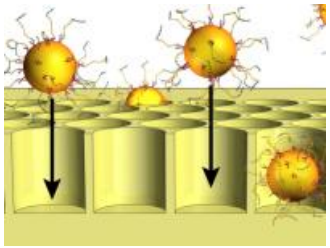
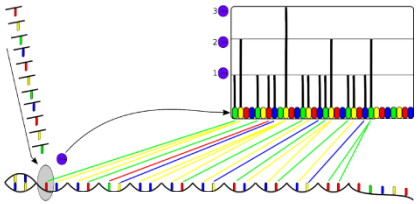
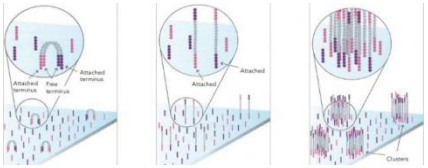
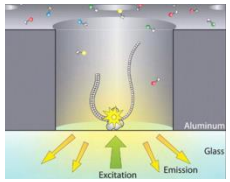
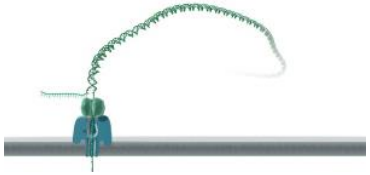
DNA sequence analysis: data handling, visualization and (some) multivariate statistics in R

R roundtable 21.1.2016 – Christiane Hassenrück

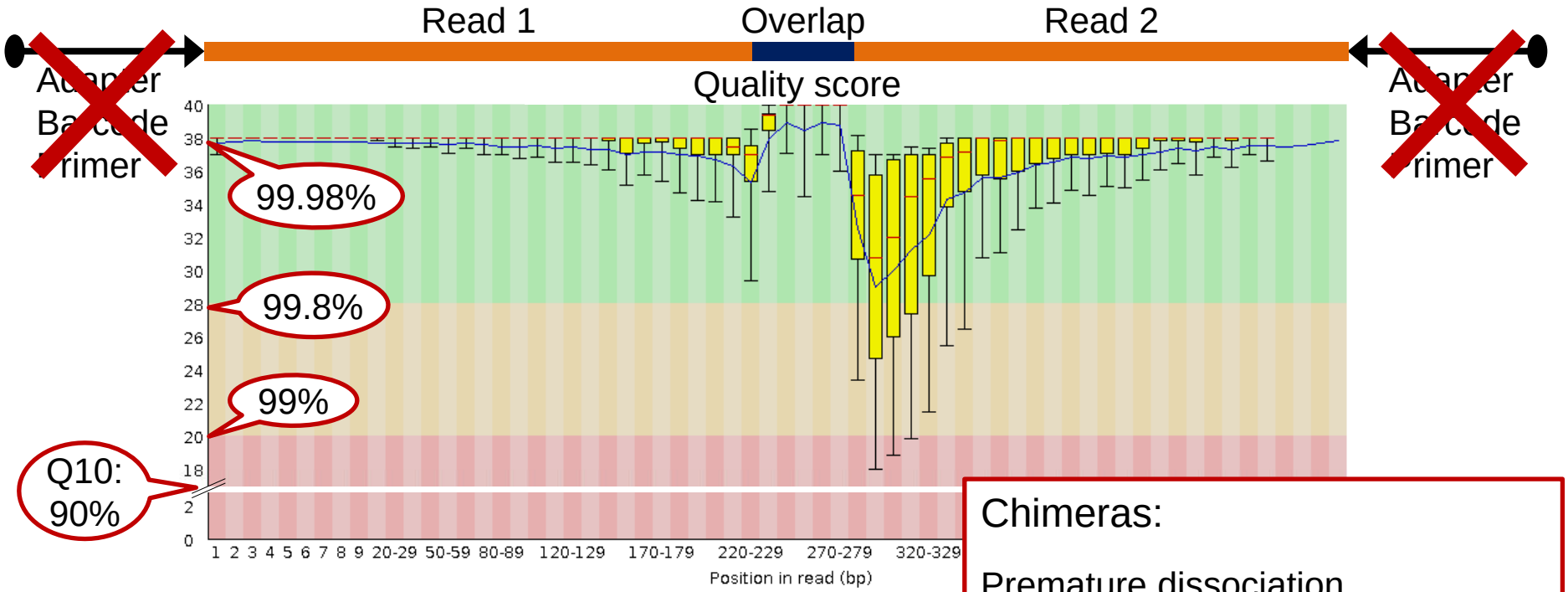
outline

1. Sequencing technologies and sequence analysis pipelines	Major steps, sequencing errors, OTU clustering algorithms, taxonomy
2. Data handling in R	Reading data: Data subsetting, parsing taxonomic paths
3. Diversity concepts	Alpha and beta diversity, random subsampling, effective species number, symmetric vs. asymmetric diversity indices
4. Plotting community data in R	Custom functions (alpha diversity, abundant taxa, networks), ordination plots
5. Pitfalls of sequence analysis	Compositionality, sampling design and replication
6. Multivariate statistics	Overall patterns (ANOSIM, PERMANOVA, RDA), differential OTU abundance
7. Estimating bacterial functions	Tax4Fun

