# Analysis of next generation sequencing data for microbial diversity studies

13.06.2017

Christiane Hassenrück, PhD
Tropical Marine Microbiology
Leibniz Centre for Tropical Marine Research
*christiane.hassenrueck@leibniz-zmt.de*

University of the Philippines
Marine Science Institute

ZMT LEIBNIZ CENTRE for Tropical Marine Research

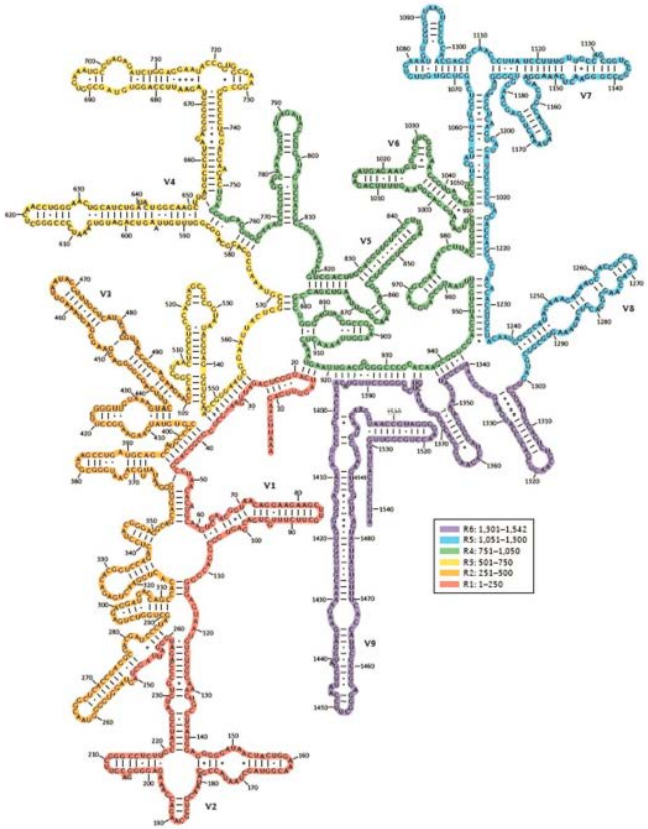
# Assessing microbial diversity

- Cultivation
- Clone libraries
- Next generation sequencing

} • Sequence-based

# Assessing microbial diversity

- Cultivation
- Clone libraries
- Next generation sequencing




- Sequence-based
- Marker gene: small-subunit ribosomal DNA
  - Universal
  - Conserved and hypervariable regions
  - Mutation rate close to species divergence

# Operational taxonomic units

- OTUs...

...are defined as sequences of sufficient similarity that are distinct from other sequences

...are dependend on the amplified region and analysis method

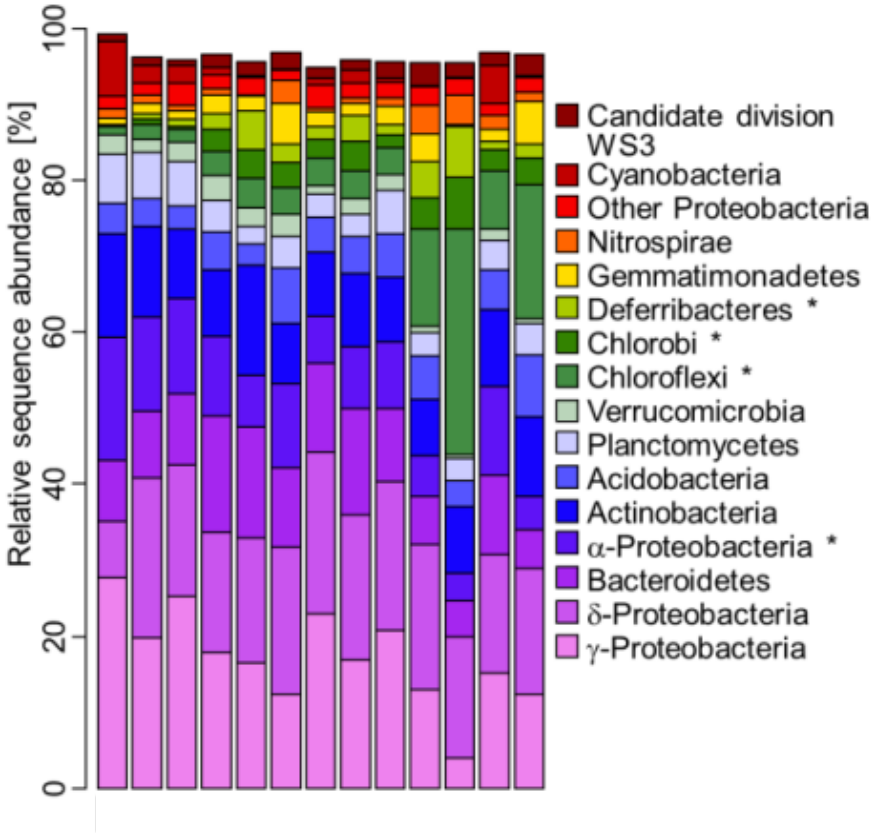
...are NOT comparable across studies

...do NOT represent species

...do NOT represent genome divergence

# Illumina paired-end sequencing

- 16S screening
- PCR-based (amplicon) sequencing
- 2x300bp reads
- High sample throughput
- Low sequencing depth



# Primer selection

- Target group: bacteria, archaea, universal prokaryotic, eukaryotes
- PCR bias: primer coverage
- Fragment length: insert size
- Taxonomic resolution: sequence variability

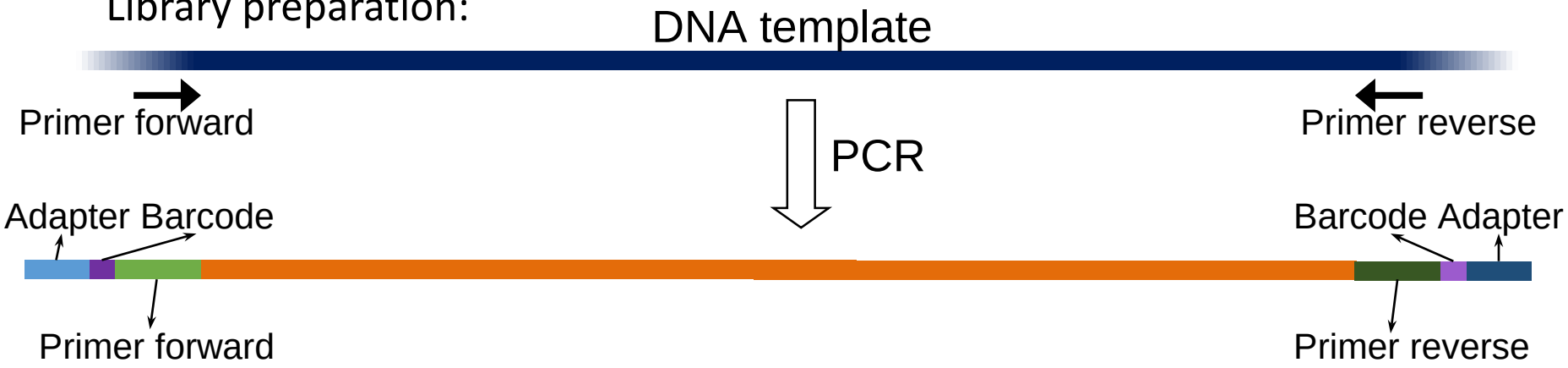
# Primer selection

- Target group: bacteria, archaea, universal prokaryotic, eukaryotes
- PCR bias: primer coverage
- Fragment length: insert size
- Taxonomic resolution: sequence variability

| Target group | Region | Primer pair | Fragment size | Primer sequence | Reference |
|---|---|---|---|---|---|
| Archaea | V4-V6 | Arch349F<br>Arch915R | ~510bp | GYGCASCAGKCGMGAAW<br>GTGCTCCCCCGCCAATTCCT | Amann et al. 1990<br>Klindworth et al. 2013 |
| Bacteria | V3-V4 | 341F<br>785R | ~420bp | CCTACGGGNGGCWGCAG<br>GACTACHVGGGTATCTAATCC | Klindworth et al. 2013 |
| Prokaryotes | V4-V5 | Bact515F<br>926R | ~410bp | GTGYCAGCMGCCGCGGTAA<br>CCGYCAATTYMTTTRAGTTT | Parada et al. 2016 |
| Eukaryotes | V4 | TAReukFor<br>TAReukRev | ~380bp | CCAGCASCYGCGGTAATTCC<br>ACTTTCGTTCTTGATYRA | Logares et al. 2012 |

