16S-sequencing analysis workshop

Matti Kankainen, PhD

Medical and Clinical Genetics,
HUS diagnostic center,
Helsinki and Uusimaa Hospital District

Hematology Research Unit Helsinki,
University of Helsinki &
Helsinki University Hospital Comprehensive Cancer Center

contact: matti.kankainen@hus.fi

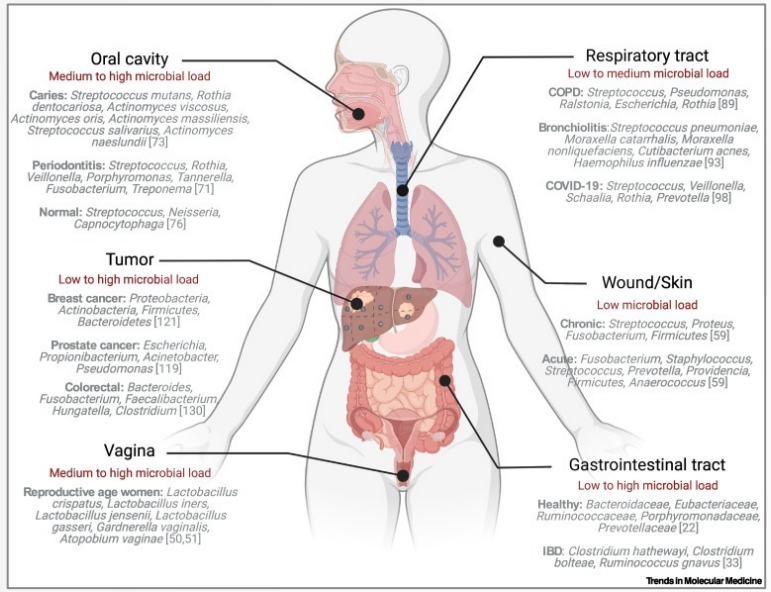
Content

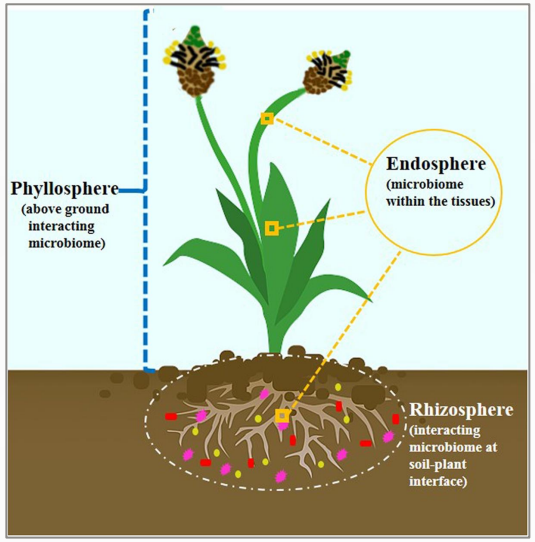
- Microbiome (and host) sequencing
- Design of 16S experiments
- Primary analysis of 16S sequencing data
- Secondary analysis of 16S sequencing data
- Tool suggestions

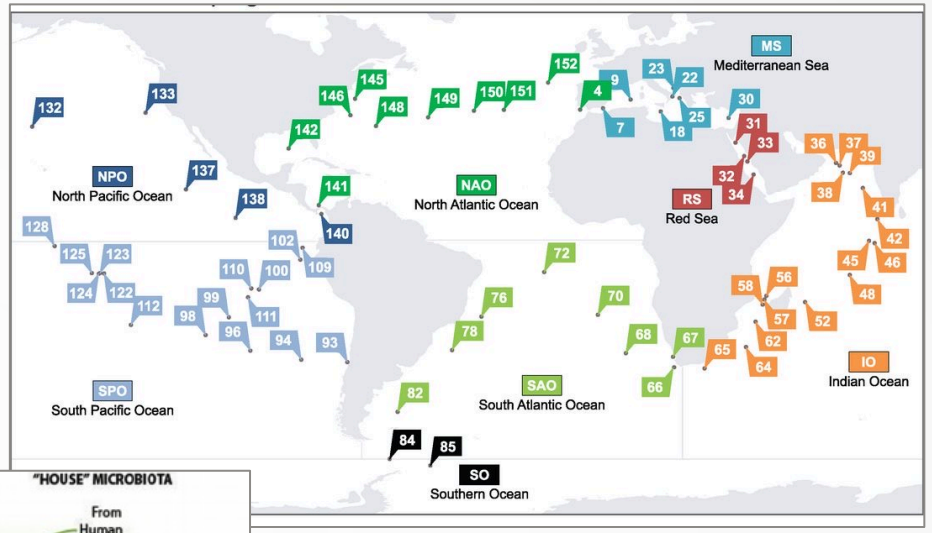
Aims of microbiome analysis

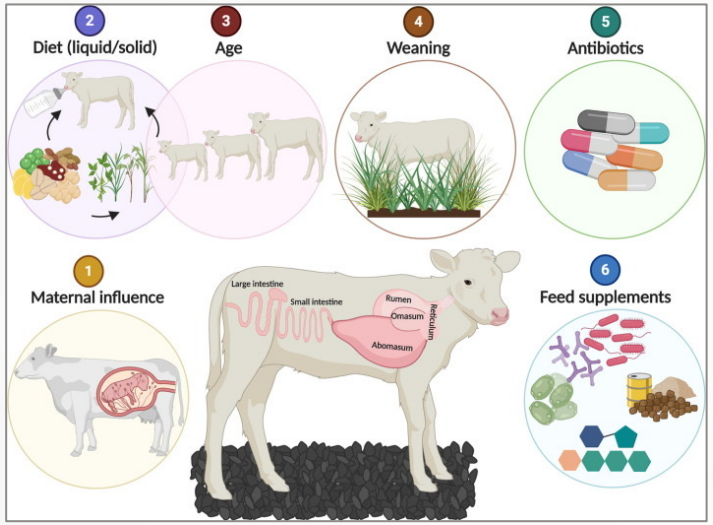
- Identify the taxa present within a sample with a high degree of accuracy
- Investigate the system level functionality and variability of the microbiome
- Study the alterations in the microbiome in health and disease
- Integration of microbiome datasets with other data sources

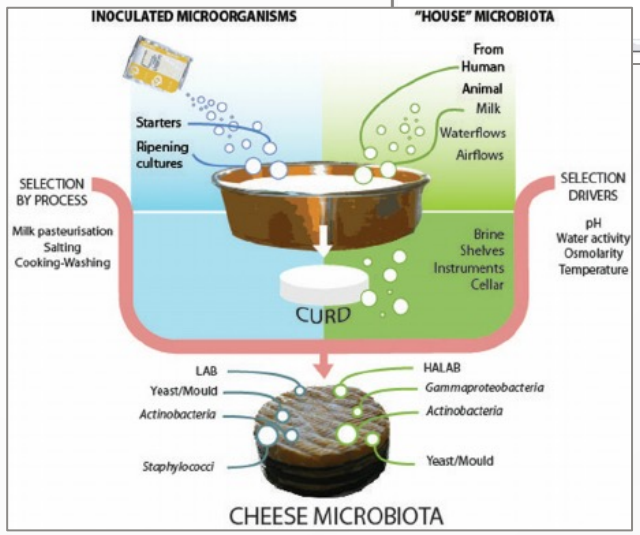
Popular research topics





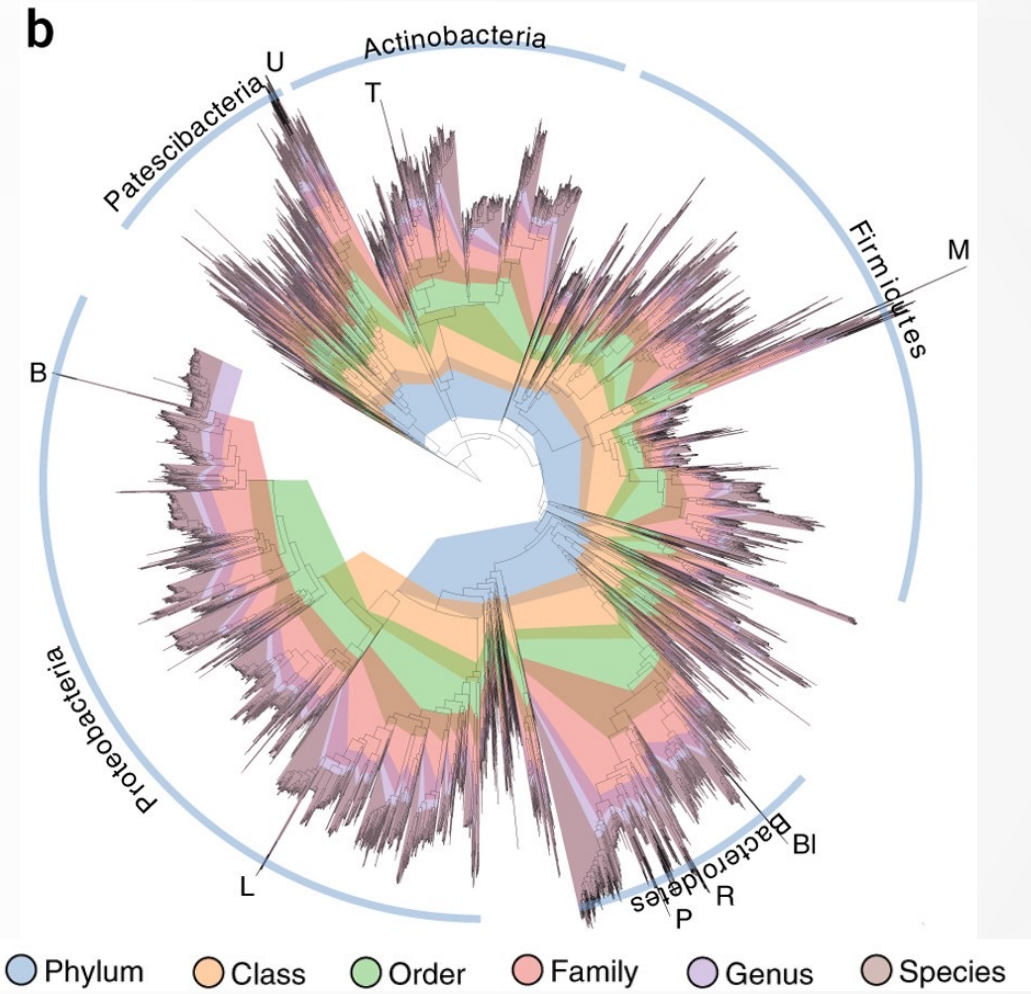






Microbial taxonomy

- Microbial taxonomy may include millions of species (human gut ~4600 species with ~14 million different proteins). Most microbial niches not yet fully characterised
- Microbial organisms arranged to a tree based on similarity: domain > phylum > class > order > family > genus > species > strain
- Microbes change genomic material with their neighbours via horizontal gene transfer
- No function yet assigned to most microbial encoded genes (AI tools like AlphaFold2 may change this)



Microbiome analysis methods

- Amplicon sequencing: targeted sequencing of marker genes like 16S rRNA
- Metagenomics: random sequencing of bacterial genomes
- Meta-transcriptomics: random sequencing of bacterial mRNAs
- Inference of microbiomes from off-target transcriptome / WGS reads
- Single-cell RNA-sequencing