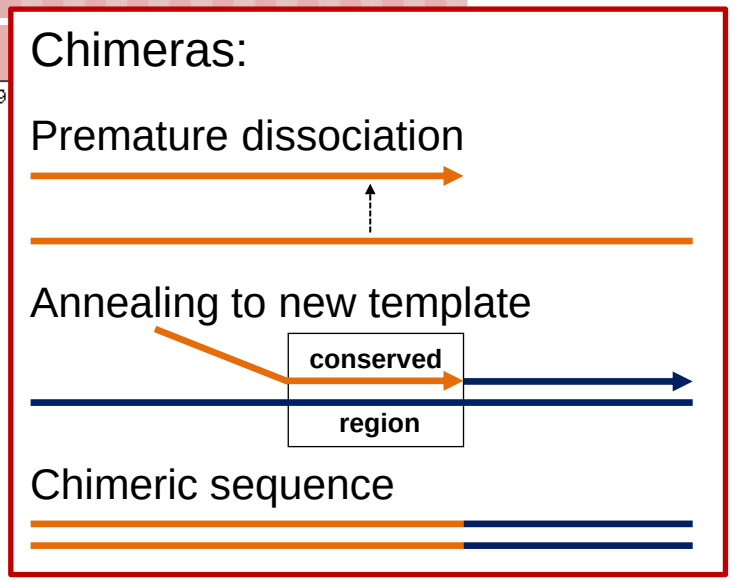
NGS technologies

Sequencing technology	Principle	Read length	Errors	Comments
454		< 450 bp SE	homopolymers	discontinued
Ion torrent		< 400 bp SE	homopolymers	
Illumina		< 300 bp SE < 550 bp PE	substitutions	Currently preferred for amplicon and shotgun sequencing
PacBio		> 10 kb	~ 10% error rate (single pass)	
nanopore		> 10 kb	< 4% error rate	In development

Bioinformatic analysis



1) Adapter, barcode, primer removal
2) Quality trimming
3) Merging
4) OTU clustering
5) Taxonomic classification



OTU clustering methods

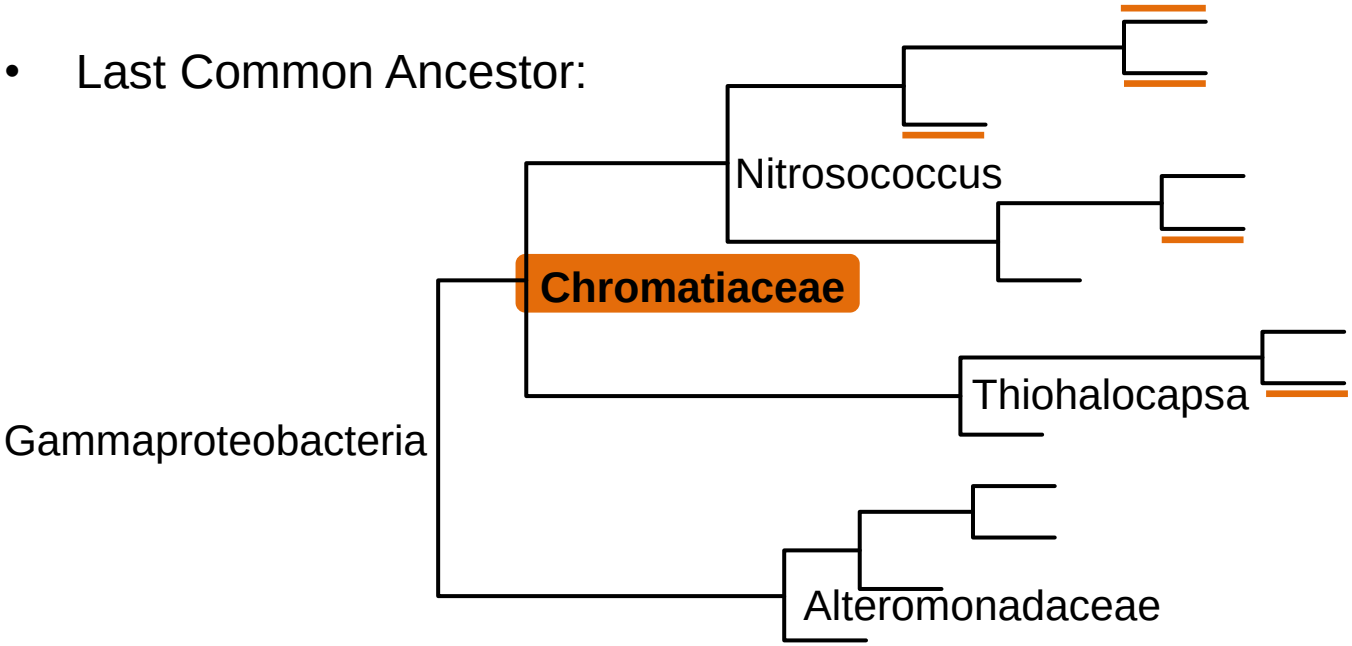
Algorithm	Pro's	Con's
Hierarchical	• Better defined OTUs than heuristic clustering	• Very slow
Heuristic (greedy)	• Fast compared to hierarchical clustering	• Low reproducibility
Swarm	• Fast • Variable OTU cut-off • High reproducibility	• Large swarms
Oligotype	• Fast • Omits stochastic variation • Sub-species resolution (SNPs)	• No rare biosphere