# Sequence generation

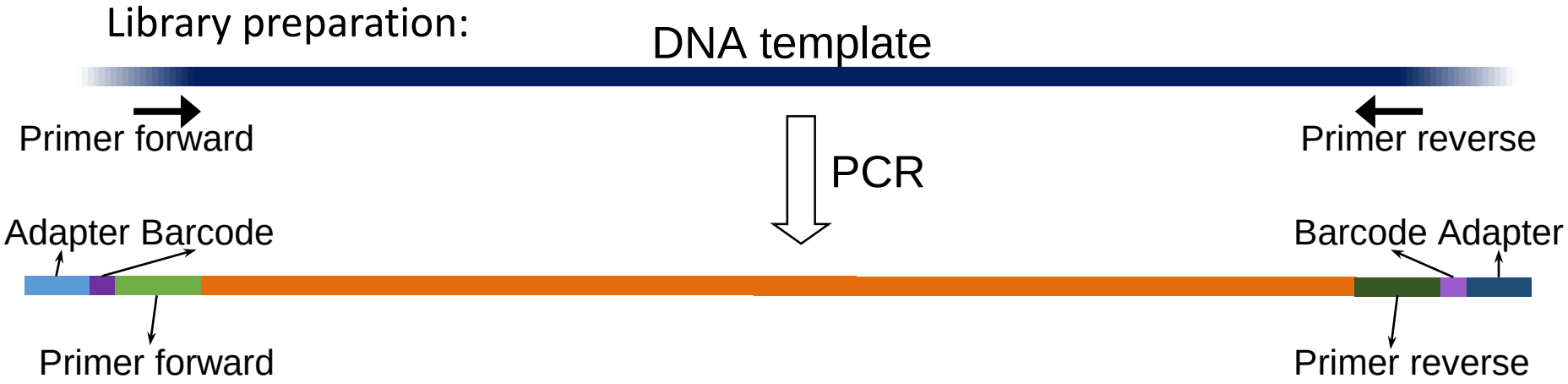
Library preparation:



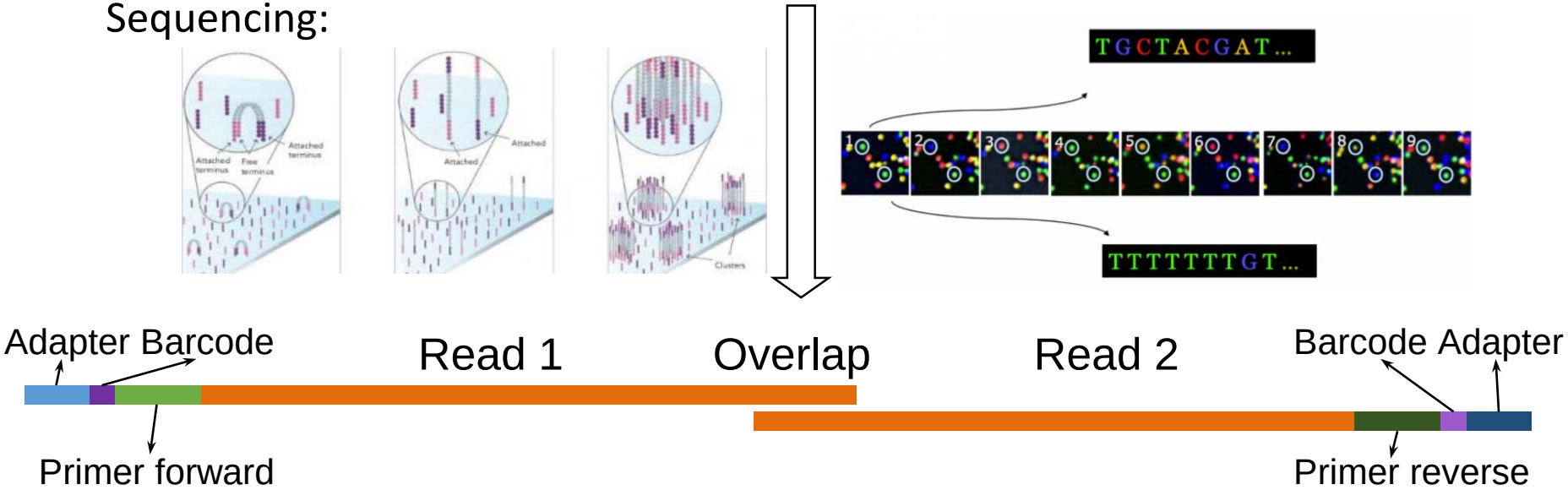

Sequencing:

# Sequence generation

Library preparation:




Sequencing:



# Fastq format

@Sequence accession
Sequence...
+
Base quality scores....

```
@SNL168:111:H2YY7BCXY:1:1101:11589:1946 1:N:0
TGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCCATGCCGCGTGTATGAA
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

→ Simple text files!

# Bioinformatic data analysis

1. Sequence preprocessing:
   - Adapter, Barcode and Primer removal
   - Quality control
   - Merging

2. OTU generation

3. Taxonomic classification

# Sequencing errors

**Errors during PCR-library preparation**

Chimeras

Incorporation of wrong base

**Errors during sequencing**

Wrong base call

- Single nucleotide polymorphisms (SNPs)

# Sequencing errors

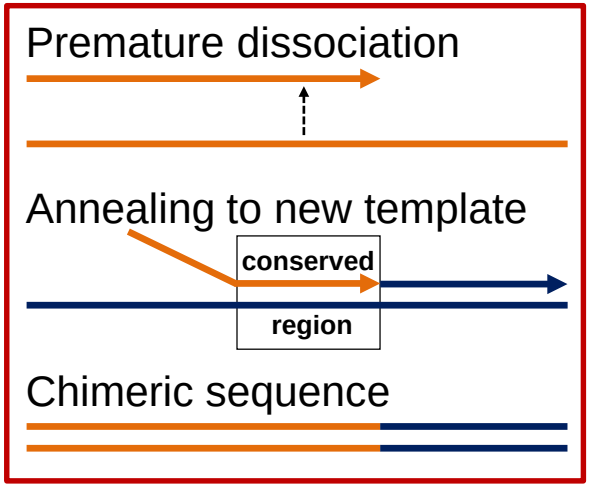
**Errors during PCR-library preparation**

Chimeras

Incorporation of wrong base

**Errors during sequencing**

Wrong base call




- Single nucleotide polymorphisms (SNPs)

# Sequencing errors

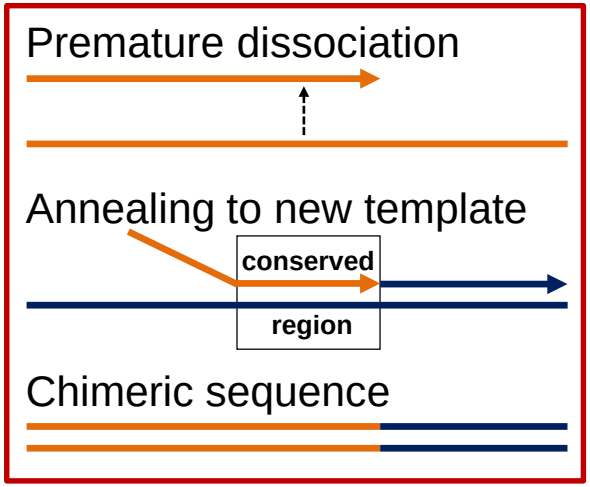
**Errors during PCR-library preparation**

Chimeras

Incorporation of wrong base

**Errors during sequencing**

Wrong base call




- Single nucleotide polymorphisms (SNPs)

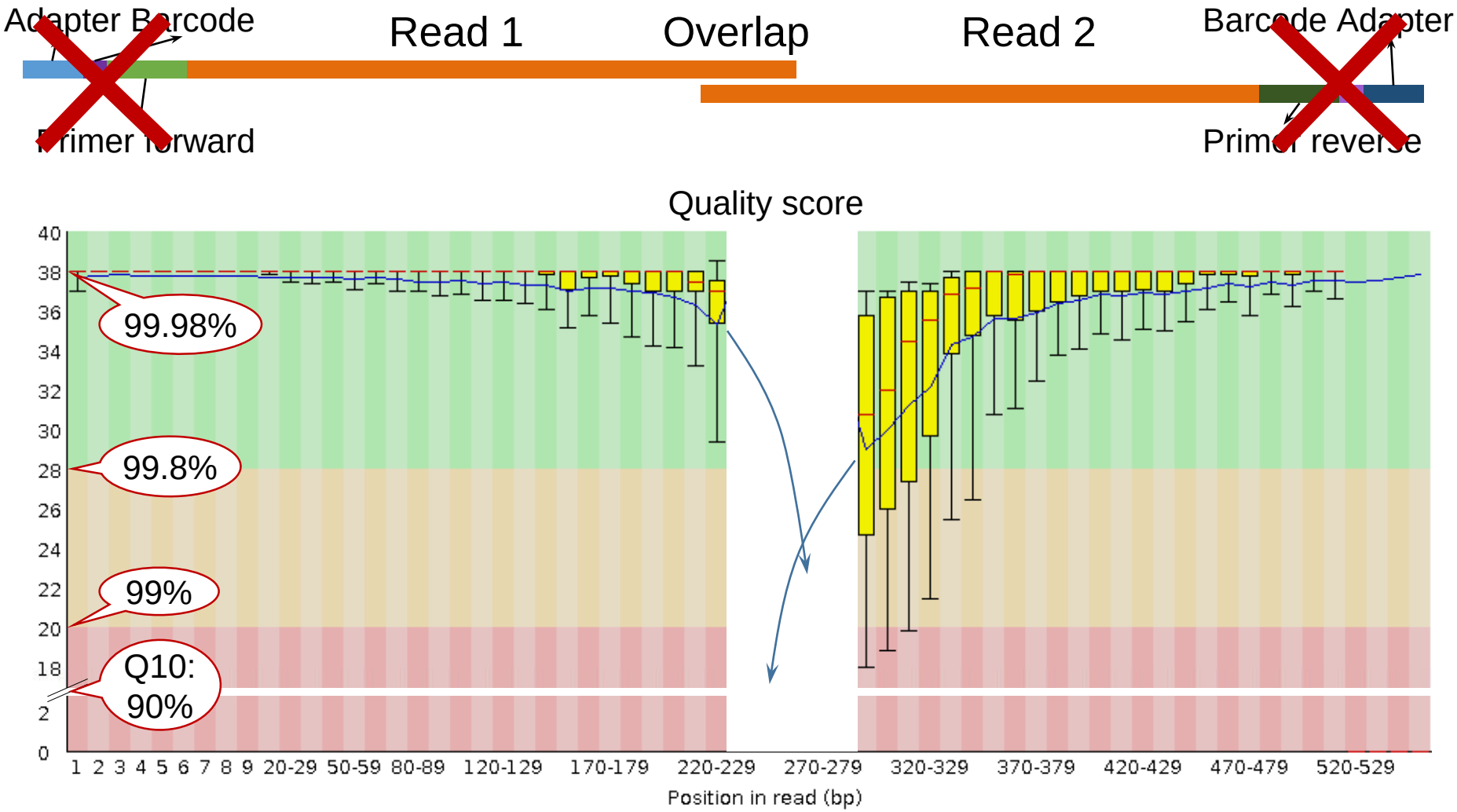
- In downstream processing estimated based on sequence abundance and alignment quality

- Removal based on base quality scores

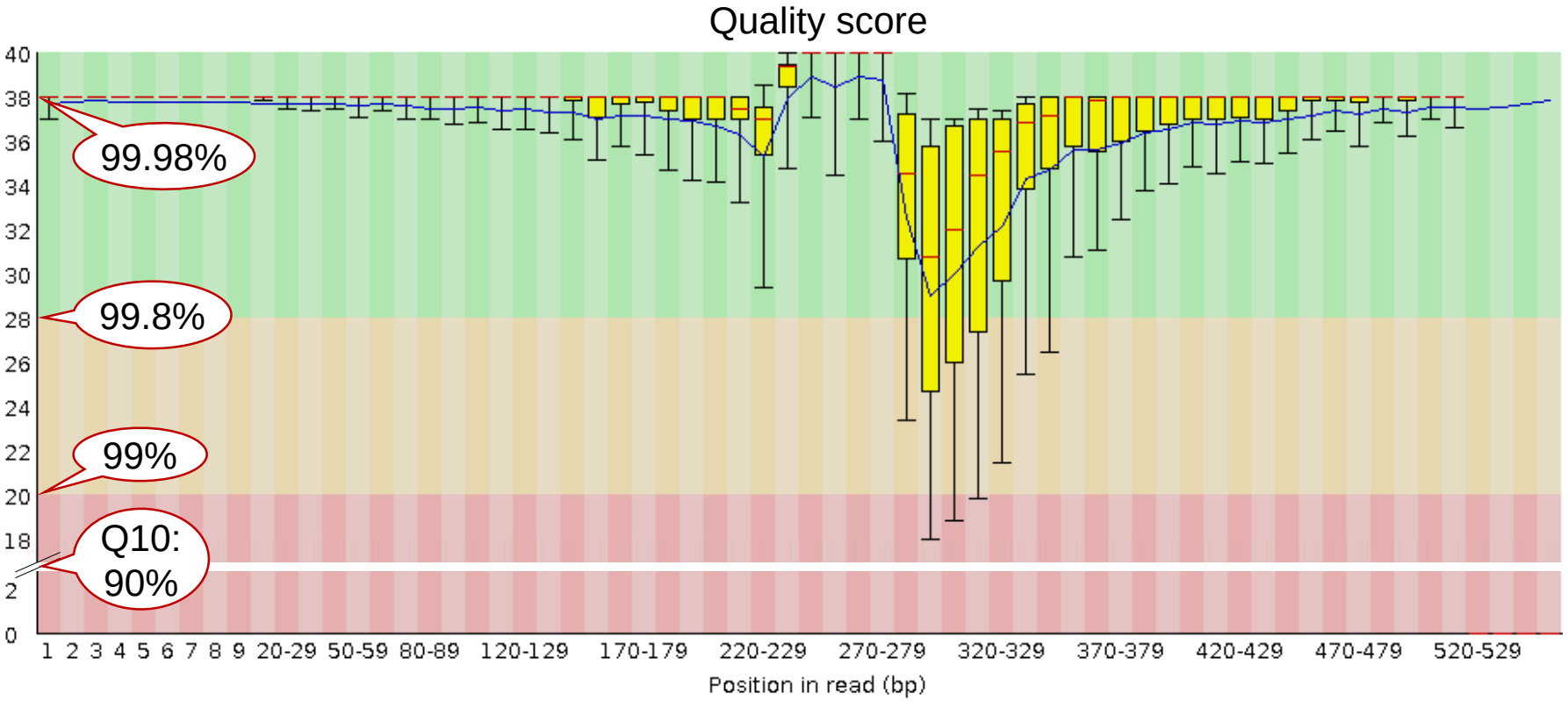
# Sequence pre-processing

# Sequence pre-processing

# Sequence pre-processing

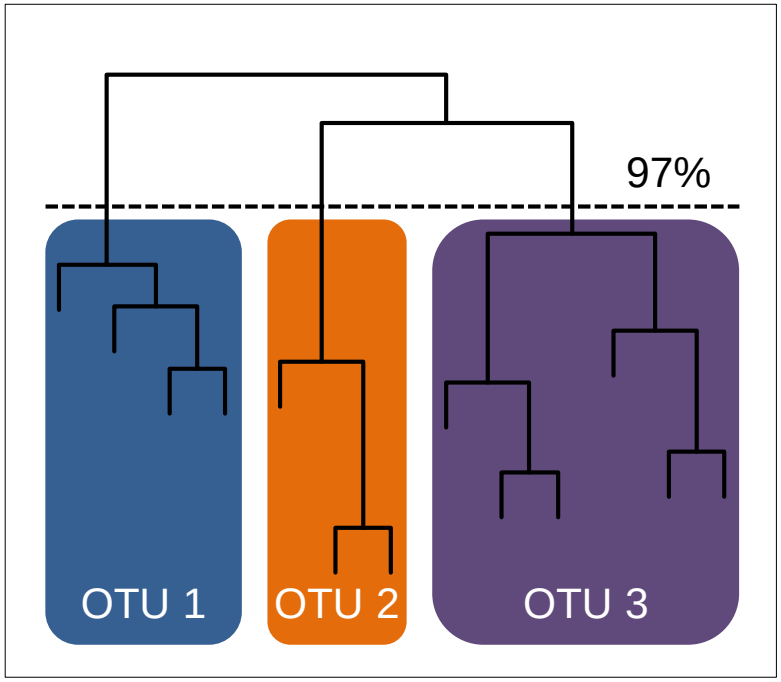


# Sequence pre-processing

# OTU generation

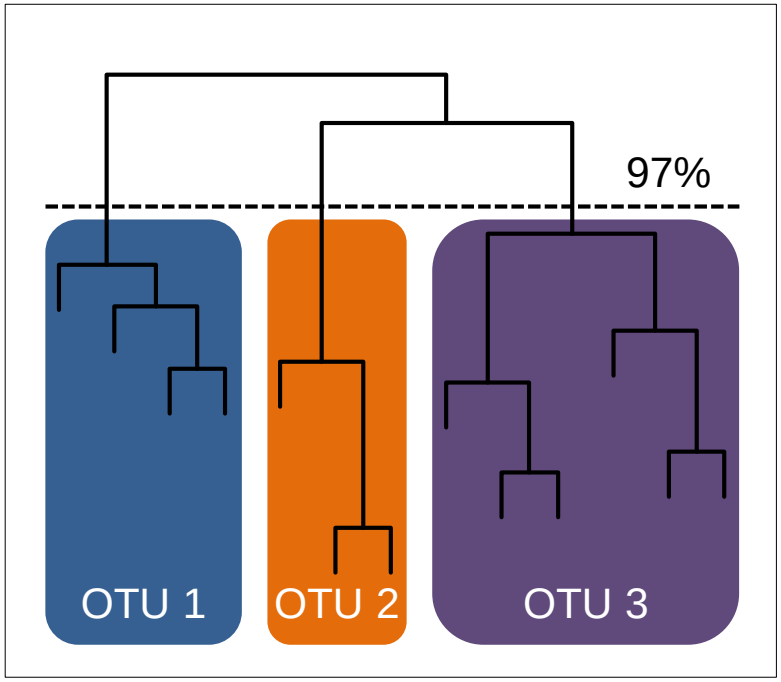
**Hierarchical clustering**




- Better defined OTUs than heuristic clustering
- Very slow
- *mothur*

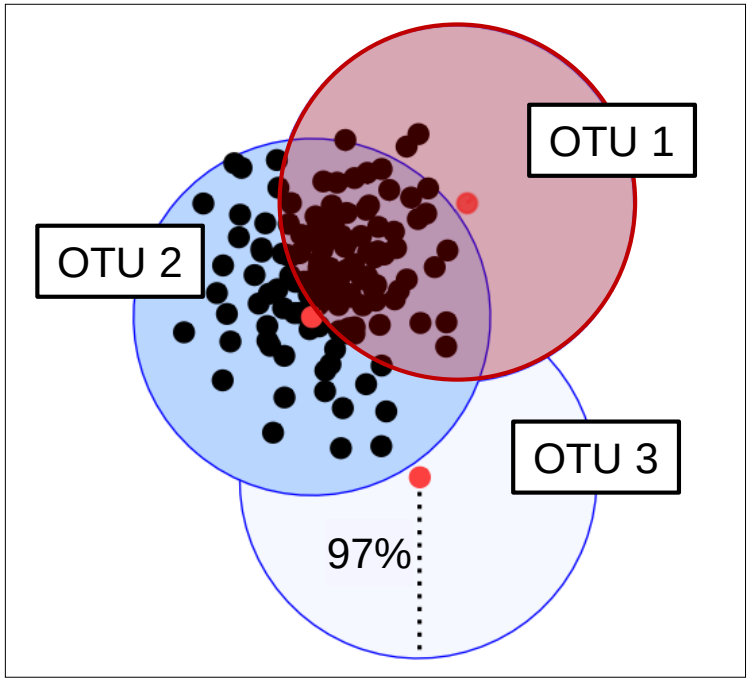
# OTU generation

**Hierarchical clustering**




- Better defined OTUs than heuristic clustering
- Very slow
- *mothur*

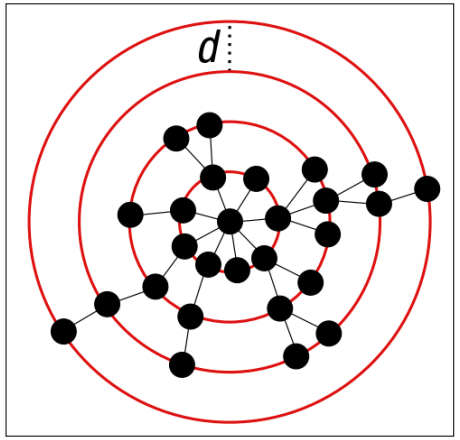
**Heuristic clustering**



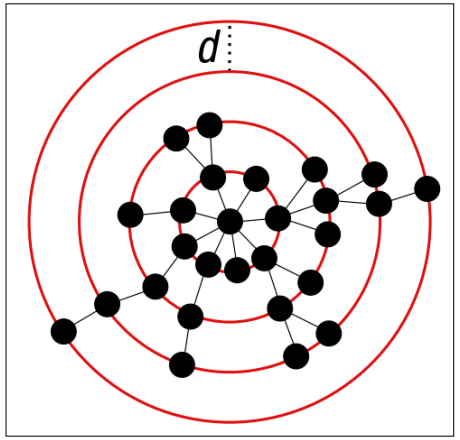

- Fast compared to hierarchical clustering
- Low reproducibility
- *usearch, qiime*

# OTU generation

- Swarming (*swarm*, *OBITools*)
  - Fast
  - Variable OTU cut-off