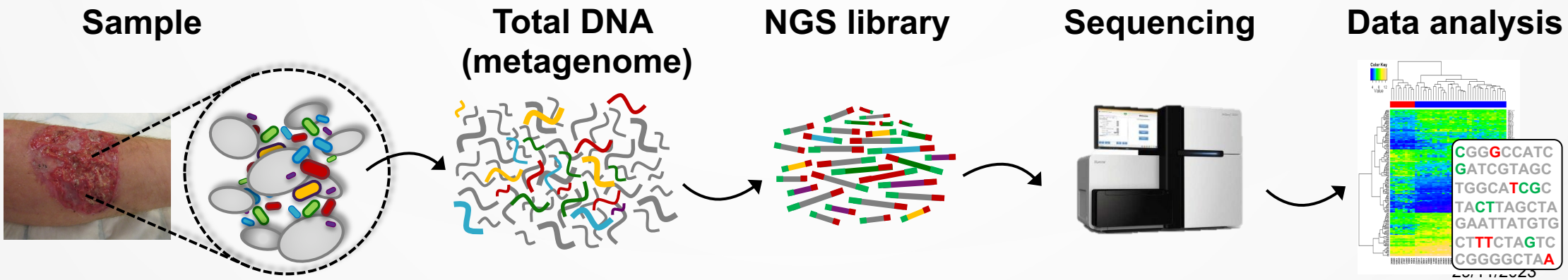
Amplicon sequencing

- Enrichment of a region of interest shared by all organisms (like 16S rRNA and IST) by PCR and high-throughput sequencing of PCR products
- Genus-level classification of bacterial communities. Cost effective (~50 € per sample) and applicable to low-quality samples
- Provides no direct information of functional / gene repertoire of the community. Highly Sensitive to contamination (from air, reagents, instruments, researchers)

Primers **Sampling** **Extraction** **Amplification** **Sequencing** **Bioinformatics**

Metagenomics

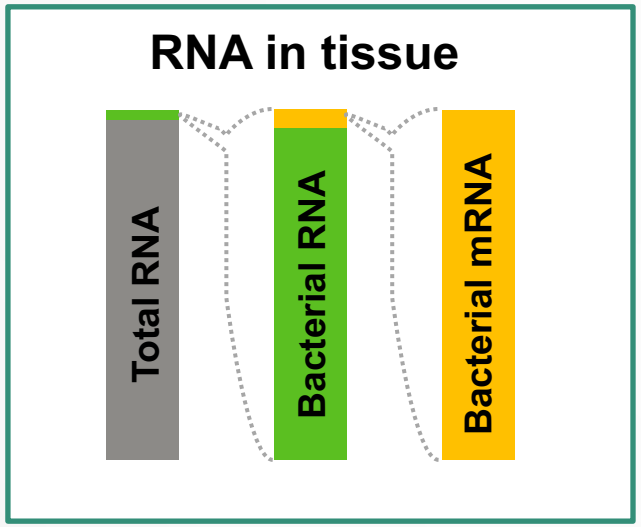
- Random sequencing of microbial DNA from the sample by NGS
- Strain level-assessment of bacteria communities. Expensive (~500 € per sample) and requires average quality samples. Insufficient depth can leave organism and genes uncharacterized
- Information on genes and/or functional repertoires
- Samples rich in host DNA troublesome as microbial DNA cannot easily be enriched

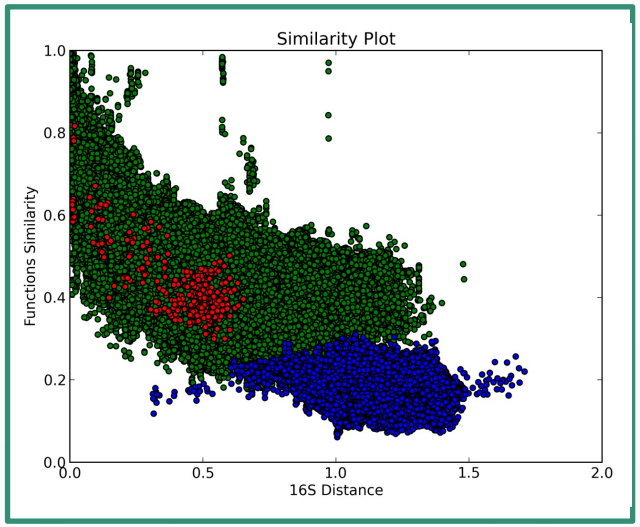


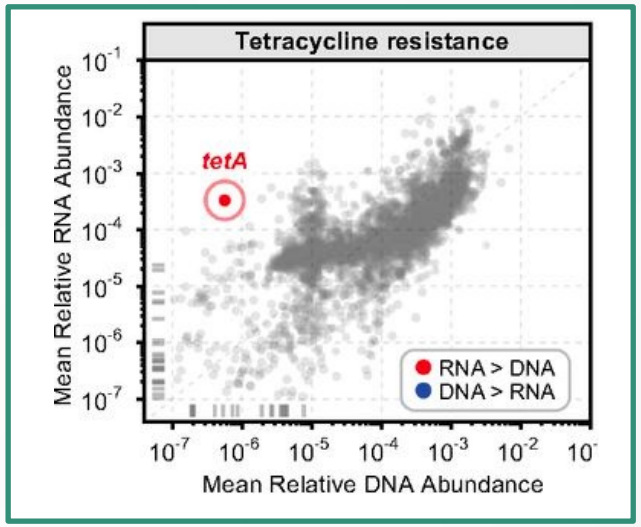
Metatranscriptomics

- Portrait of microbial transcripts in a sample providing information on active microbes and their gene activities

- Strain level-assessment of bacteria communities. Moderate cost (~300 € per sample) and requires good quality samples (RIN > 9). Amount of host RNA vs. microbial RNA in host-microbe samples: 10-30 pg vs. 100 fg per cell. Host RNA can be depleted



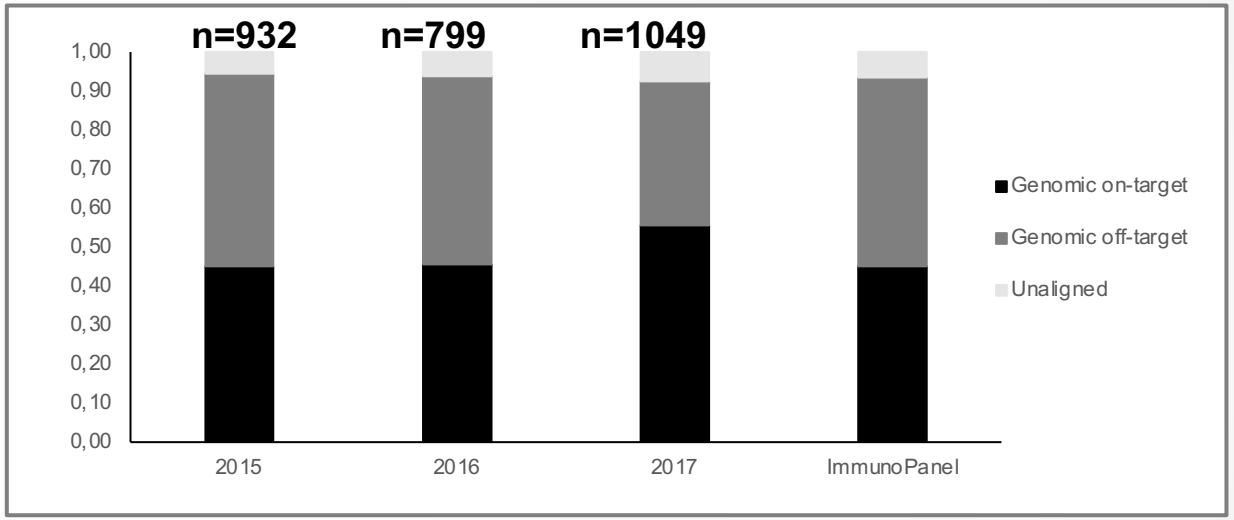




Human genomic data

- WGS and ribo-depletion/total RNA-sequencing don’t differentiate host and microbe molecules and contain microbial reads in the same portion as there were microbes in the sample (up to 30% in buccal samples)

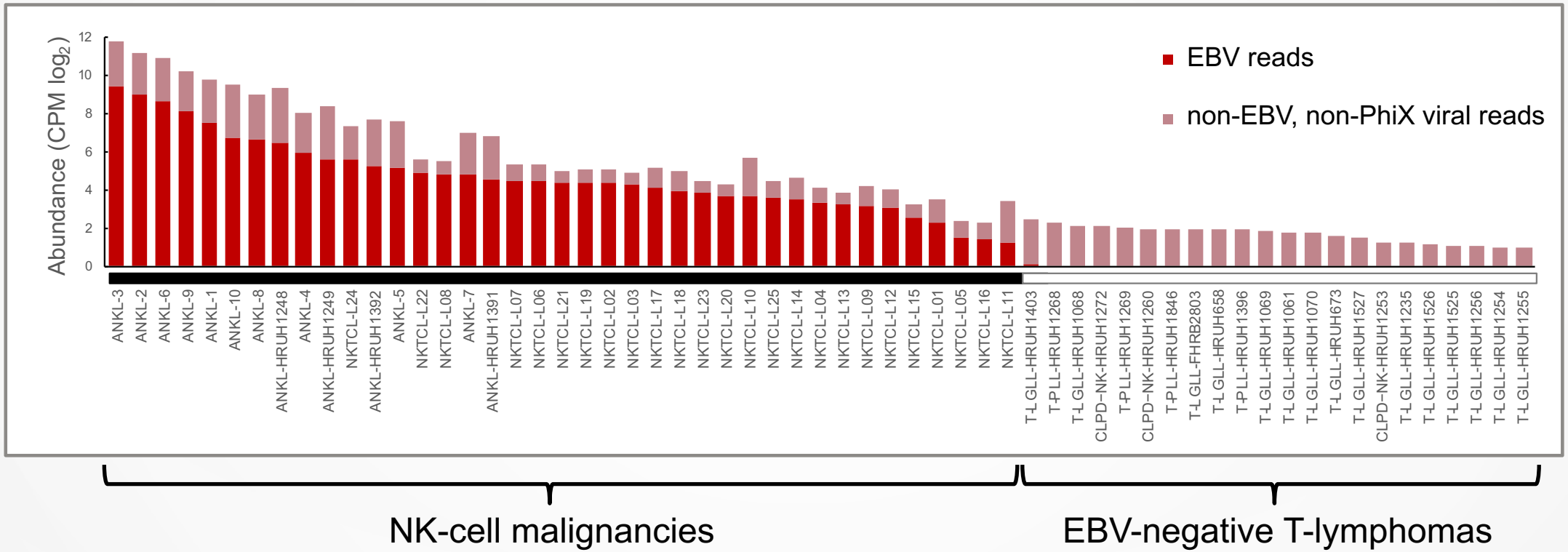
- 20-50% of reads in WES and poly-A capture RNA-sequencing analyses are from unintended regions. Used often in human genetic studies to infer copy-number variants. Provide also source material for microbial inference



- Same resolution as metagenomics and metatranscriptomics (strain). Add-on analysis that is free-of-charge. Insufficient microbial depth leaves organism uncharacterized

Microbe inference from host WES data

- WES applied to 39 EBV-positive and 22 EBV-negative cancers. EBV mapping reads found only in EBV-positive samples. Differences in hybridization may exclude use of EBV load estimates in abundancy estimation

