EMBS 2015
50th European Marine Biology Symposium

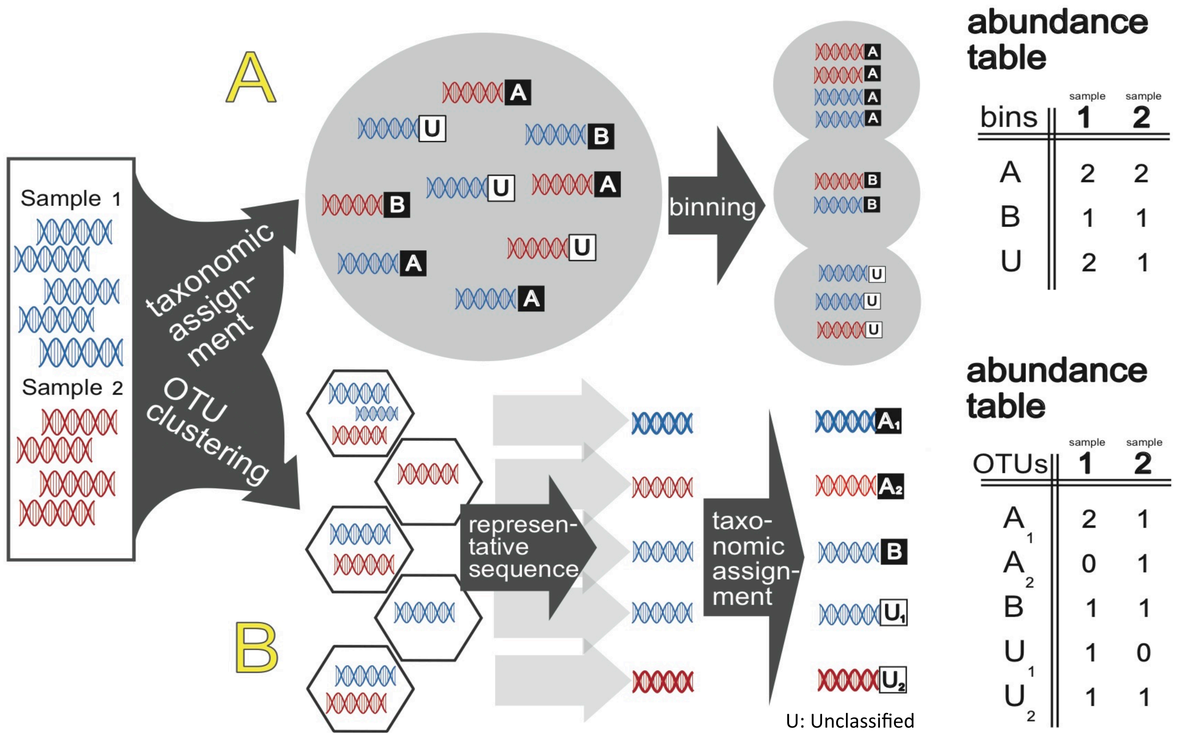
ALFRED-WEGENER-INSTITUT
HELMHOLTZ-ZENTRUM FÜR POLAR- UND MEERESFORSCHUNG

Stefan Neuhaus, Paul Kloss and Stephan Frickenhaus

From sequence data to biodiversity information – towards time series data

Current bioinformatics analyses for biodiversity from molecular sequence data are discussed on the background of next generation high throughput sequencing technologies. In particular, for creating time series of community composition data from amplicon sequencing approaches, different methods and their implications are compared: OTU-clustering vs. phylotyping, tree-based taxonomic assignment vs. assignment based on only-sequence characteristics. Software and hardware requirements as well as aspects of sustainable bioinformatics support are discussed. As a concrete example of analyses support, details of the AWI pipeline QZIP are shown.

Biodiversity with amplicon data: Phylotypes or OTUs?



Phylotypes (A) are reference dependent. If reference lacks of taxonomic groups, resolution is unsufficient. **OTU (B)** cluster base on %-similarity and reflect ecotype distribution more accurately. But %-threshold is arbitrary and often does not correspond to species separation. Modern approaches (Swarm, MED) based on single nucleotide differences provide better ecological insights.

