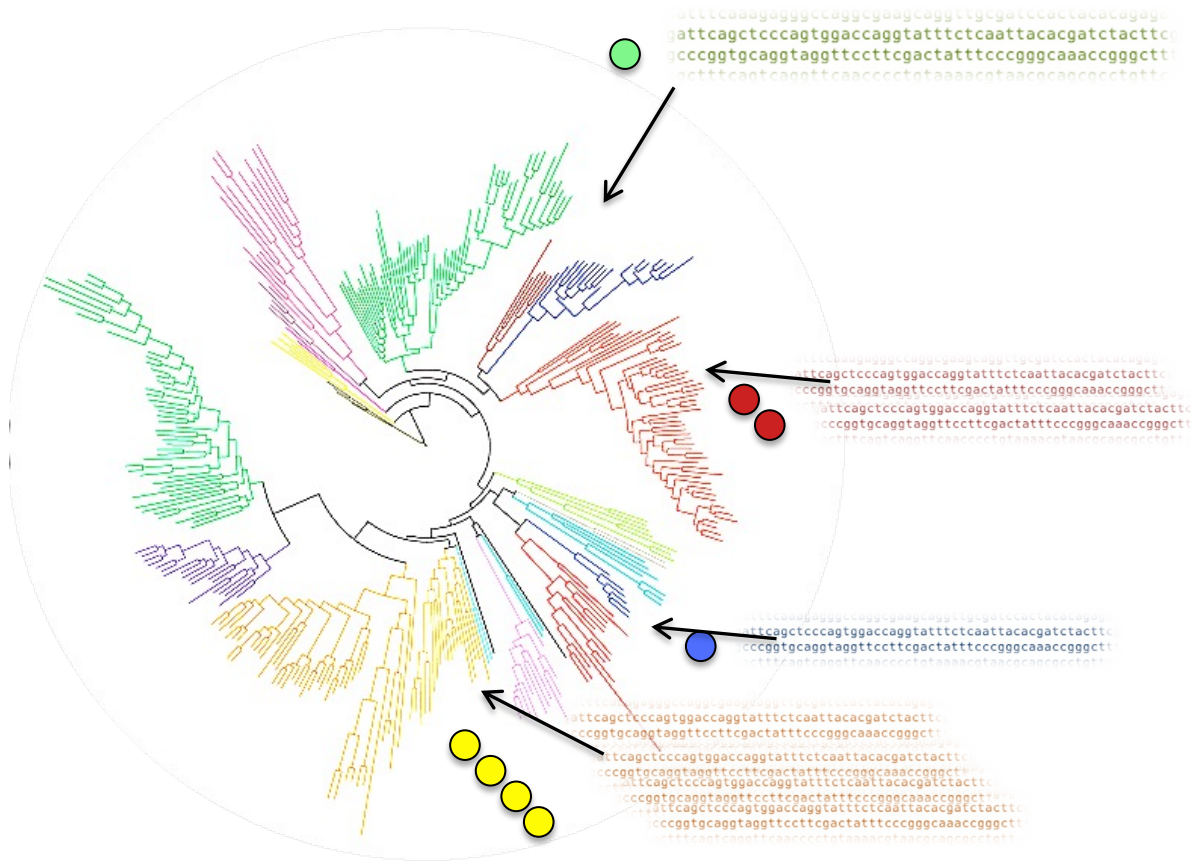
Taxonomic classification of ASVs

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

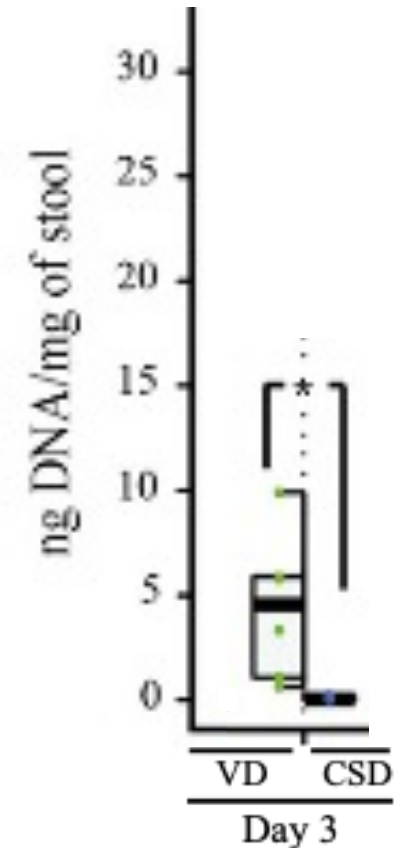
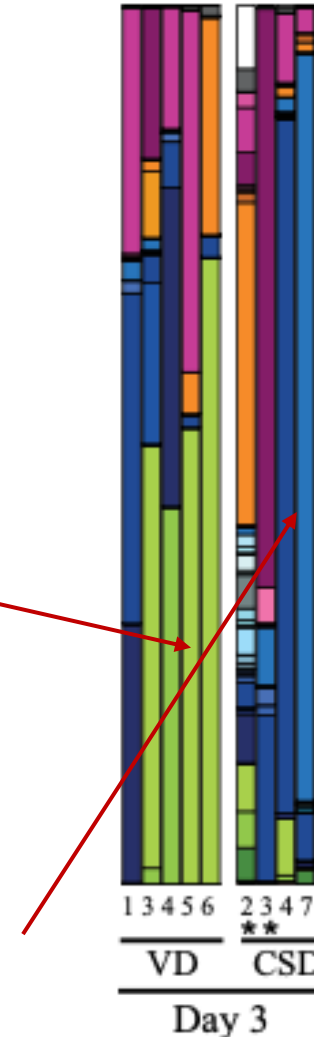
The problems with always counting reads out of a total