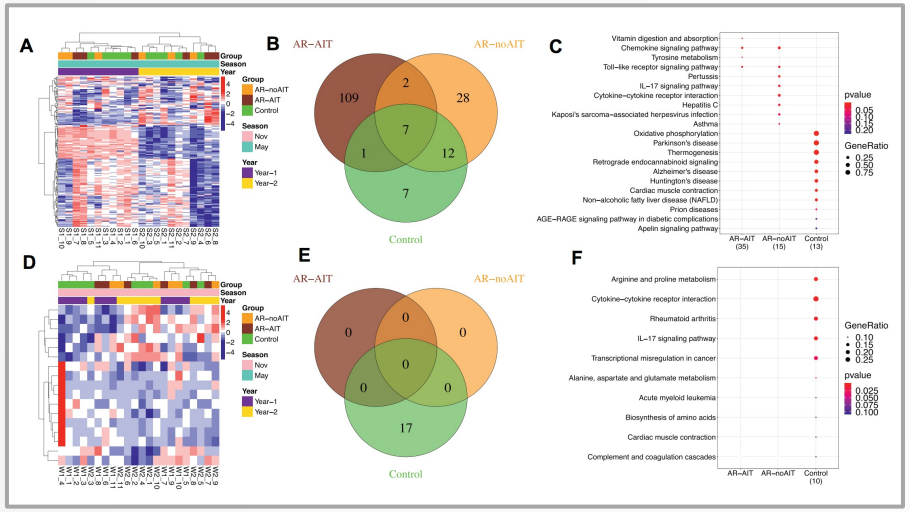


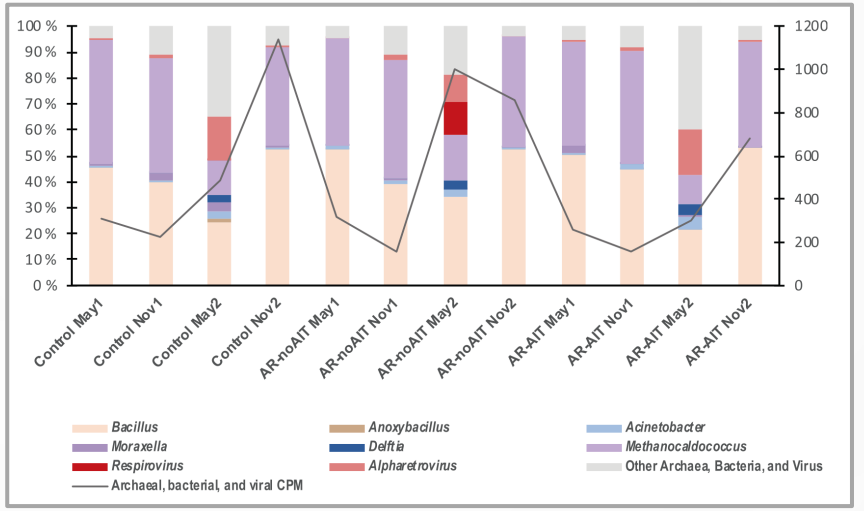
Dufva *et al.*, Nat. Comm., 2018

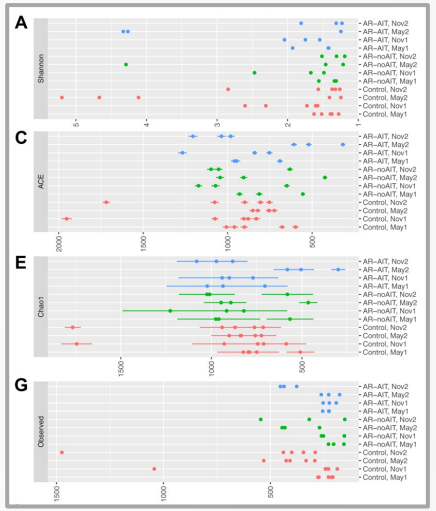
Microbe inference using host RNA-seq

- Host gene expression analysis of nasal epithelial samples from controls and allergic patients with (AR-AIT) or without (AR-noAIT) immunotherapy at 4 consecutive seasons identified largest transcriptional reprogramming and restoration of gene expression towards normal in AIT group

- Nasal samples contained on average 16340 (0.05%) reads mapping to microbes per sample. Microbial diversity between winters increased most in AR-AIT group

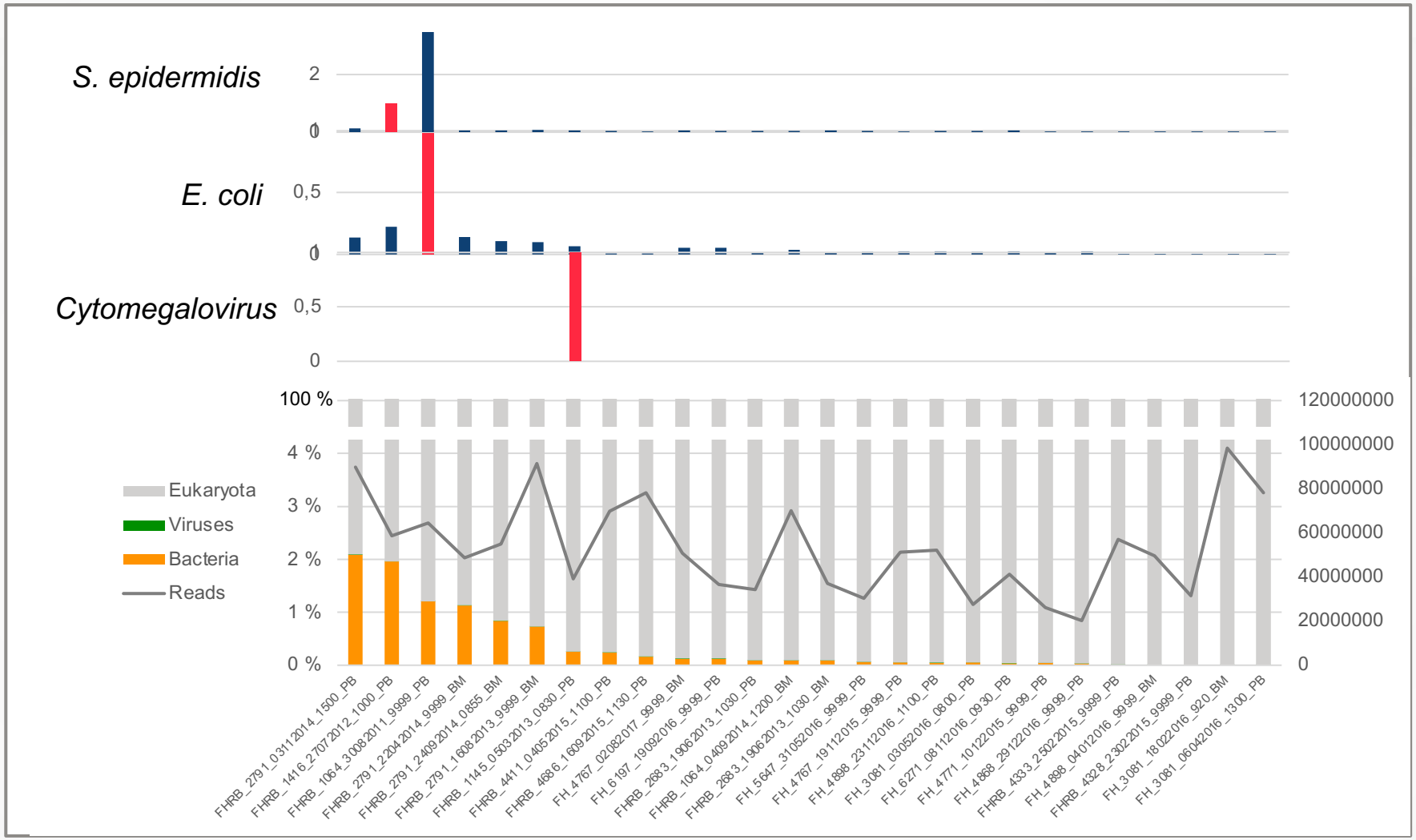






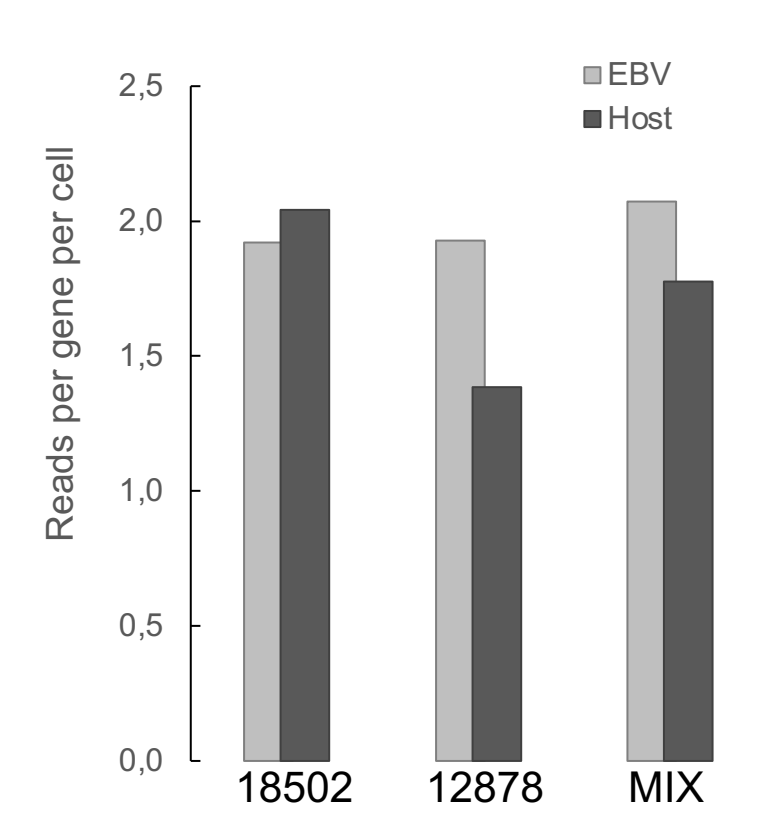
Hanif *et al.*, JACI, 2019

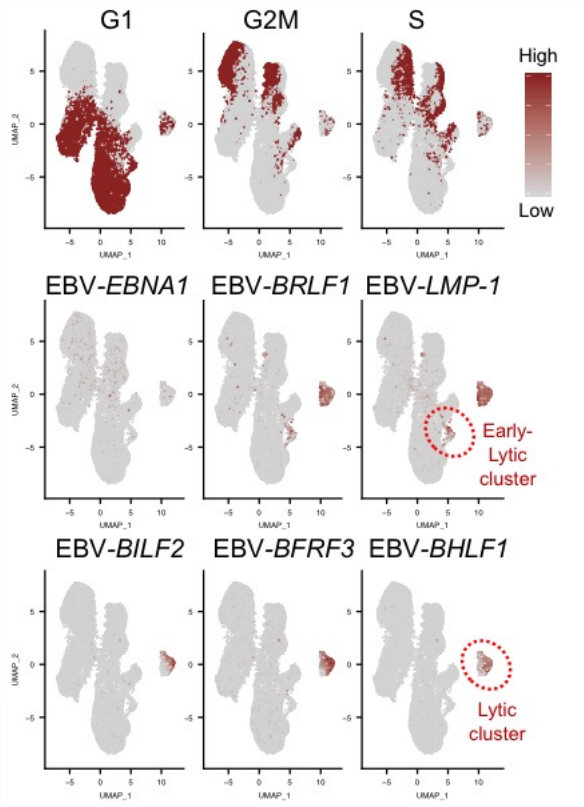
RNA-seq reproduces clinical MB calls



sc-RNA-seq pathogen classification

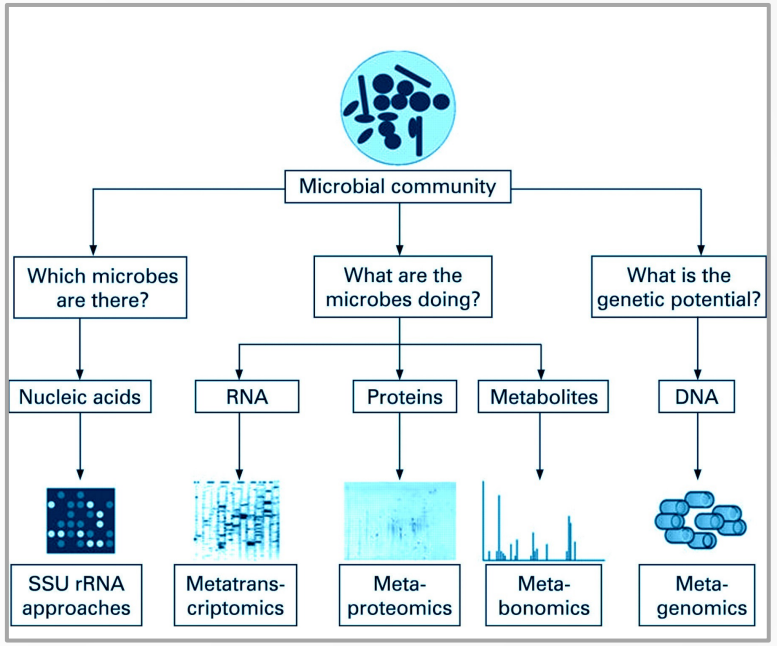
- scRNA-seq data has similar amounts of human and EBV reads (per gene) and allows to access expression of viral genes at the cell-level similar to human gene expression