Taxonomic classification

Domain;Phylum;Class;Order;Family;Genus

Bacteria;Proteobacteria;Gammaproteobacteria;Chromatiales;Chromatiaceae;Nitrosococcus

- Last Common Ancestor:



- Truncated paths:

  Bacteria;Proteobacteria;Gammaproteobacteria;Chromatiales;Chromatiaceae;**unclassified**

- Incomplete paths:

  Bacteria;Proteobacteria;Gammaproteobacteria;**Incertae Sedis**;**Incertae Sedis**;Sedimenticola

Let‘s get to R

Sample-by-OTU table

OTU-by-taxonomy table

Environmental data

Consistent sample and OTU names!

Alpha diversity

- Community richness and eveness per sample
- Classical indices:
  - OTU number
  - Shannon
  - Chao 1
  - Inverse Simpson

- Hill numbers: effective species number
  - 1 formula for all indices → only changing 1 parameter (q)

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

^{0}D = OTU number

^{1}D = exp (Shannon)

^{2}D = inverse Simpson

Influence of rare biosphere

- Rarefaction curves
- Unequal sequencing depth → random subsampling to compare alpha diversity indices across samples

Beta diversity

- Community (dis)similarity between samples

	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

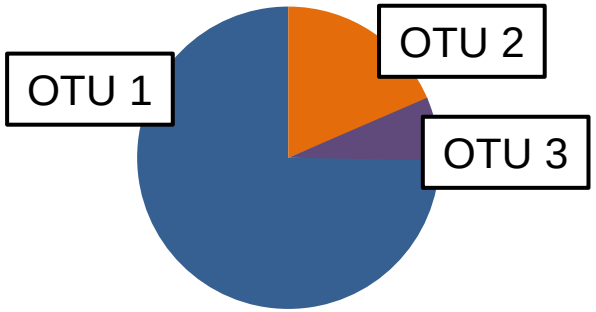
- Zeros in ecology: Is this species really not there or did we just not find it?
  → double zeros not relevant

Plotting in R

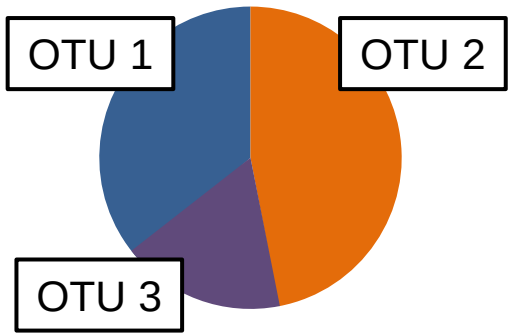
Alpha diversity indices
Abundant taxa
OTU networks
Ordination plots

Pitfalls of sequence analysis

- Compositionality: OTU abundances not independent
- Centered log-ration transformation (clr): $\log(x_i) - \log(\mathrm{n}\sqrt{product(x_1 \ldots xn)})$