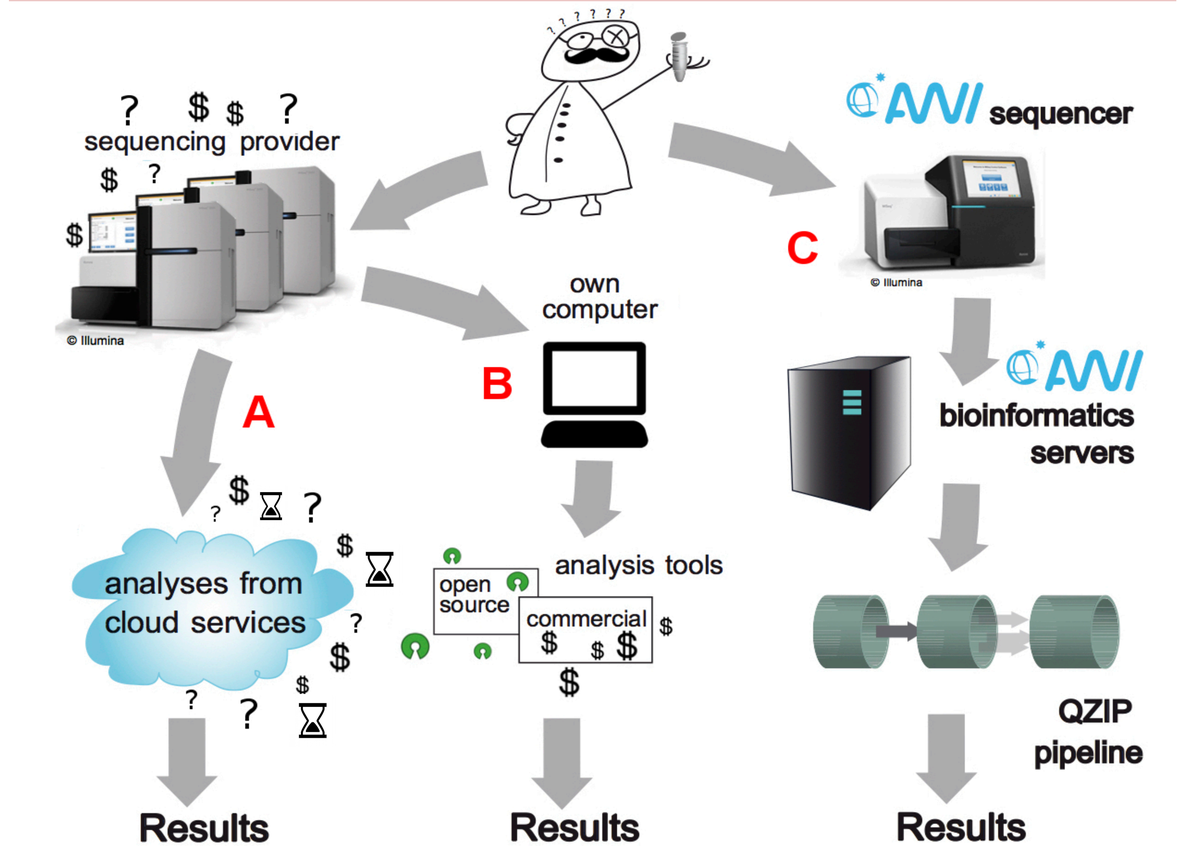
Taxonomy by tree or by only-sequence characteristics?

1. **Alignment-based similarity comparison** of query against set of reference sequences: **Blastn:** assignment of full taxonomy of best hit. **Uclust consensus:** assignment of common taxonomy prefix of distinct number of best hits (with distinct minimum query similarity).
2. **Machine learning classification algorithm based on sequence sub-word (k-mer) profiles: RDP:** training with reference set determines consensus k-mer profiles of taxonomic groups; classification by profile comparison. For each taxonomic rank of the assignment an uncertainty value is provided.
3. **Placement** of queries onto labelled and **fixed backbone tree** synthesized from subset of well-selected reference sequences: **Phyloassigner:** depending on taxonomic coverage of reference and preset uncertainty value, queries are placed close to leafs or close to inner nodes. Labels collected rootwards give taxonomic assignment.

Lack of well-sampled, equally taxon-distributed, error-corrected reference sequence sets: Blastn sensitive to single erroneous reference sequences. **Uclust consensus** and **RDP** provide acceptable results only at high taxonomic ranks. **Phylogeny-based classification methods (Phyloassigner) generally perform more accurately, even if reference lacks of taxonomic groups.**