  - High reproducibility
  - Dimension of swarms depending on sequencing space

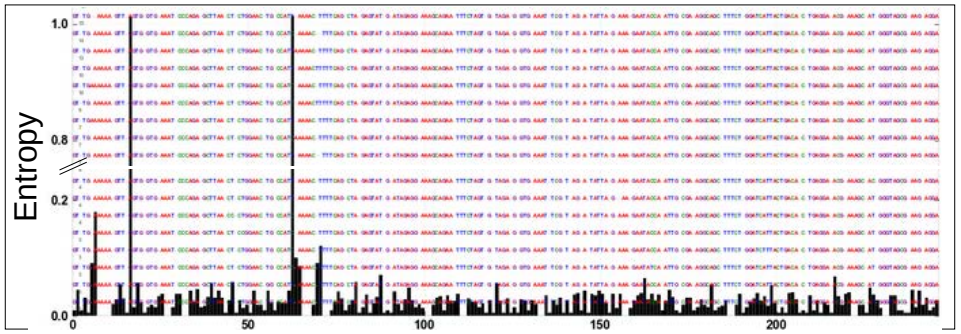


# OTU generation

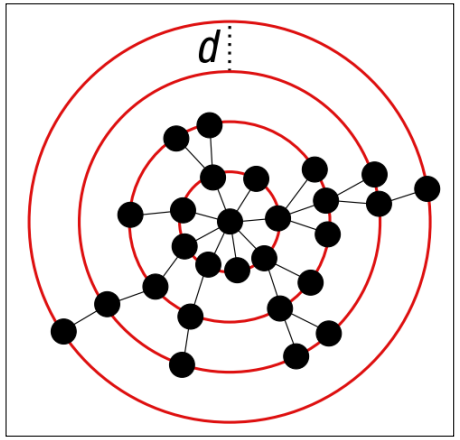



- Swarming (*swarm*, *OBITools*)
  - Fast
  - Variable OTU cut-off
  - High reproducibility
  - Dimension of swarms depending on sequencing space

- Minimum entropy decomposition (MED)
  - Fast
  - Omits stochastic variation
  - Sub-species resolution (SNPs)
  - No rare biosphere

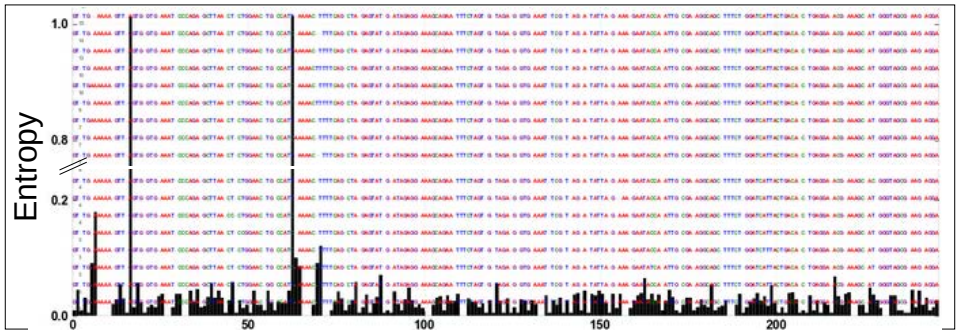


# OTU generation




- Swarming (*swarm*, *OBITools*)
  - Fast
  - Variable OTU cut-off
  - High reproducibility
  - Dimension of swarms depending on sequencing space

- Minimum entropy decomposition (MED)
  - Fast
  - Omits stochastic variation
  - Sub-species resolution (SNPs)
  - No rare biosphere




- Denoising (*dada2*)
  - Probability that any unique sequence was created by sequencing error
  - High taxonomic resolution
  - Less rare (spurious?) OTUs than swarm
  - Requires very high quality sequences as input

# Taxonomic classification

Domain;Phylum;Class;Order;Family;Genus

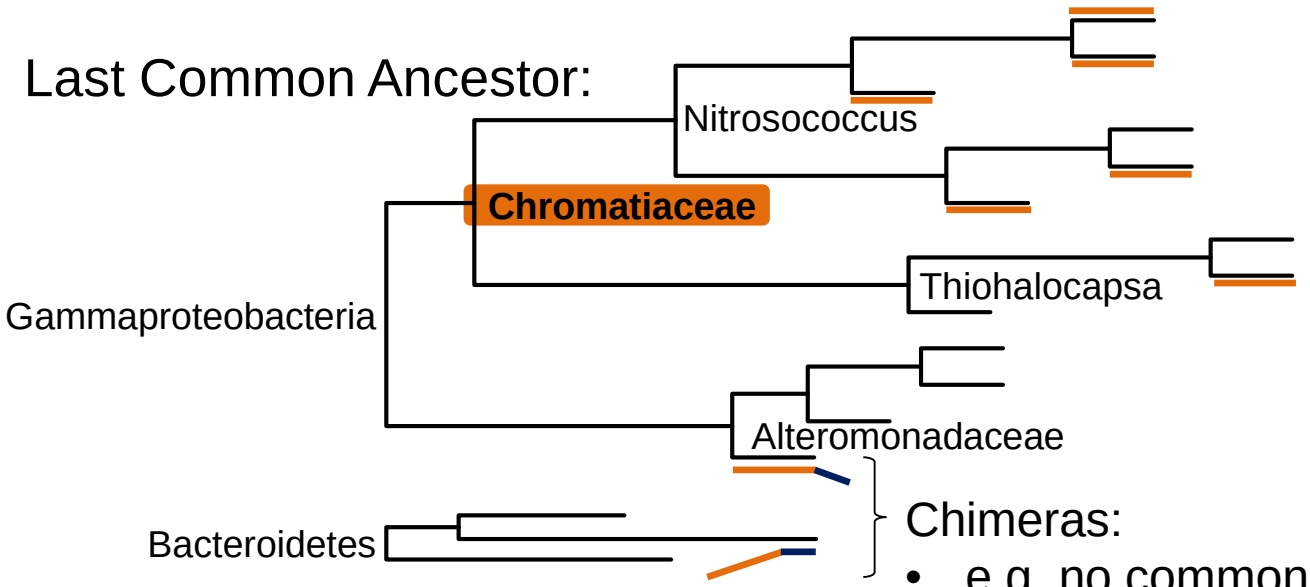
Bacteria;Proteobacteria;Gammaproteobacteria;Chromatiales;Chromatiaceae;Nitrosococcus

# Taxonomic classification

Domain;Phylum;Class;Order;Family;Genus

Bacteria;Proteobacteria;Gammaproteobacteria;Chromatiales;Chromatiaceae;Nitrosococcus

- Last Common Ancestor:




Chimeras:

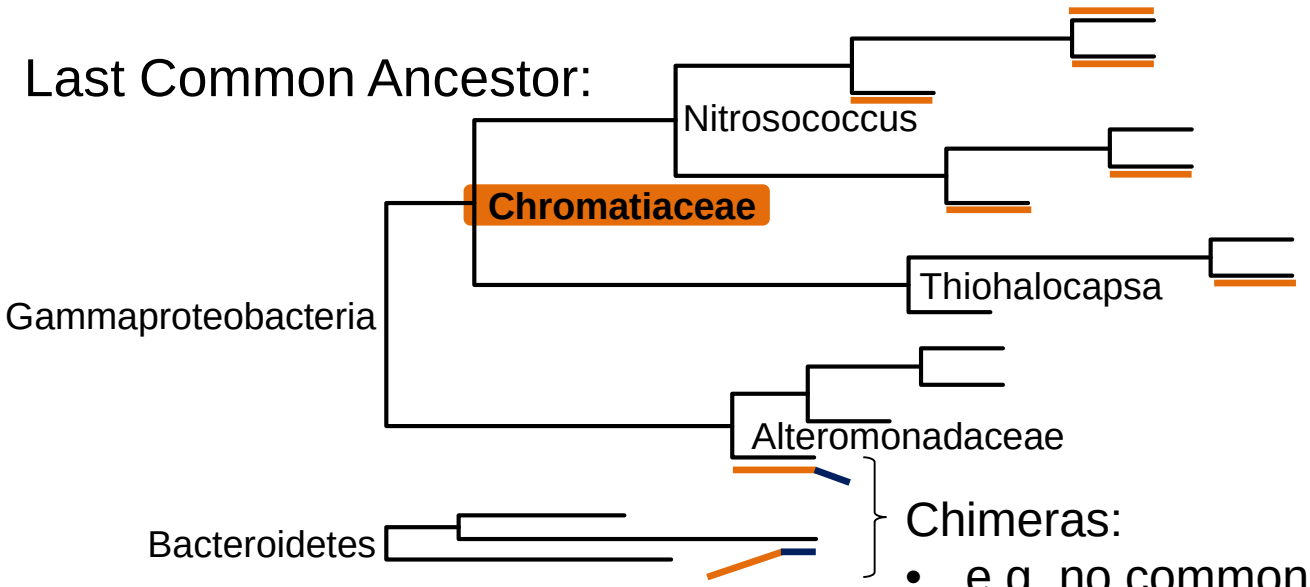
- e.g. no common path beyond domain level?
- Chimera detection programs

# Taxonomic classification

Domain;Phylum;Class;Order;Family;Genus

Bacteria;Proteobacteria;Gammaproteobacteria;Chromatiales;Chromatiaceae;Nitrosococcus

- Last Common Ancestor:




Chimeras:

- e.g. no common path beyond domain level?
- Chimera detection programs

- Taxonomic reference databases: **Silva**, RDP, Greengenes, etc.

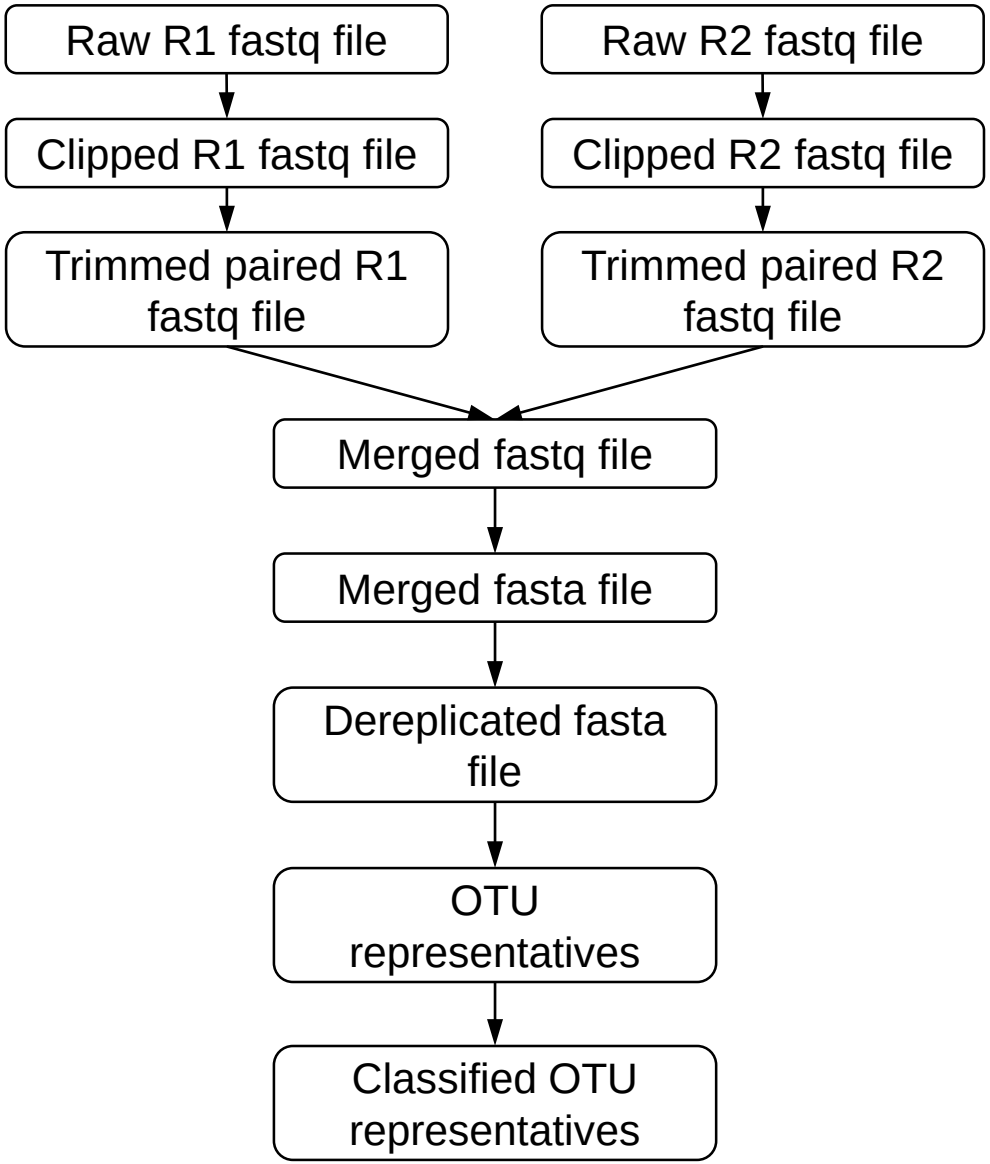
# Available programs

- Complete analysis pipelines:
  - Mothur: https://github.com/mothur/mothur, https://www.mothur.org/
  - Qiime (Qiime2): http://qiime.org/, https://qiime2.org/