Method comparison

- Price: 16S (50 €) < Metatranscriptome (300 €) < Metagenome (500 €)
- Resolution: 16S (genus) < Metagenome (strain) < Metatranscriptome (strain)
- Sample quality: 16S (low) < Metagenome (moderate) < Metatranscriptome (good)
- Functional information: 16S (no) < Metagenome (yes) < Metatranscriptome (yes)



- Data analysis: 16S (easy) < Metatranscriptome (moderate) < Metagenome (difficult)

Discuss with you neighbour(s)

You are about to study human gut microbiota before and after use of antimicrobial drugs. Which of the many methods you would use to achieve this.

Content

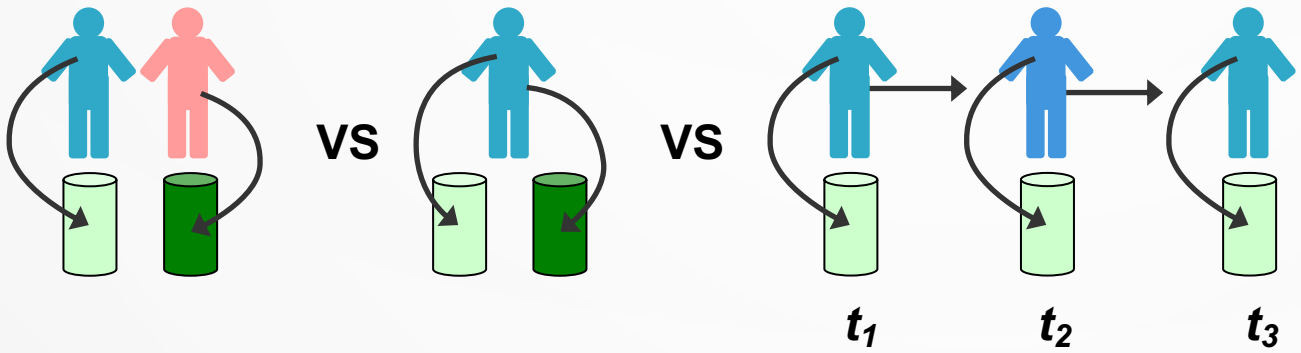
- Microbiome (and host) sequencing
- Design of 16S experiments
- Primary analysis of 16S sequencing data
- Secondary analysis of 16S sequencing data
- Tool suggestions

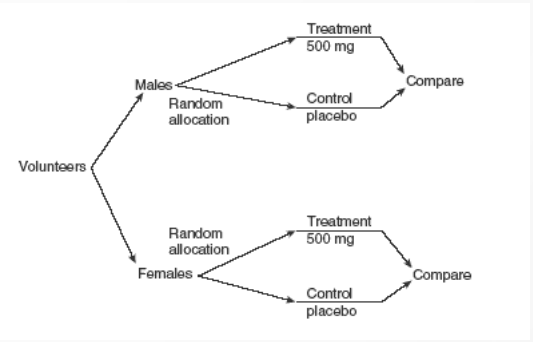
Study design

- Conduct power analysis to determine required sample number. Use literature search to estimate variable number (i.e. OTU number) and effect size (e.g. mean abundance in treatment vs. control group of OTUs of interest)

- Prefer paired samples and/or blocking in sample design instead of independent study subjects. Prefer age, gender, body-site, etc matched independent study subjects

- Avoid generation of batches (e.g. samples prepared using two different kits) and use of technical replicates and/or mixing of technical and biological replicates

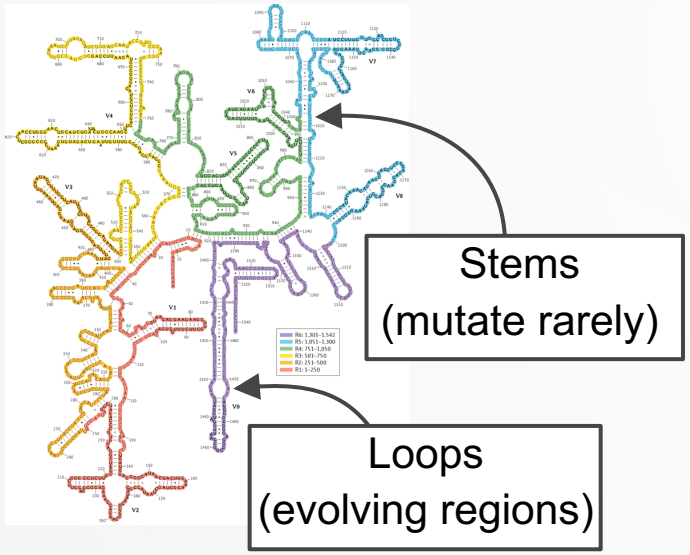


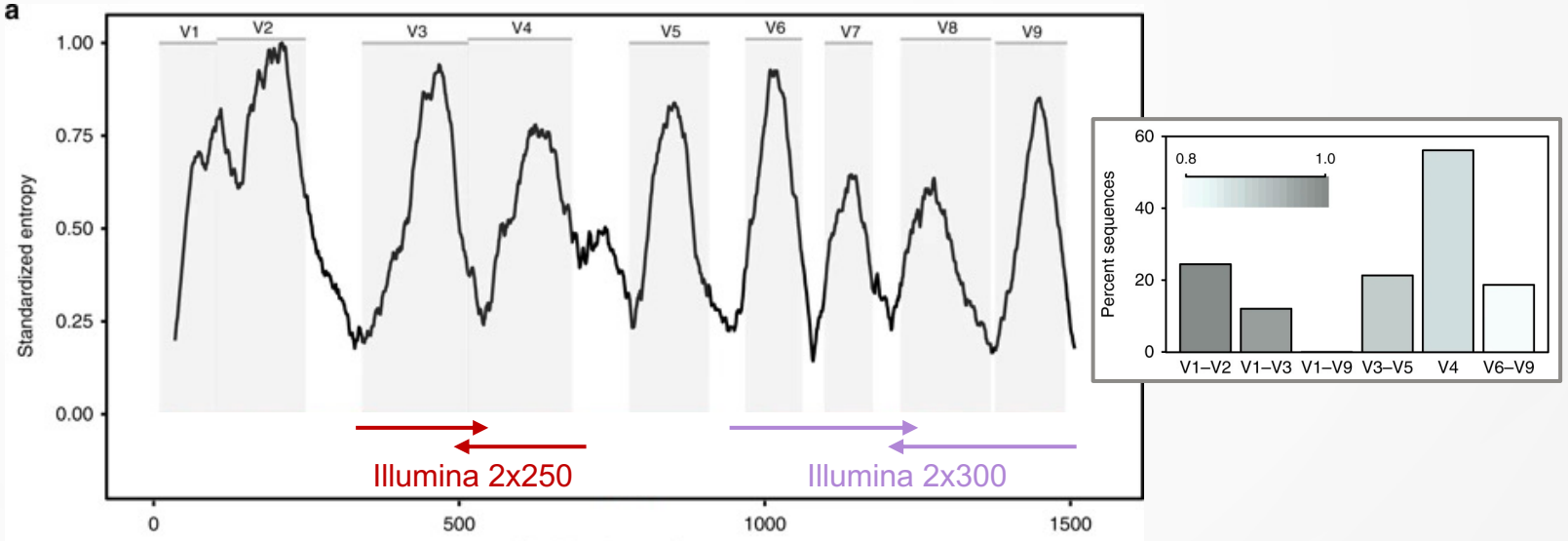


Primer design

- Bacterial 16S rRNA gene includes 9 hyper variable (loop) regions (V1-V9) interspersed throughout the highly conserved (stem) sequences

- Longer regions provide more accurate and comprehensive classifications, but sample fragmentation and/or limitations of the chosen sequencing strategy (i.e. >450-550 bp in Illumina) may prevent their analysis





16S controls

- Use of experimental controls is critical to avoid false findings from contaminations especially when analysing samples with low bacterial/fungal/viral biomass

- Controls should be prepared in parallel to study samples, undergo subsequent steps, and be sequenced

- Common controls to include:

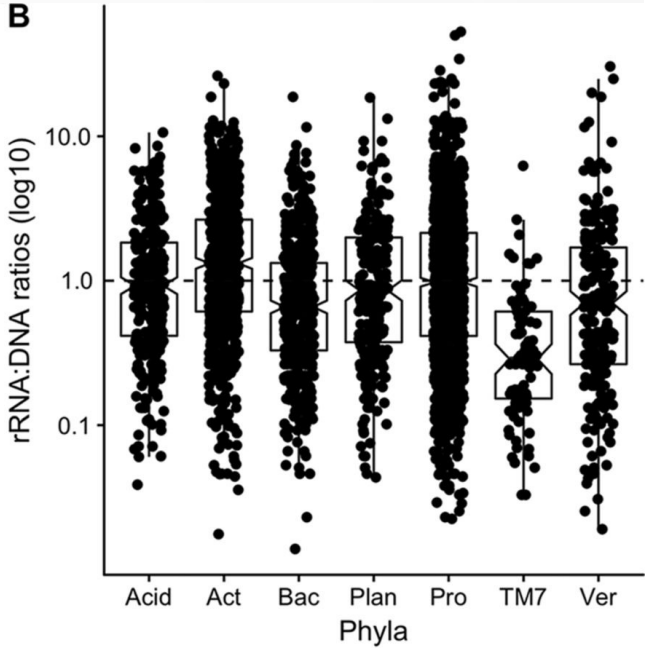
  1. Negative sampling controls like sterile swab opened in the sampling site
  2. Negative extraction controls of blank DNA extractions without input material
  3. Negative PCR amplification controls of blank amplifications without input material
  4. Positive mock community and/or others spike-in bacterial controls

16S rRNA vs. rDNA

- DNA is stable and DNA from nonviable microbial cell can remain in the sample up to 14 days. Provide means to characterize unchanging and rather stable microbiota