- 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?

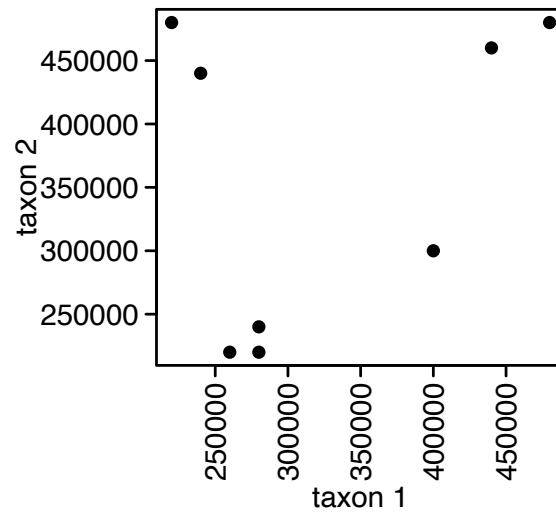
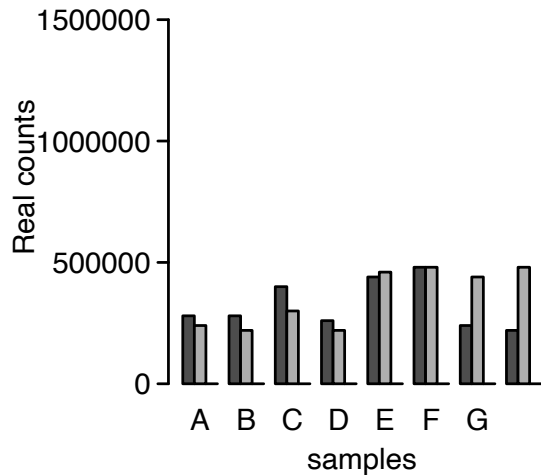
OR

are these a lot more abundant in here?

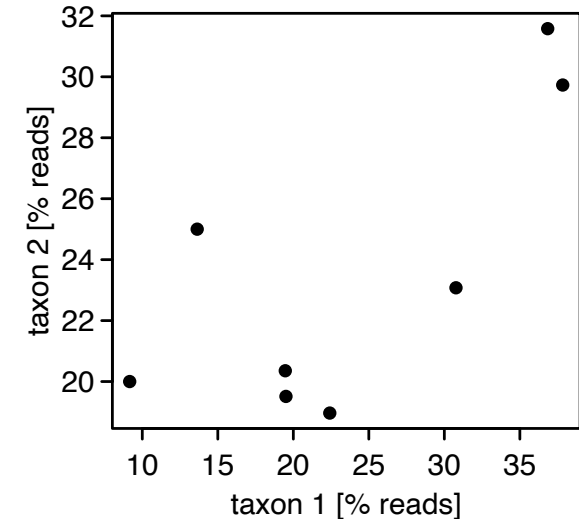
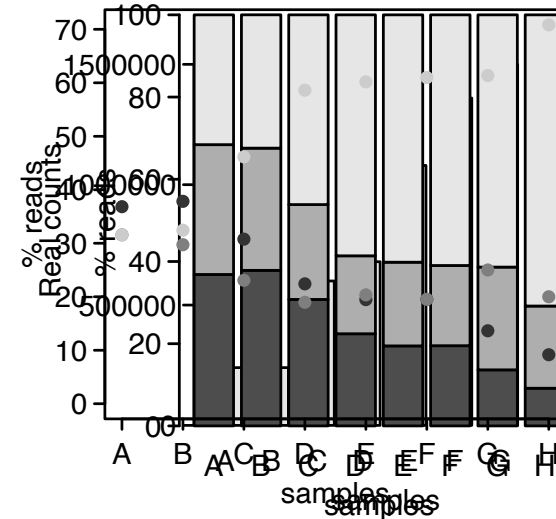