  - Dada2: https://github.com/benjjneb/dada2, http://www.nature.com/nmeth/journal/v13/n7/full/nmeth.3869.html
  - Silvangs: https://www.arb-silva.de/ngs/

- Stepwise analysis:
  - Cutadapt: https://github.com/marcelm/cutadapt
  - Trimmomatic: http://www.usadellab.org/cms/?page=trimmomatic
  - PEAR: https://sco.h-its.org/exelixis/web/software/pear/
  - Bbmap: https://sourceforge.net/projects/bbmap/
  - Swarm: https://github.com/torognes/swarm
  - MED: http://merenlab.org/2014/11/04/med/
  - SINA: https://www.arb-silva.de/aligner/

# 16S Example workflow

| | |
|---|---|
| Primer removal | cutdapt |
| Quality trimming | trimmomatic |
| Read merging | PEAR |
| Quality check | FastQC |
| Conversion to fasta | reformat.sh |
| Dereplication | bash scripts |
| OTU clustering | swarm, MED |
| Taxonomic classification | SINA |



# Useful links and literature

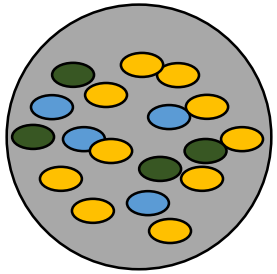
- Tutorials:
  - Gitlab: https://gitlab.leibniz-zmt.de/chh/bioinf
  - Amplicon workflow bioconductor: https://f1000research.com/articles/5-1492/v2
- Forums:
  - http://seqanswers.com/
  - https://omictools.com/
- Primer selection:
  - Parada et al. 2016: https://www.ncbi.nlm.nih.gov/pubmed/26271760
  - Eloe-Fadrosh et al. 2016: https://www.ncbi.nlm.nih.gov/pubmed/27572438
  - Silva test prime: https://www.arb-silva.de/search/testprime/
- Taxonomy:
  - https://www.arb-silva.de/aligner/
  - https://www.arb-silva.de/browser/

# Data analysis

# Concepts of microbial diversity

Alpha diversity:

- Diversity within one sample
- Richness
- Evenness



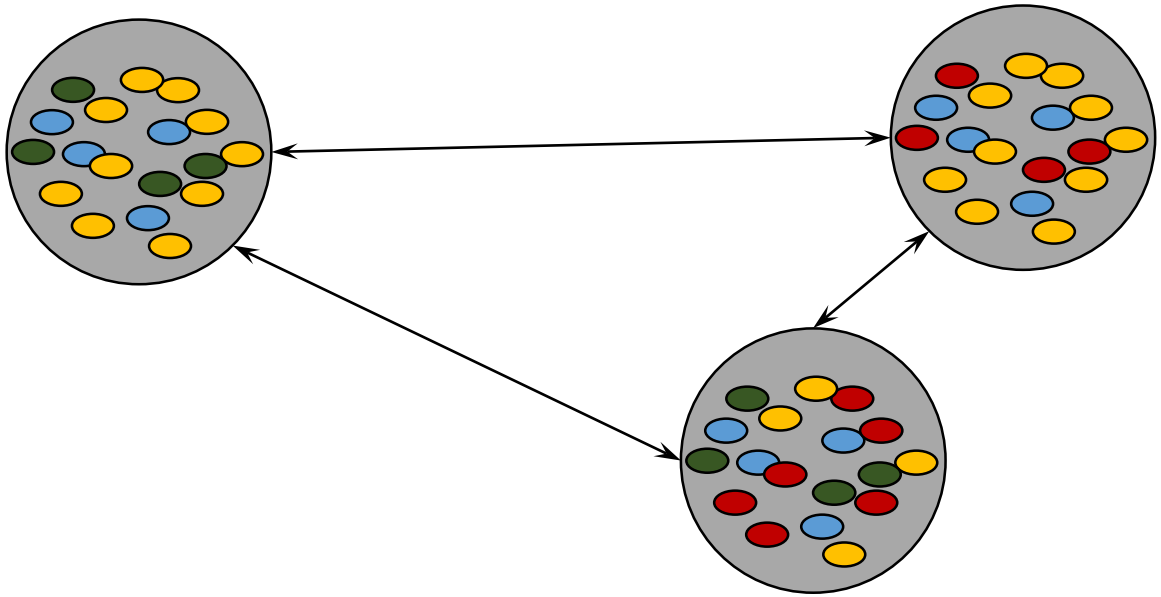
# Concepts of microbial diversity

Alpha diversity:

- Diversity within one sample
- Richness
- Evenness

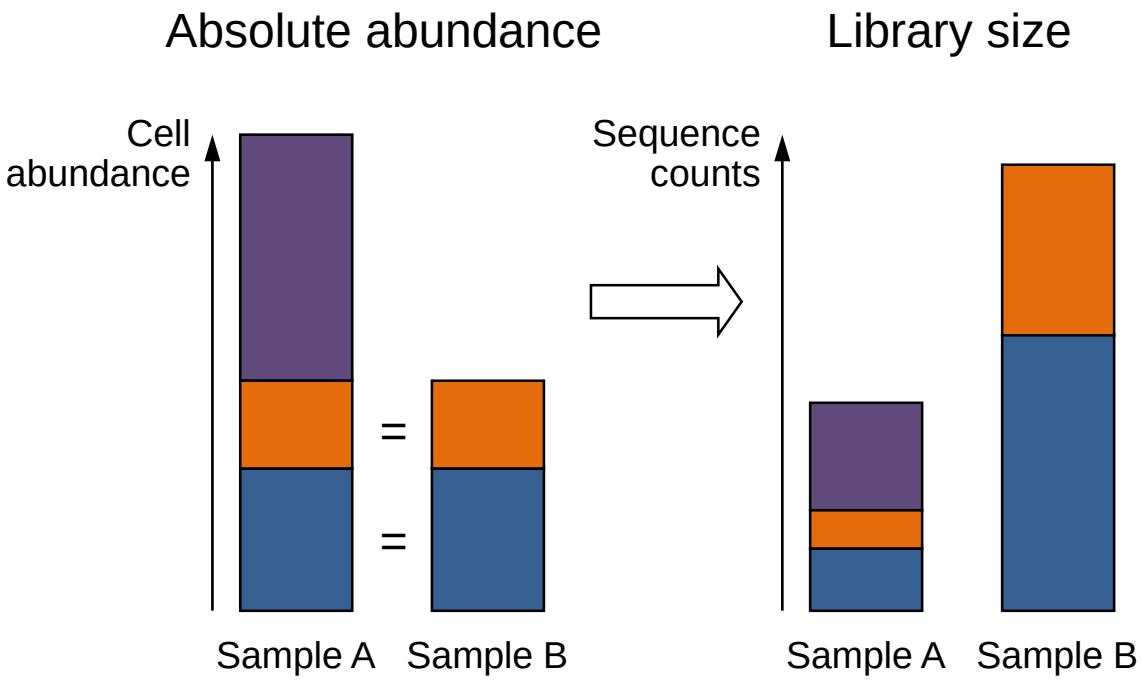
Beta diversity:

- Diversity between samples (comparison)
- Dissimilarity
- Shared OTUs



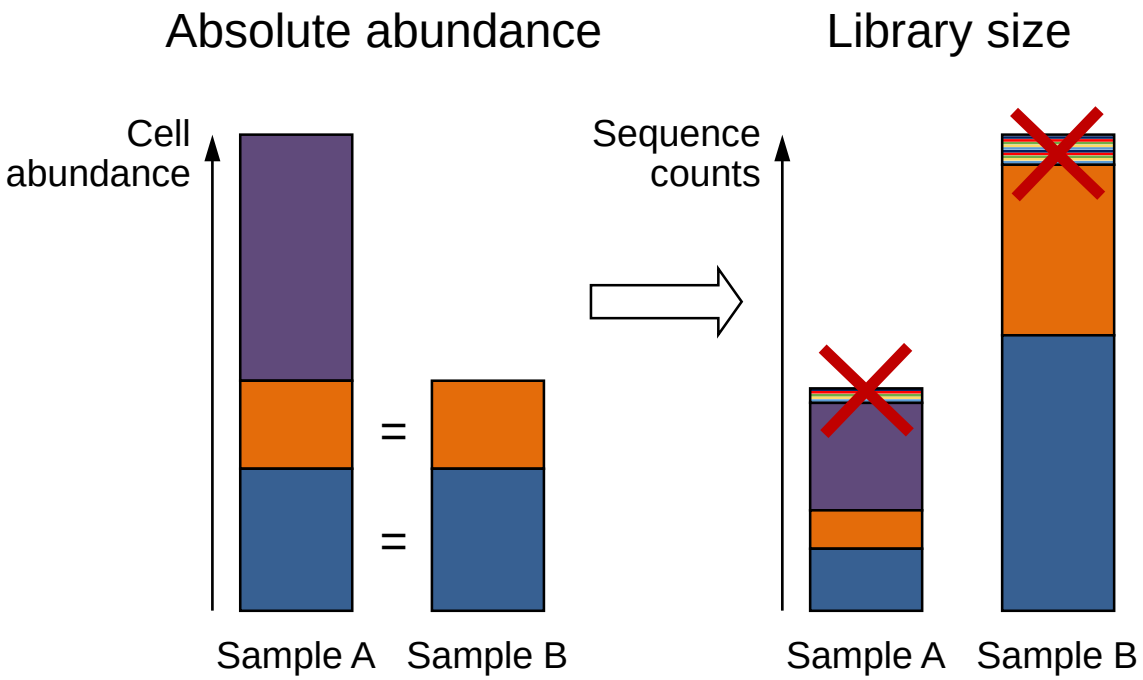
# Pitfalls of NGS data analysis

- Library size bias: coverage of microbial diversity



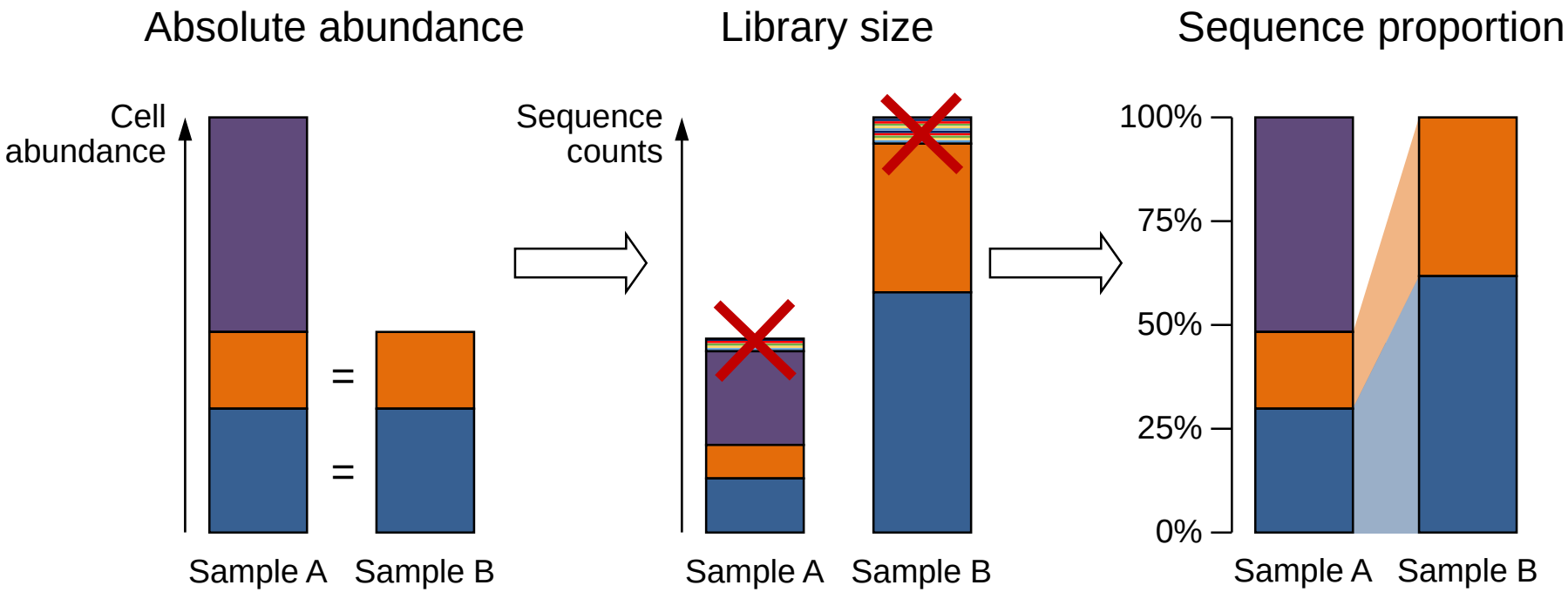
# Pitfalls of NGS data analysis

- Library size bias: coverage of microbial diversity