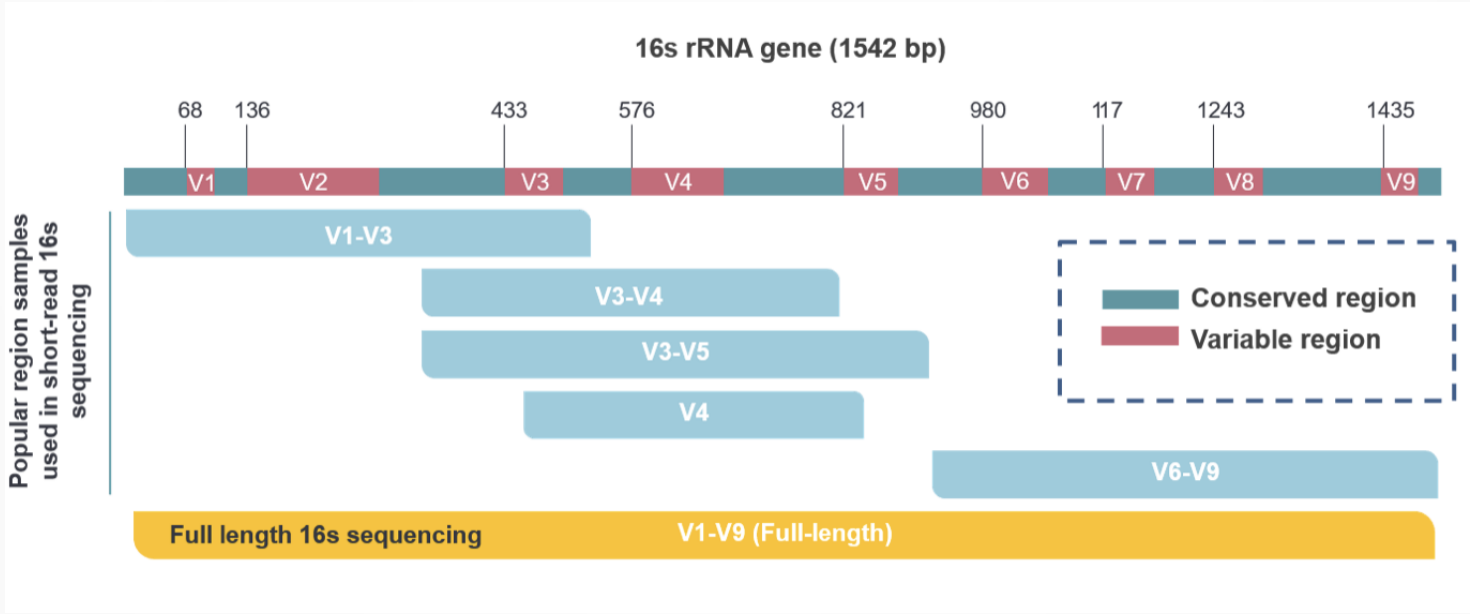
- rRNA has short half-life (minutes to hours) and serves as a proxy of recent metabolic activity and/or growing or active communities. Provide means to characterize sudden changes in activity

- rRNA and rDNA estimates typically correlate. rRNA:dDNA ratios used to find active bacteria, but the robustness of ratios has been questioned / criticized



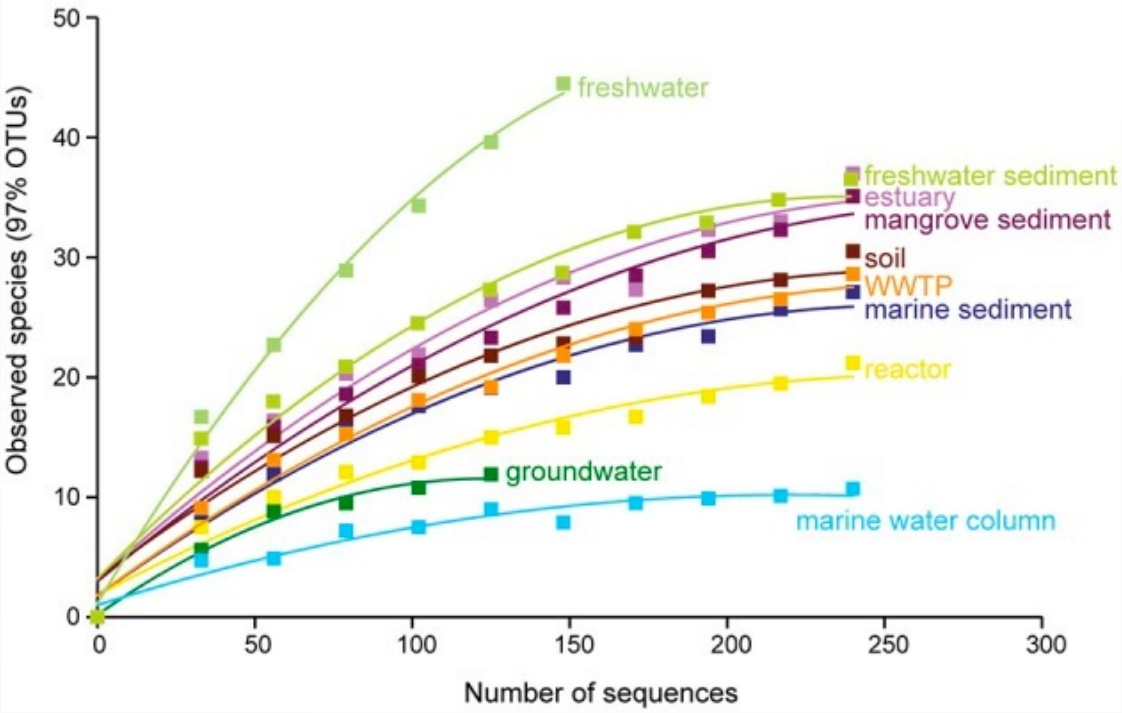
Sequencing instrument

- Short-read sequencing instruments (i.e. Illumina) still most popular. Typical sequencing read-length is ~250 bo although 100 bp read-length also used
- Long-read technologies allow nowadays sequencing of almost complete and closed bacterial genomes / target amplicons. Their usage improve accuracy and minimize the need of assembly algorithms in further steps



Sequencing depth

- Adequate sequencing depth needed to cover the full microbial diversity. Proper depth is study specific, but can be estimated by doing pilot sequencing and rarefaction analysis
- A good starting guess could be 100k Illumina paired-end reads. If full range of species not achieved, consider generating more reads and/or re-sequencing



Other aspects to consider

- Recommendable to use stabilization buffers can preserve microbial community DNA and RNA
- Amount of host DNA/RNA (1% in stool vs. 80% in skin biopsy) and microbial DNA/RNA (100 cell per ml in blood vs. 10^{14} in the colon) in the sample and strategy to enrich microbial biomaterial
- Sample logistic and storage. Freezing of samples immediately and/or employing a fixative that stops microbial growth and stabilizes biomolecules

Discuss with you neighbour(s)

You are about to study human gut microbiota before and after use of antimicrobial drugs using 16S amplicon sequencing. How would you design your experiment.

Content

- Microbiome (and host) sequencing
- Design of 16S experiments
- Primary analysis of 16S sequencing data
- Secondary analysis of 16S sequencing data
- Tool suggestions

16S analysis pipeline

Input: NGS reads

Read trimming
(Trimmomatics)

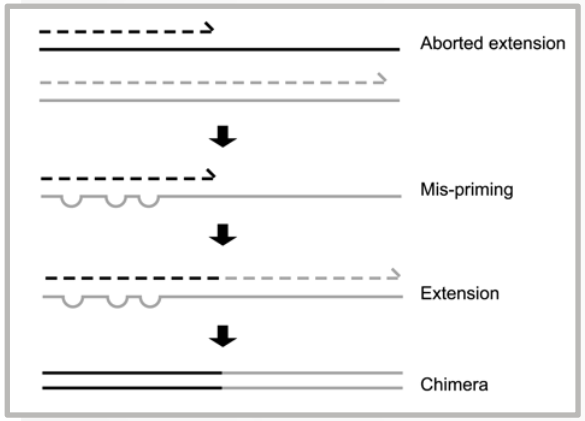
Paired-end assembly
(FLASH/PEAR)

Remove artifacts
(Mothur)

Unique/Cluster reads
(Mothur)

Remove chimeras
(VSEARCH)

Output: Good reads



Good reads

Classify reads
(Wang)

OTU clustering
(opticlust / UPARSE)

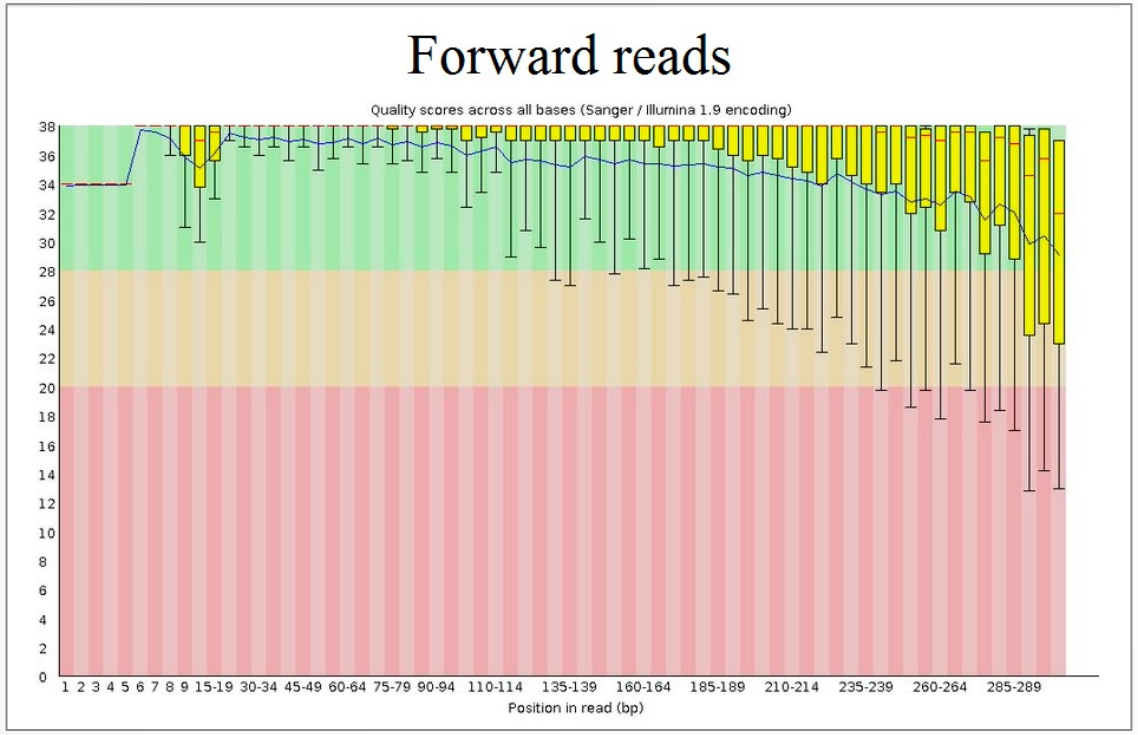
Taxonomy classification
(Mothur)