Compositionality - example

- taxa 1 and 2 have no special relationship

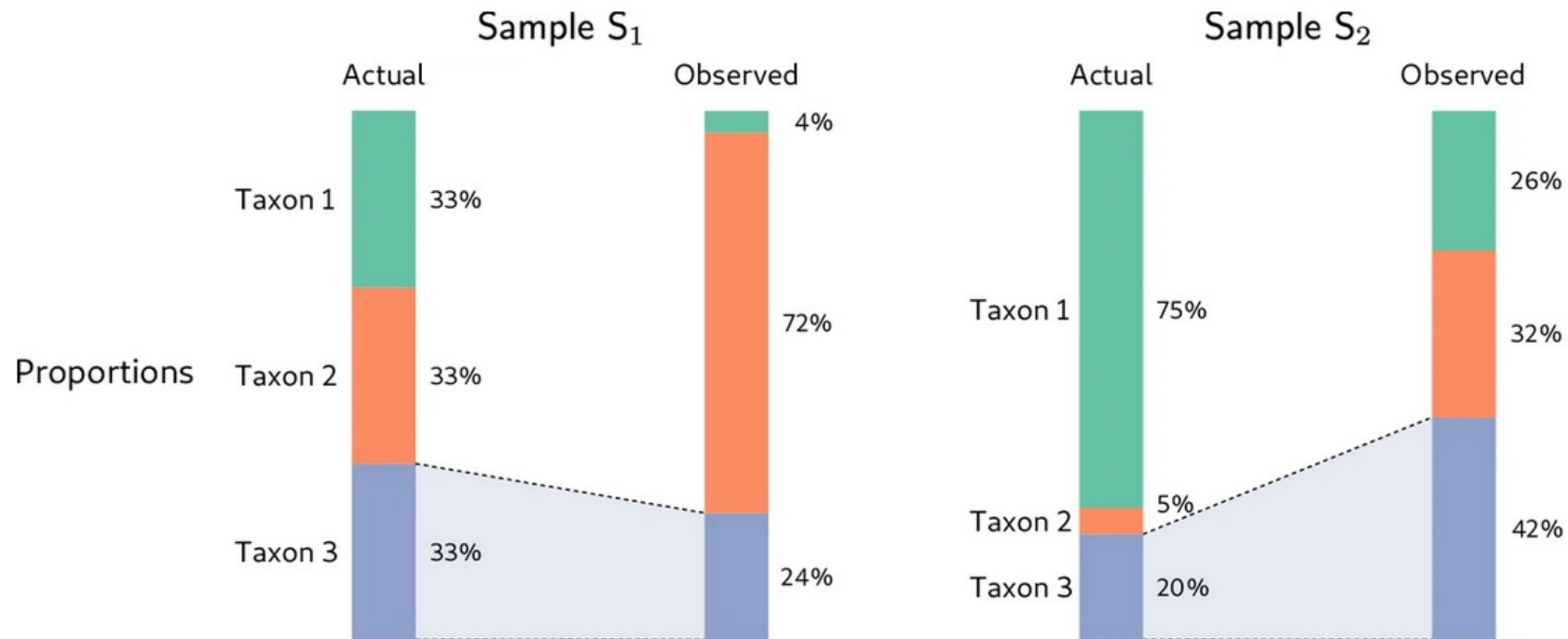


- taxon 3 **introduces a positive correlation**



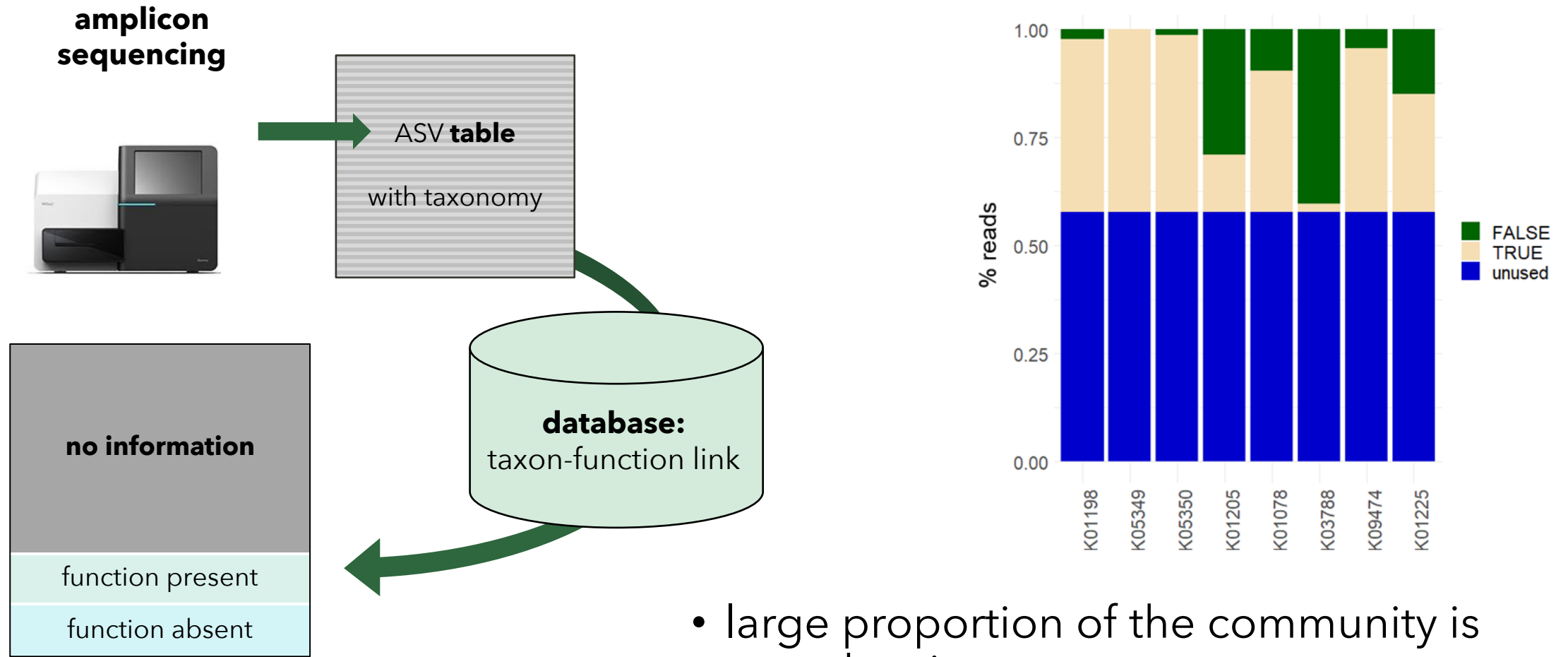
Add to that: bias

- assuming a constant bias for every ASV/taxon:



- you can transform data and correct for biases