Decrease in OTU 1 →



Sample 1	%	clr
OTU 1	75	1.83
OTU 2	18	-0.23
OTU 3	7	-1.59

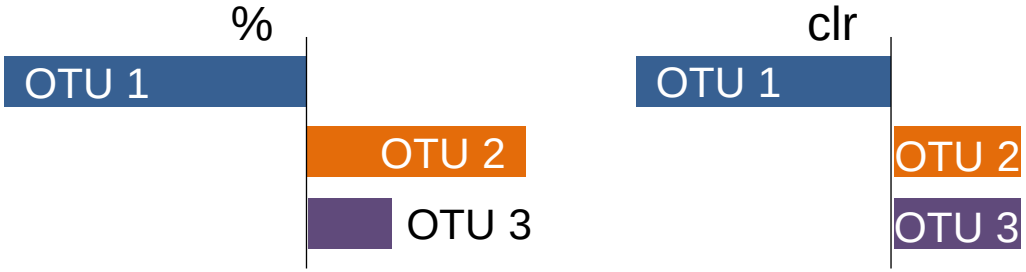
%: Δ 11
clr: Δ 1.36

Sample 2	%	clr
OTU 1	35	0.14
OTU 2	47	0.61
OTU 3	18	-0.75

%: Δ 29
clr: Δ 1.36

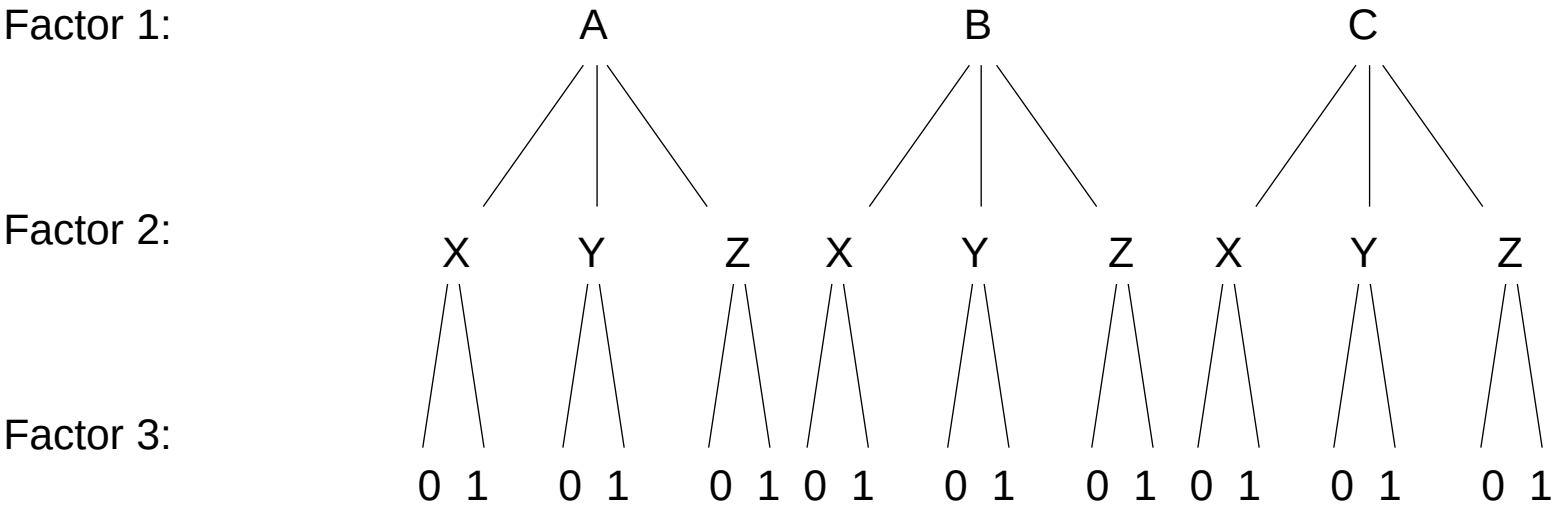
→ Difference between OTUs independent of library size

- Difference between samples:



Pitfalls of sequence analysis

- Sampling design



Replication: 3 x 3 x 2 x 5 = 90
Estimated costs: ~ 5000 € consumables

- To reduce costs, i.e. only sequencing selected samples:

Better to remove a condition than to reduce the number of replicates!

Multivariate statistics in R

ANOSIM and PERMANOVA

Redundancy analysis

Differential OTU abundance (ALDEx2)

Path analysis

→ check out: http://mb3is.megx.net/gustame

Estimating bacterial functions

Tax4Fun: http://tax4fun.gobics.de/

Sample-by-OTU table

Taxonomic paths

Taxonomic reference database

Useful links

- Web links:
  - http://www.arb-silva.de/
  - http://www.arb-silva.de/download/archive/
  - http://mb3is.megx.net/gustame
  - http://tax4fun.gobics.de/
  - https://github.com/chassenr/NGS

- References:

  - Aßhauer, K.P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015) Tax4Fun : predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* **31**: 2882–2884.
  - Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., and Ellison, A.M. (2014) Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* **84**: 45–67.
  - Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2014) Minimum entropy decomposition : Unsupervised oligotyping for sensitive partitioning of high- throughput marker gene sequences. *Isme J* **9**: 968–979.
  - Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., and Gloor, G.B. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**: 15.
  - Mahé, F., Rognes, T.T., Quince, C., de Vargas, C., and Dunthorn, M. (2014) Swarm : robust and fast clustering method for amplicon-based studies. *PeerJ* 1–12.
  - Sinclair, L., Osman, O.A., Bertilsson, S., Eiler, A. (2015) Microbial Community Composition and Diversity via 16S rRNA Gene Amplicons: Evaluating the Illumina Platform. *PLoS ONE* 10: e0116955.