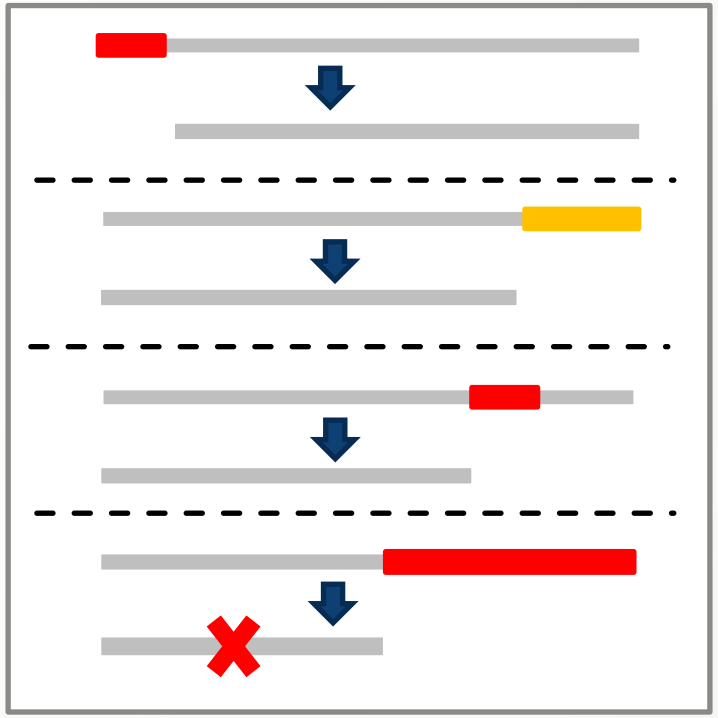
Function prediction
(PICRUSt 2.0)

OTU-+ function table

Read trimming

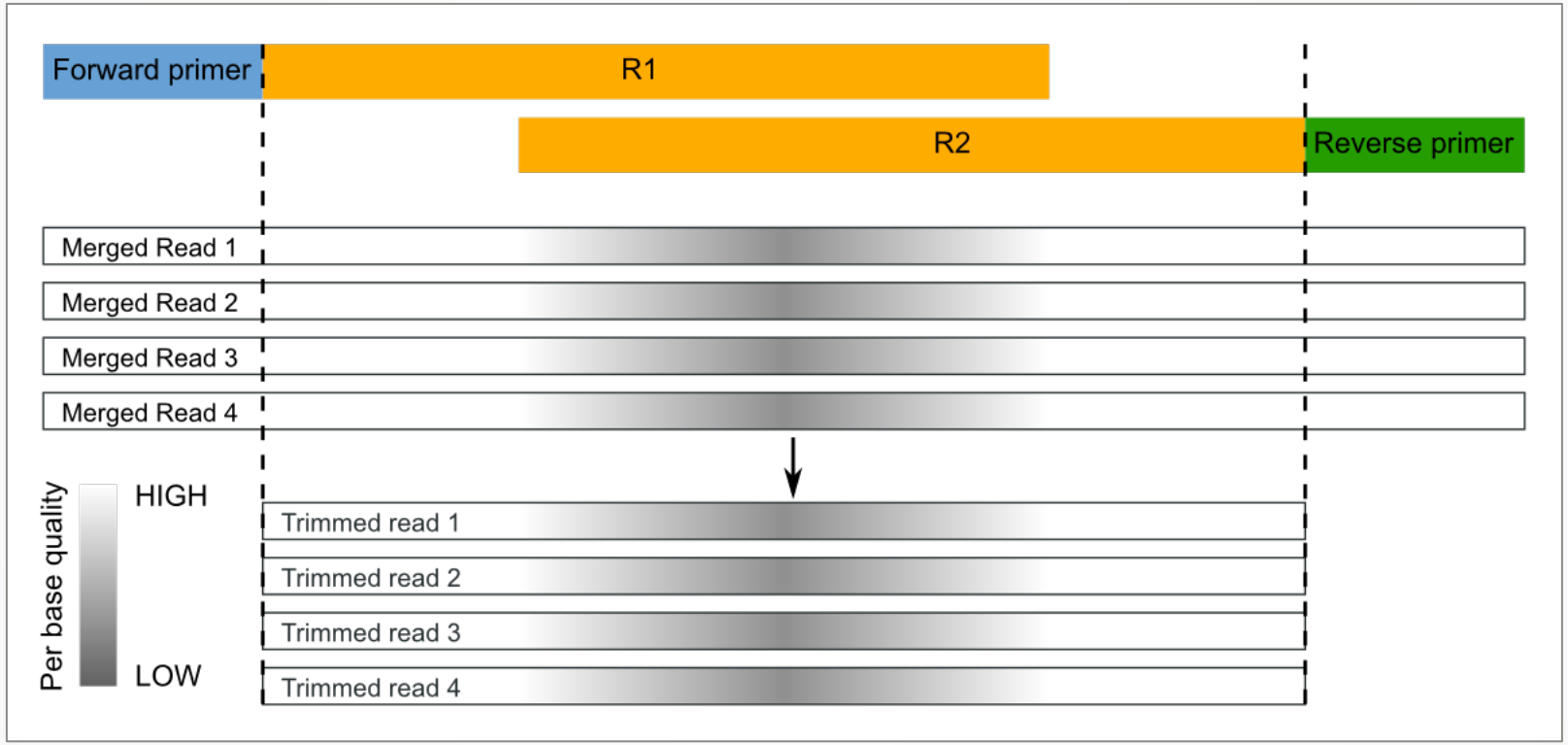
- Inadequate read regions and reads removed prior further analyses to minimize false findings by, for example, removing low-quality, technical adapters, overly short reads, and reads with many mismatches. Tools like Trimmomatic, Mothur, etc commonly used





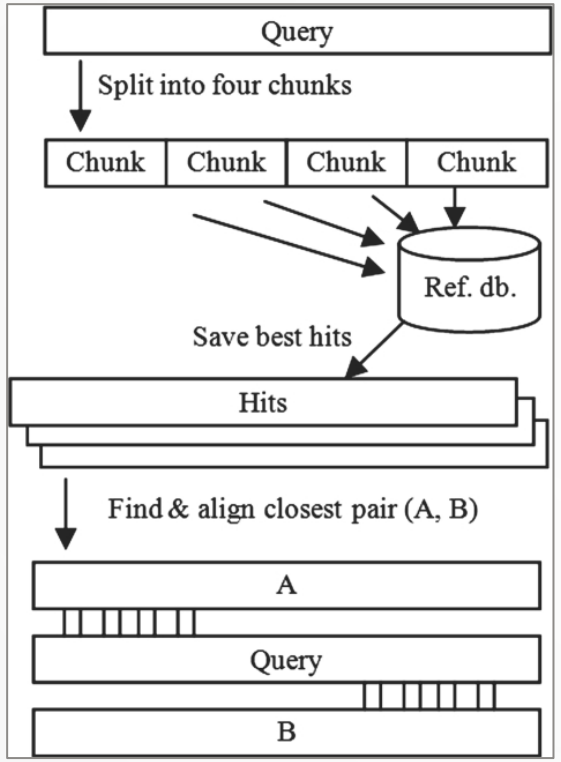
Paired-end assembly

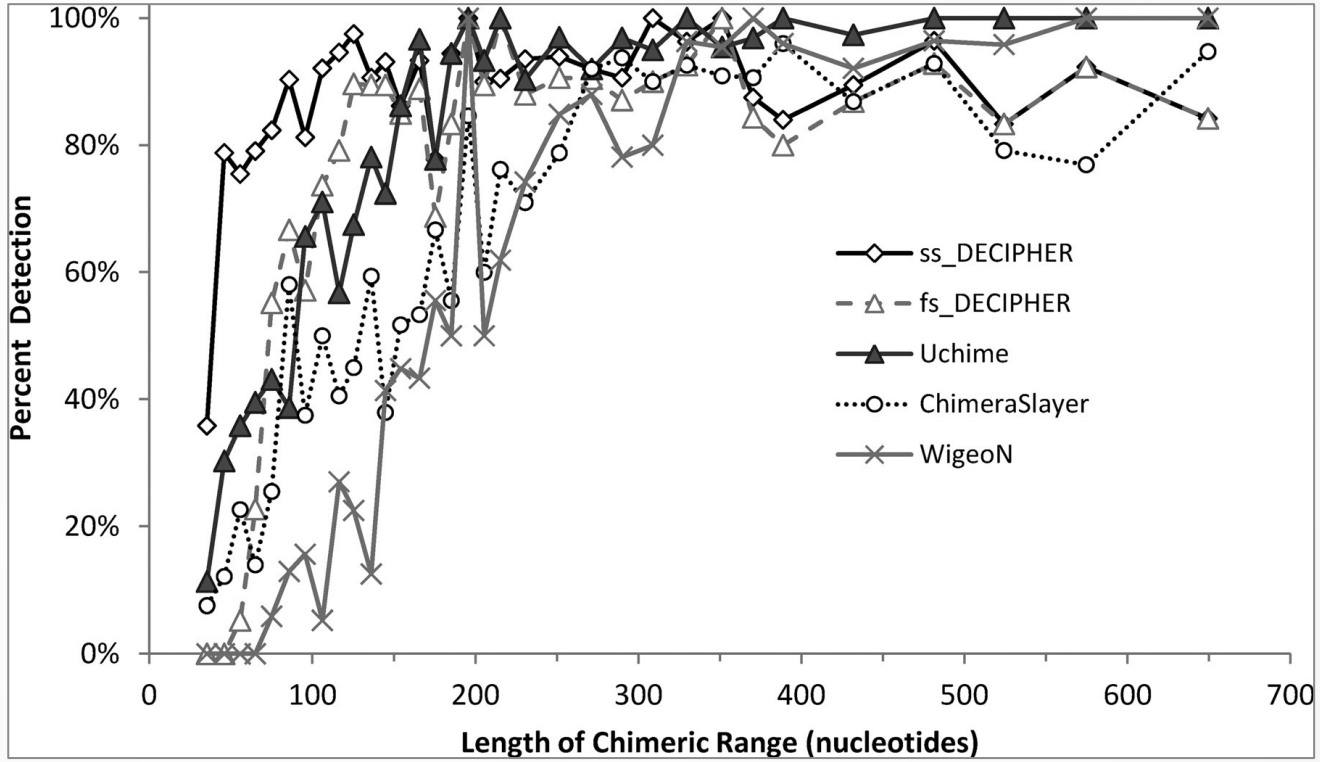
- Generation of single full-length sequences from read-pairs and ignorance of read-pairs failing to assembly improves accuracy and coverage in sequence classification. Tools like Pear, FLASH, and Panda-seq commonly used in paired-end assembly



Chimera removal

- Chimeras are (PCR) artefact sequences formed by two or more biological sequences incorrectly joined together. Detected using tools VSEARCH. UCHIME2, or DECIPHER searching for reads mapping into distinct references