- Rare sequences: detection limit of method



# Pitfalls of NGS data analysis

- Library size bias: coverage of microbial diversity
- Rare sequences: detection limit of method
- Compositionality: spurious correlations



# Alpha diversity

- Alpha diversity indices:
  - Chao1
  - ACE
  - Richness
  - Shannon
  - Inverse Simpson

! Influence of rare OTUs

# Alpha diversity

- Alpha diversity indices:
  - Chao1
  - ACE
  - Richness
  - Shannon
  - Inverse Simpson

Influence of rare OTUs

- Unifying concept: Hill numbers
  - Richness (q = 0)
  - Exponential Shannon (q = 1)
  - Inverse Simpson (q = 2)

$$ {}^{q}D = \left( \sum_{i=1}^{S} p_i^{q} \right)^{1/(1-q)} $$

# Alpha diversity

- Alpha diversity indices:
  - Chao1
  - ACE
  - Richness
  - Shannon
  - Inverse Simpson

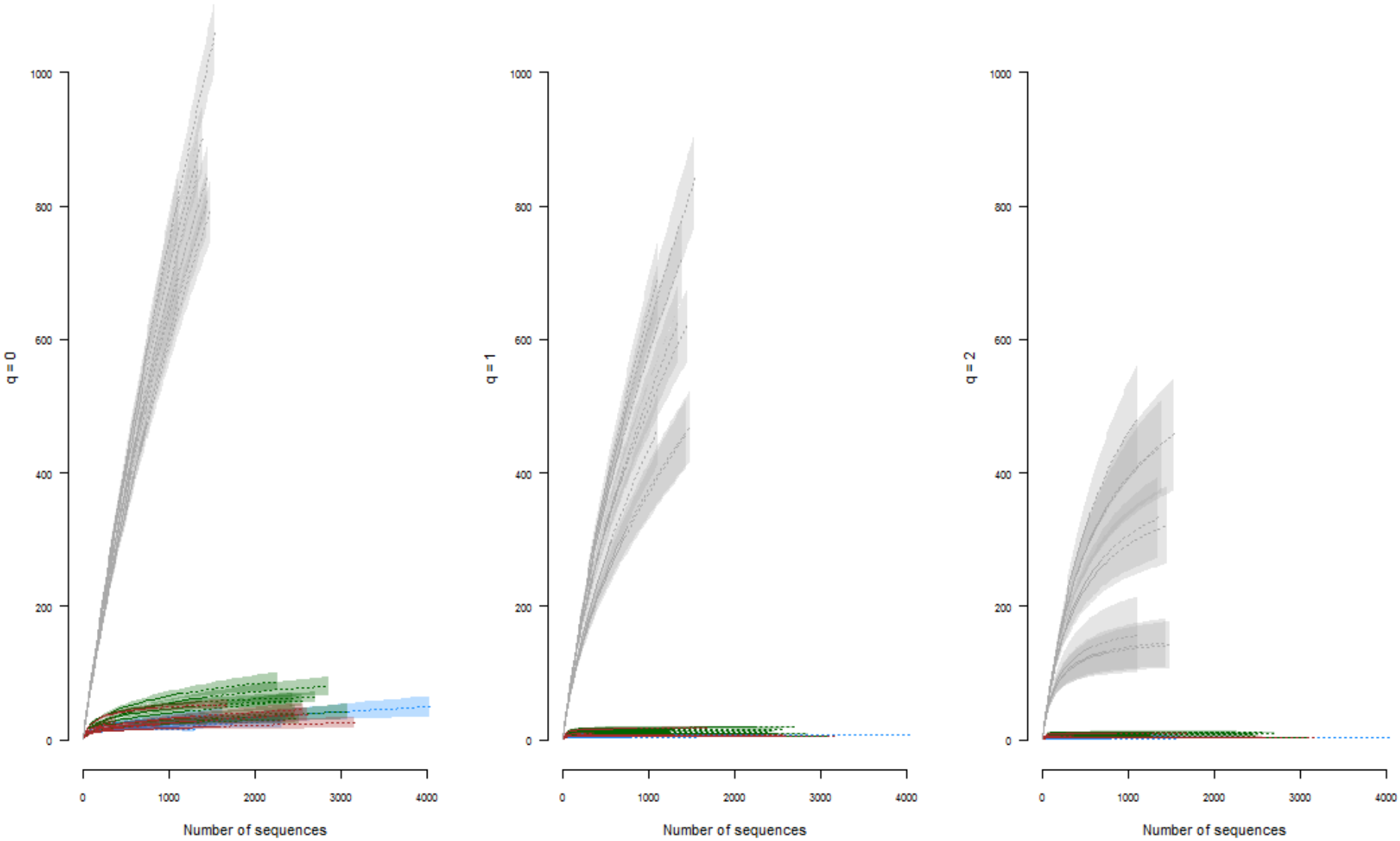
Influence of rare OTUs

- Unifying concept: Hill numbers
  - Richness (q = 0)
  - Exponential Shannon (q = 1)
  - Inverse Simpson (q = 2)

$$ {}^{q}D = \left( \sum_{i=1}^{S} p_i^{q} \right)^{1/(1-q)} $$

- Sequencing depth and rare biosphere?
  - Subsample sequences to equal library sizes
  - Correct the number of singletons per sample
  - Use rarefaction curves to estimate covered diveristy by available sequencing depth

# Rarefaction curves



# Beta diversity

- Dissimilarity vs. distance

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| S1 | 14 | 2 | 14 | 14 |
| S2 | 10 | 14 | 0 | 8 |
| S3 | 0 | 5 | 0 | 2 |
| S4 | 0 | 0 | 1 | 0 |

# Beta diversity

- Dissimilarity vs. distance

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| S1 | 14 | 2 | 14 | 14 |
| S2 | 10 | 14 | 0 | 8 |
| S3 | 0 | 5 | 0 | 2 |
| S4 | 0 | 0 | 1 | 0 |

Asymmetrical vs. symmetrical

**Bray-Curtis** vs. euclidean

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | | | |
| S2 | 0.5 | 0 | | |
| S3 | 0.8 | 0.6 | 0 | |
| S4 | 1.0 | 1 | $\boxed{1}$ | 0 |

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | | | |
| S2 | 19.8 | 0 | | |
| S3 | 23.3 | 14.7 | 0 | |
| S4 | 23.8 | 19 | $\boxed{5.5}$ | 0 |

# Beta diversity

- Dissimilarity vs. distance