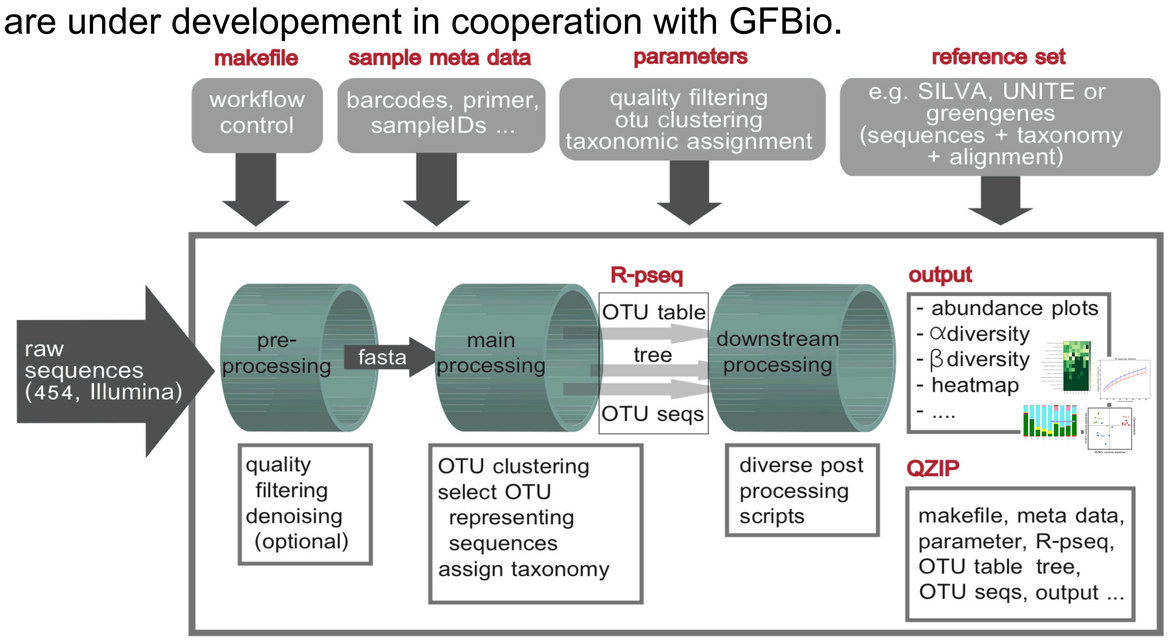
Strategy: **1. OTU clustering of full sample set 2. Preclassification based on RDP or Uclust consensus classifier 3. Placement of preclassified sequences onto taxonomic group specific phylogenetic trees (Phyloassigner)**

How can one analyze amplicon data of biodiversity surveys and obtain trustworthy and reproducible results?



A and **B**: External sequencing might be expensive, raw data are sometimes manipulated and feedback is often limited. No or restricted raw data storage concepts and analysis data submission support exist.
A: Cloud-based analysis services are either not for free or job inquiry is often queued for weeks. Analysis control and transparency are limited.
B: Transfer of data from sequence provider to scientists does not always follow data safety and security recommendations. Analysis tools are commercial or frequently hard to operate. The computer resources in terms of performance and data storage capacities are often not sufficient.

C: The AWI sequencer enables researchers to keep control of the raw data generation. Raw data are directly copied to the compute server for analysing and to tapes for long-term storage. The analyses base on open source software, which can be applied to eukaryotic and prokaryotic input data and are driven by experienced service scientists. Support and consulting is provided to optimize quality of analyses. Data submission is assisted and sustainable concepts and exchange formats are under developement in cooperation with GFBio.



Controlled execution of analysis workflows is ensured by the Qiime-dependent QZIP-pipeline. Standard (U/Vsearch) and modern (SWARM) cluster algorithms, chimera detection and methods for normalization of OTU data are supported. For further analyses an R-object (pseq) is created. OTU sequences are ready to be fed into Phyloassigner. Meta data, control logic, important result data, parameter and logging files are archived (QZIP) and accessible by browser-navigation. Thus transparency, user-friendliness and interchangeability are improved.

AWI bioinformatics services improves sustainability and reproducibility and is therefore suitable for long term data

References
1. Altschul et al. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410.
2. Caporaso et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7(5): 335-336.
3. DeSantis et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microb 72(7): 5069-5072.
4. Edgar 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460-2461.
5. Edgar et al. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27(16):194-200.
6. Eeren et al. 2015. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J. 9, 968-79.
7. Mahé et al. 2014. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ 2:e593.
8. McMurdie and Holmes 2013. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8(4).
9. Quast et al. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596.
10. Wang et al. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microb 73(16): 5261-5267.

ALFRED-WEGENER-INSTITUT HELMHOLTZ-ZENTRUM FÜR POLAR- UND MEERESFORSCHUNG
HELMHOLTZ GEMEINSCHAFT

Stefan Neuhaus
Stefan.Neuhaus@awi.de
0049471-4831-2329
Am Handelshafen 12
27570 Bremerhaven
www.awi.de