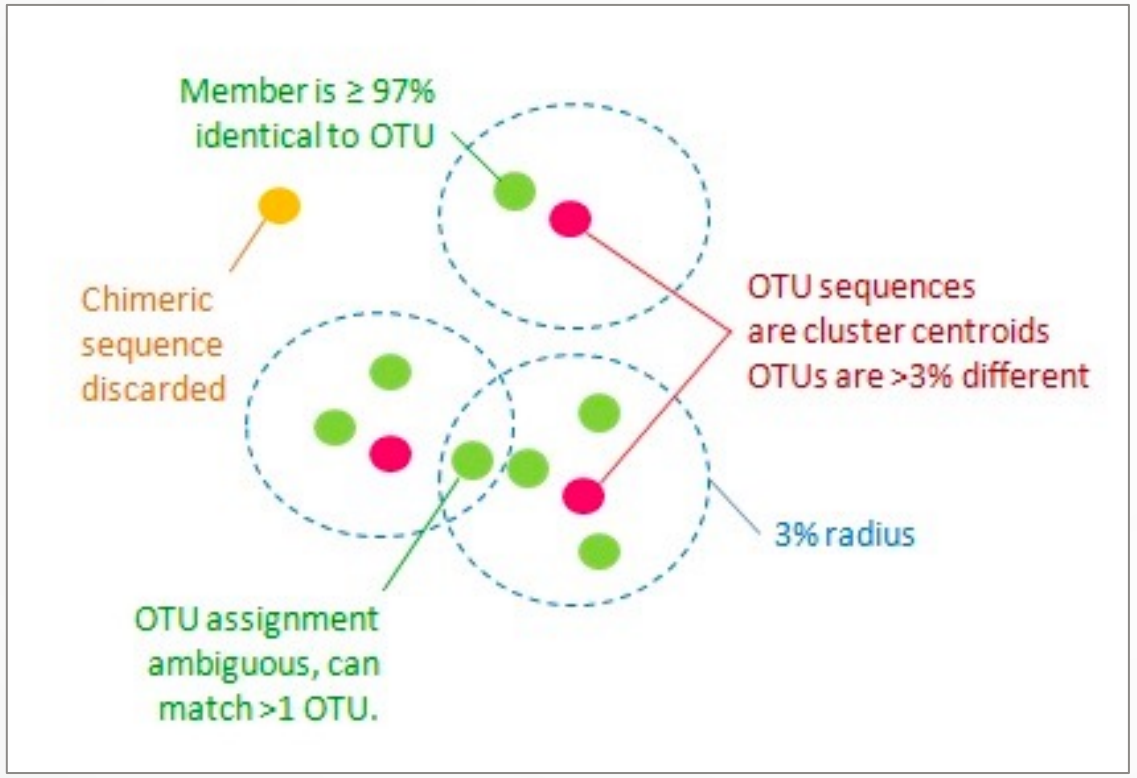


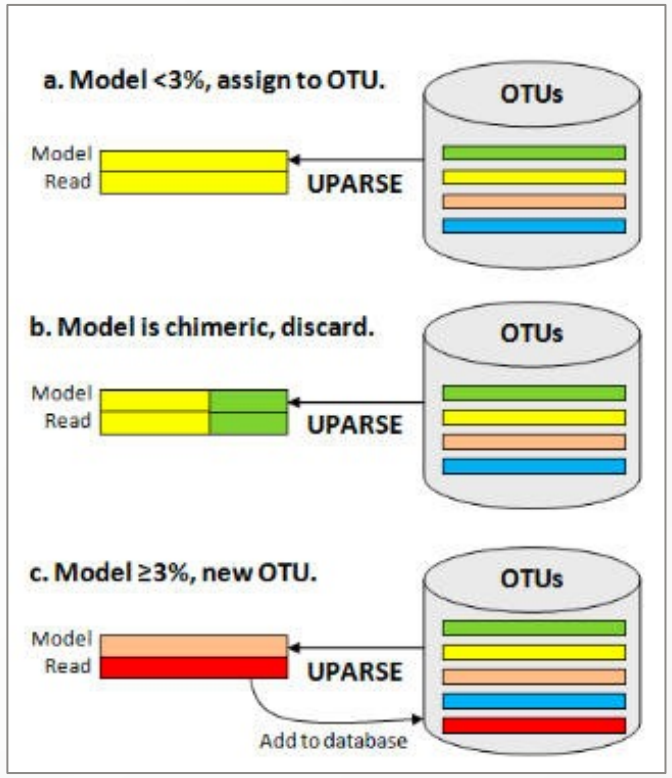
OTU clustering

- OTU clustering groups sequences representing the same taxonomic unit of a bacteria species (>97%) or genus (>95%) together and provides source for abundance counts
  - De-novo OTU: sequence reads sufficiently similar to one another (e.g. 97% similarity) clustered together. All reads within one cluster >97% similar to each other
  - Phylotyping/closed-reference clustering: sequences compared to a curated database and sequences matching the same reference assigned to the same OTU. Reads within the cluster >97% similar to the reference, but can be more dissimilar from each other
  - Open-reference clustering: combination of above. Sequences without reference match de-novo clustered
  - Amplicon sequence variant (AVS) correction: infer the biological sequences and distinguish sequence variants by assuming them to be rare and randomly distributed

De-novo OTU clustering

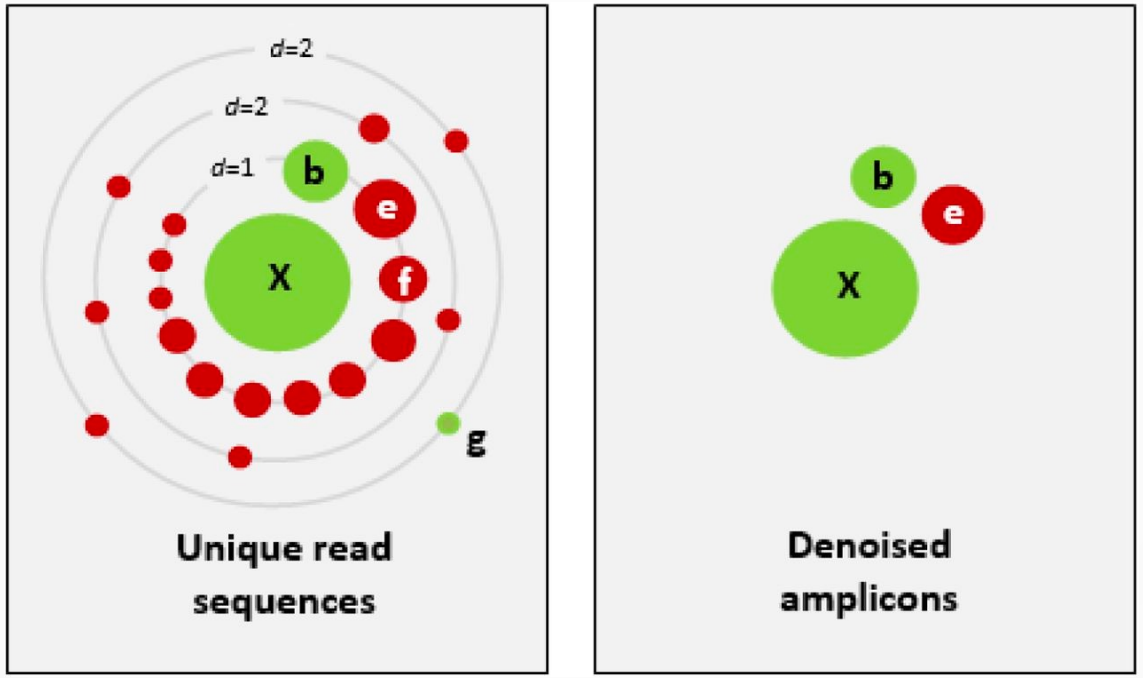
- De-novo OTU clustering probably (still) most popular strategy. Several computational solutions exists including hierarchical clustering, Bayesian clustering, greedy search algorithms





Amplicon sequence variant (AVS)

1. Construct a graph where nodes represent unique reads and edges represent number of base differences between reads
2. Nodes with few reads considered sequencing errors if in vicinity of a larger node and merged to the larger nodes



Taxonomy classification

- De novo OTU: representative / all sequence reads within the OTU compared against database and information of all best hit matches used to assign taxa to the OTU. Note that *de novo* OTUs defined in two different data sets may not be comparable

- Phylotyping/closed-reference clustering: taxonomic classification of reference assigned to the OTU. Use of the same reference database allows comparison of results between runs, but **sequences lacking reference database matches** are ignored

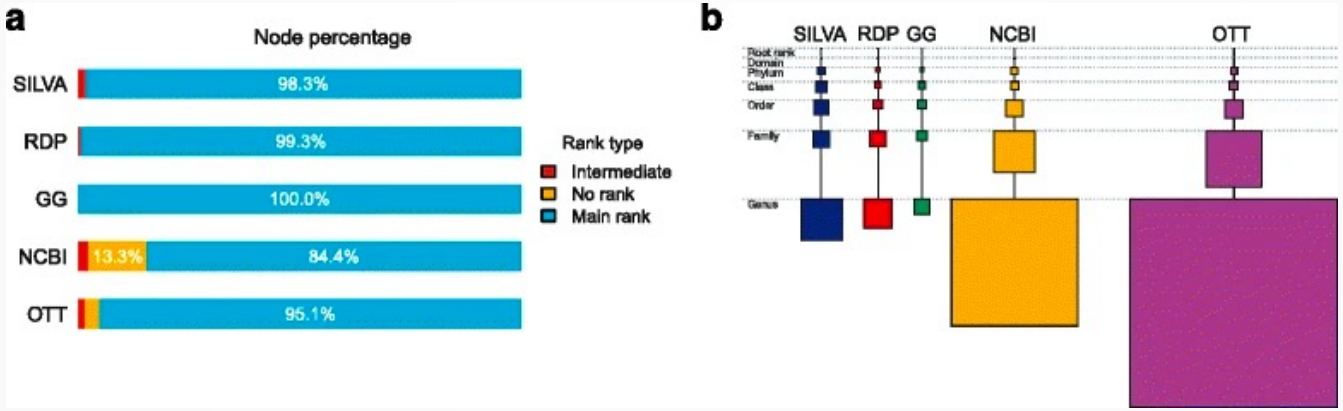
- Amplicon sequence variant (AVS) correction: representative / all sequence reads within the OTU compared against database and information of all best hit matches used to assign taxa to the OTU

Reference databases

- SILVA: Manually curated genus level database for Bacteria, Archaea and Eukarya. Taxonomic rank information for Archaea and Bacteria from *Bergey's Taxonomic Outlines* LPSN. Default in Mothur. >190k sequences

- EzBio: Manually curated species level database. Contains also 16S sequences from genome assemblies that are of higher quality than amplicon data. >63k sequences

- Greengenes (GG): Database dedicated to Bacteria and Archaea. Classifications are based on automatic *de novo* tree construction. Default in QIIME. **Not updated.** >99k sequences



Discuss with you neighbour(s)

You are about to study human gut microbiota before and after use of antimicrobial drugs using 16S amplicon sequencing. How would you analyse your amplicon data