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| S1 | 14 | 2 | 14 | 14 |
| S2 | 10 | 14 | 0 | 8 |
| S3 | 0 | 5 | 0 | 2 |
| S4 | 0 | 0 | 1 | 0 |

Asymmetrical vs. symmetrical
**Bray-Curtis** vs. euclidean

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | | | |
| S2 | 0.5 | 0 | | |
| S3 | 0.8 | 0.6 | 0 | |
| S4 | 1.0 | 1 | $\boxed{1}$ | 0 |

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | | | |
| S2 | 19.8 | 0 | | |
| S3 | 23.3 | 14.7 | 0 | |
| S4 | 23.8 | 19 | $\boxed{5.5}$ | 0 |

- Zeros in ecology: Is this species really not there or did we just not find it?

→ double zeros not relevant

# Beta diversity

- Dissimilarity vs. distance

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| S1 | 14 | 2 | 14 | 14 |
| S2 | 10 | 14 | 0 | 8 |
| S3 | 0 | 5 | 0 | 2 |
| S4 | 0 | 0 | 1 | 0 |

presence/ absence →

| | OTU1 | OTU2 | OTU3 | OTU4 |
|---|---|---|---|---|
| S1 | 1 | 1 | 1 | 1 |
| S2 | 1 | 1 | 0 | 1 |
| S3 | 0 | 1 | 0 | 1 |
| S4 | 0 | 0 | 1 | 0 |

Asymmetrical vs. symmetrical
**Bray-Curtis** vs. euclidean

**Jaccard**

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | | | |
| S2 | 0.5 | 0 | | |
| S3 | 0.8 | 0.6 | 0 | |
| S4 | 1.0 | 1 | $\boxed{1}$ | 0 |

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | | | |
| S2 | 19.8 | 0 | | |
| S3 | 23.3 | 14.7 | 0 | |
| S4 | 23.8 | 19 | $\boxed{5.5}$ | 0 |

| | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| S1 | 0 | | | |
| S2 | 0.25 | 0 | | |
| S3 | 0.5 | 0.33 | 0 | |
| S4 | 0.75 | 1 | 1 | 0 |

- Zeros in ecology: Is this species really not there or did we just not find it?

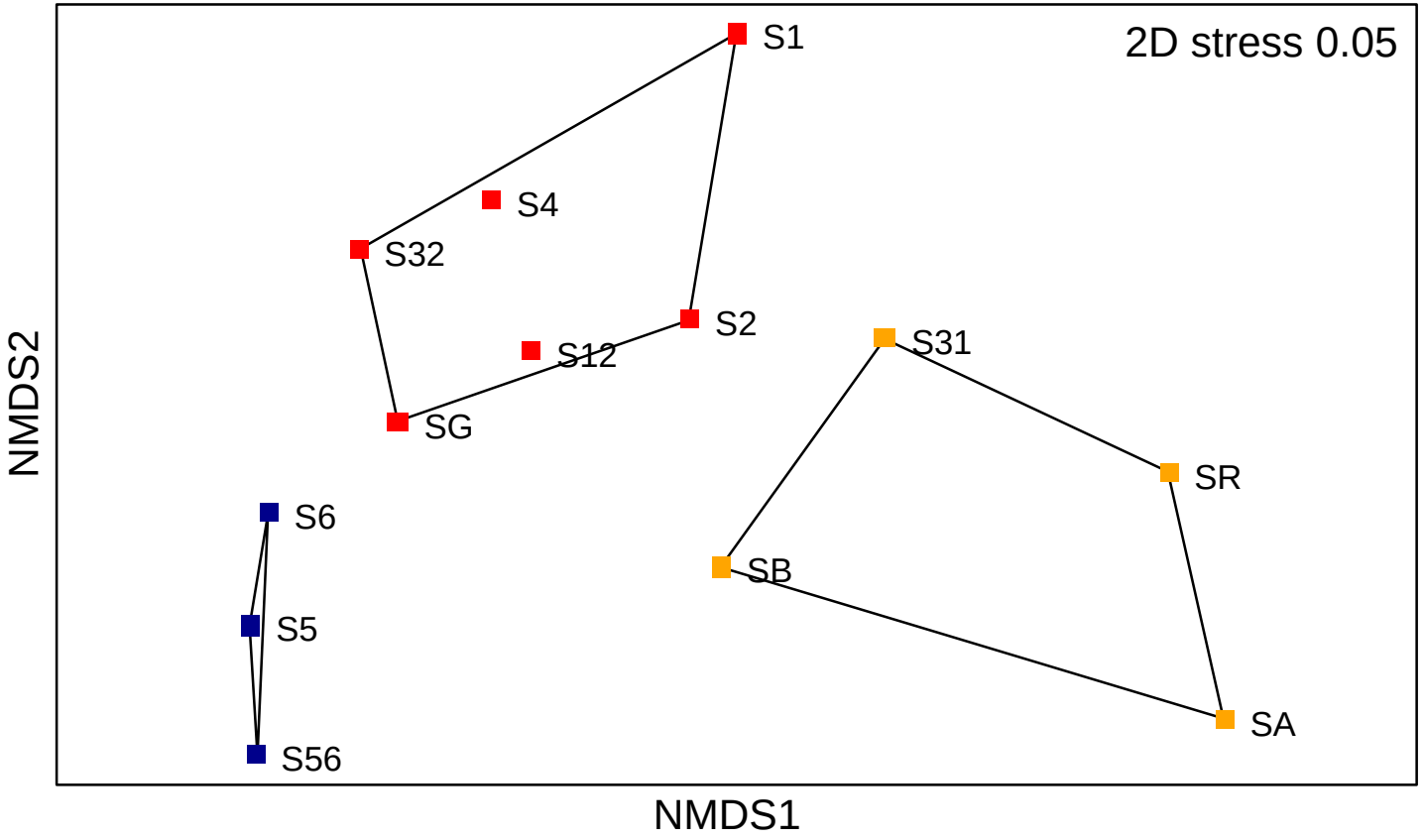
→ double zeros not relevant

# Ordination

- Visualization of a multidimensional matrix in a reduced set of dimensions (e.g. 2)

# Ordination

- Visualization of a multidimensional matrix in a reduced set of dimensions (e.g. 2)
- Non-metric multidimensional scaling (NMDS)



# Differential OTU abundance

- Testing statistical differences between environmental conditions **for each OTU**