A word on functional prediction

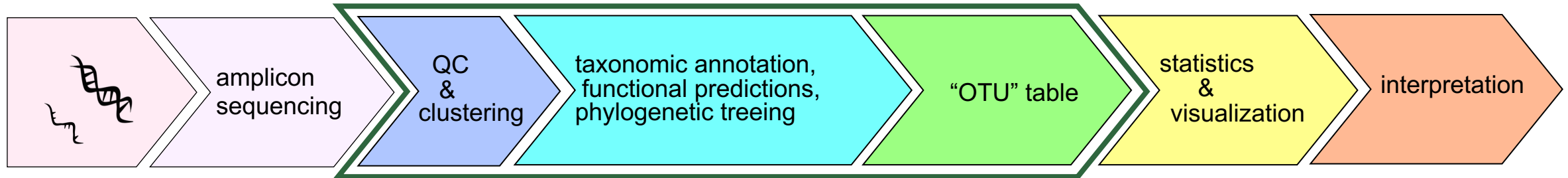


- large proportion of the community is not taken into account
- usefulness depends on the context

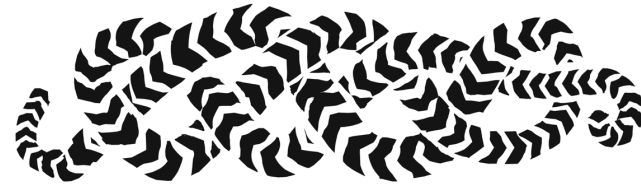
Summary

- **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
- **metabarcoding-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_