# Differential OTU abundance

- Testing statistical differences between environmental conditions **for each OTU**
- Compositionality correction: Centered log-ratio transformation (clr) $\log(x_i) - \log(n\sqrt{product(x_1 \ldots xn)})$

# Differential OTU abundance

- Testing statistical differences between environmental conditions **for each OTU**
- Compositionality correction: Centered log-ratio transformation (clr) $\log(x_i) - \log(n\sqrt{product(x_1 \ldots xn)})$
- P-value correction:

$$FWER = 1 - (1 - alpha)^n$$

Number of comparisons

Family-wise error rate

Significance threshold per comparison

| n | FWER |
|---|---|
| 1 | 0.05 |
| 3 | 0.14 |
| 1000 | ~ 1 |

# Differential OTU abundance

- Testing statistical differences between environmental conditions **for each OTU**
- Compositionality correction: Centered log-ratio transformation (clr) $\log(x_i) - \log(n\sqrt{product(x_1 \ldots xn)})$
- P-value correction:

$$FWER = 1 - (1 - alpha)^n$$

Number of comparisons

Family-wise error rate

Significance threshold per comparison

| n | FWER |
|---|---|
| 1 | 0.05 |
| 3 | 0.14 |
| 1000 | ~ 1 |

- Implementation: ALDEx2 (http://www.microbiomejournal.com/content/2/1/15, https://github.com/ggloor/ALDEx2)

# Useful links and literature

- Tutorials:
  - GUSTAME: https://sites.google.com/site/mb3gustame/
  - Github: https://github.com/chassenr/Tutorials/tree/master/R_course_MPI
  - Vegan: http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf
- Alpha diversity:
  - Chao et al. 2014: http://onlinelibrary.wiley.com/doi/10.1890/13-0133.1/abstract
  - Chiu and Chao 2016: https://www.ncbi.nlm.nih.gov/pubmed/26855872
  - iNEXT: https://cran.r-project.org/web/packages/iNEXT/iNEXT.pdf
- Recent advances in NGS data analysis
  - Mini reviews: http://www.sciencedirect.com/science/article/pii/S1047279716300722, http://www.sciencedirect.com/science/article/pii/S1047279716300734
  - SPIECEASI: https://github.com/zdk123/SpiecEasi
  - Random forests: https://rpubs.com/michberr/randomforestmicrobe
  - Balance trees: http://msystems.asm.org/content/2/1/e00162-16

# Thank you for your attention!

Any questions?