Content

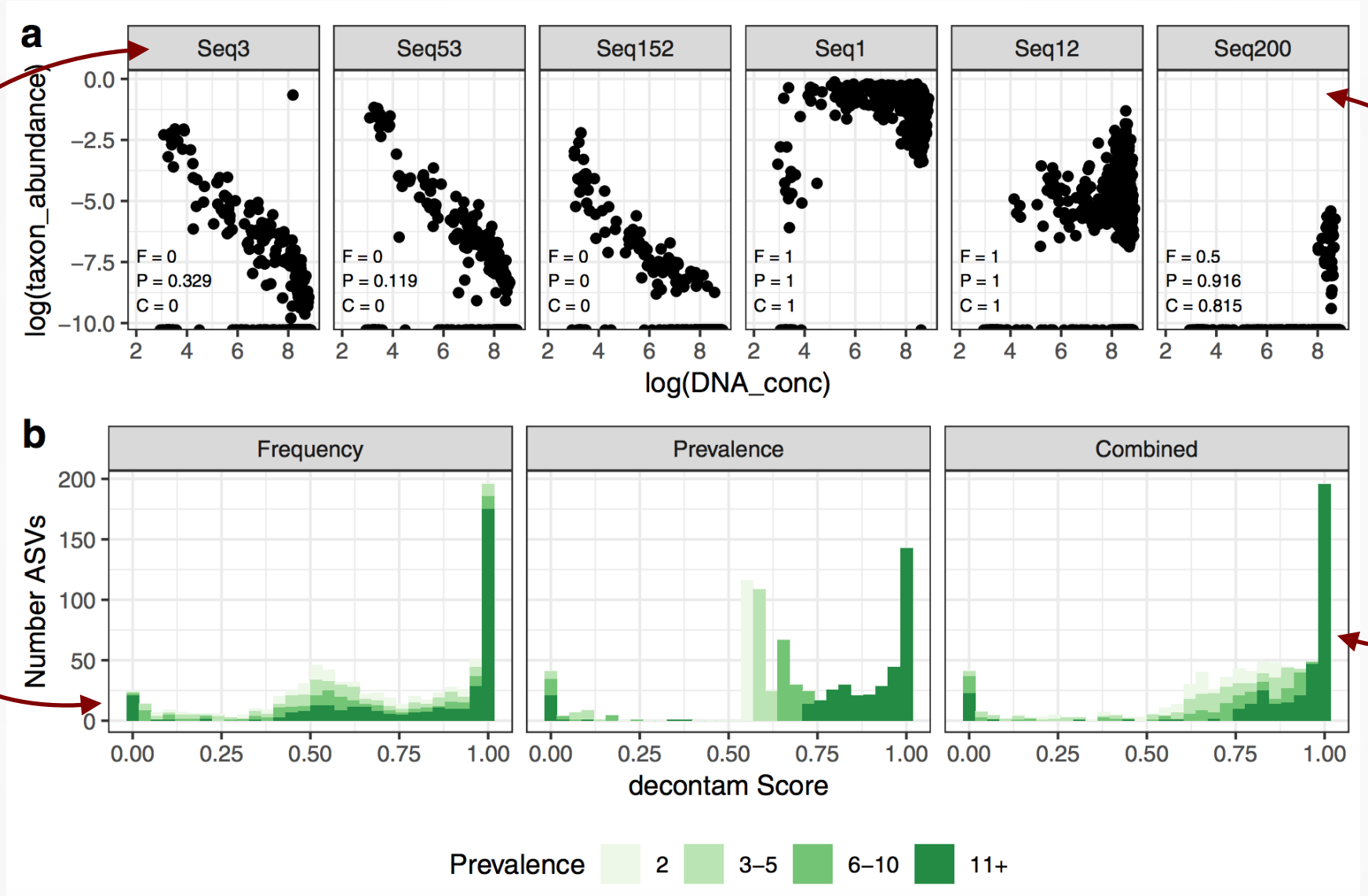
- Microbiome (and host) sequencing
- Design of 16S experiments
- Primary analysis of 16S sequencing data
- Secondary analysis of 16S sequencing data
- Tool suggestions

Contamination read filtering

- Contaminants can emerge easily especially in samples with low microbial mass and needs to be removed to avoid false findings
  - Frequency of contaminating taxa inversely correlated with sample DNA concentration
  - Contaminating taxa typically show higher prevalence in control than in true samples

- Automated algorithms for contaminant removal
  - Decontam: correlation of OTU frequency and DNA conc and chi-square test of enrichment of OTU in controls used to detect contaminants. OTUs likely to be contaminants flagged
  - SourceTracker: uses Bayesian mixtures to identify the proportion of a sample comprising taxa from known source environment / samples (e.g. fraction of hand microbiota present in nose sample)

Filtering of contamination reads

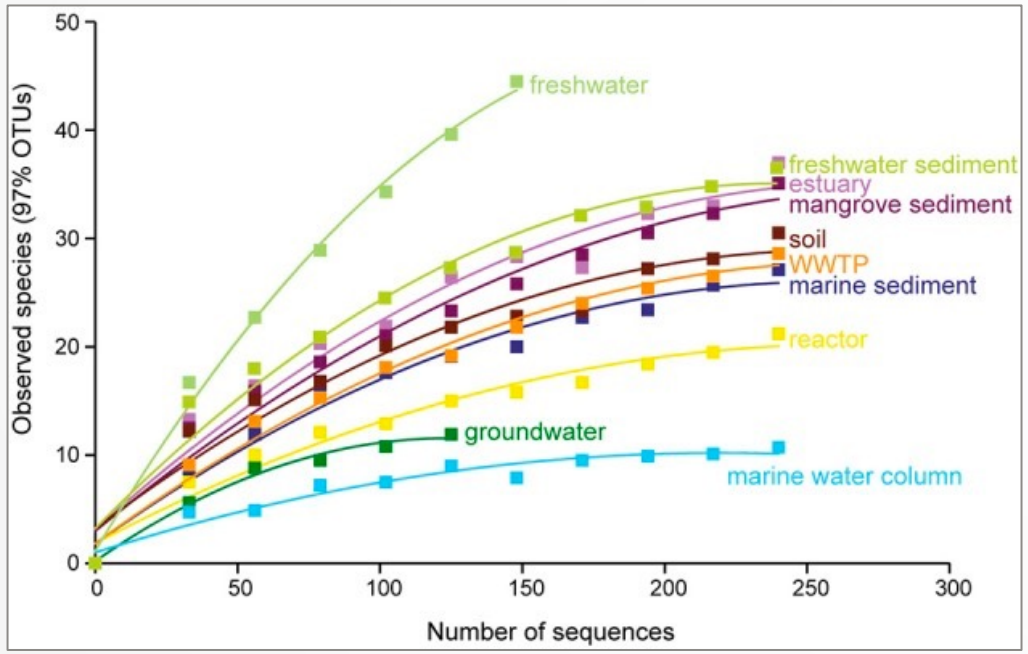


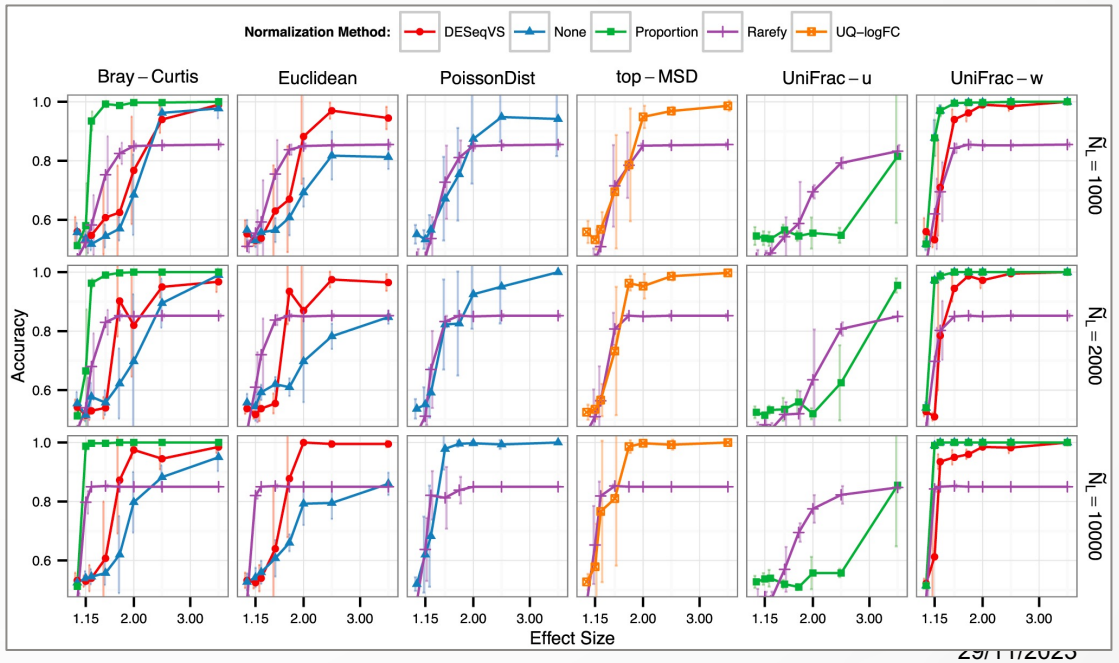
Bad OTUs with low score

Good OTUs with high score

Rarefaction

- Rarefying of counts means random selection of reads from the original read set. Used to make read depth between samples identical by sub-sampling larger libraries to the size of the smallest and/or estimate degree of completeness of species richness
- Value and/or usefulness of rarefaction to minimum depth and use of proportion data has been criticized extensively





Differential abundance testing

- Differential abundancy testing aims to identify OUTs / taxa with abundancy differences between study group (e.g. healthy vs. disease)
  - Nonparametric tests like the Mann-Whitney/Wilcoxon rank-sum test for two groups and the Kruskal-Wallis test for multiple groups. Rarefaction typically used to normalize data
  - Parametric models like limma, edgeR, and DESeq2 that use generalized linear models (GLM) and assume counts to follow negative binomial (NB) distribution. Data typically normalized by computation of size-factors
  - Zero-inflated GLM / Gaussian that address sparsity and unequal sampling depth by separately modelling zero counts. Data typically normalized by computation of size-factors

DESeq2 protocol

1. Construct DESeq object from OTU or taxa counts at certain level
2. Design appropriate statistical model
   - Unpaired t-test: ~ Group
   - Paired t-test: ~ Subject + Group
   - Nested: ~ Group + Group:Subject + Group:Treat
   - 2-factor: ~ Group + Time // ~GroupTime with cont = (g1t1-g1t2)-(g2t1-g2t2)
3. Normalize data and perform testing
4. Compute shrunken log2 fold changes (LFC)
5. Correct p-values by FDR

P-value adjustment

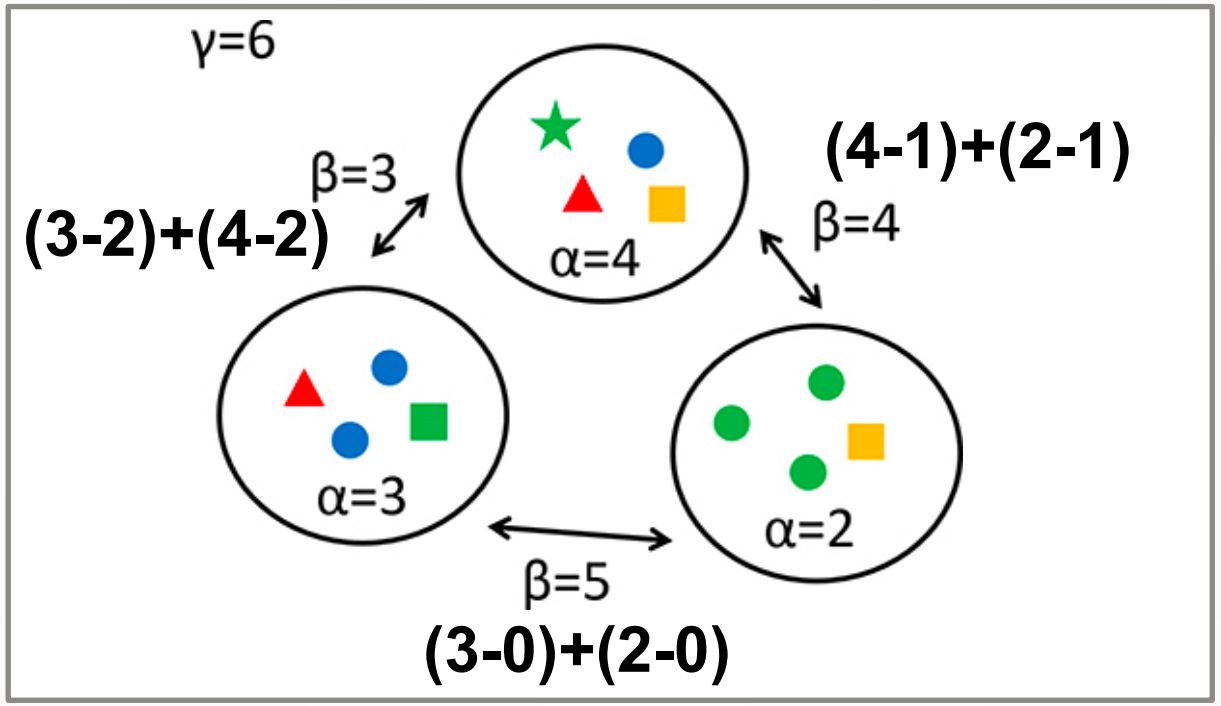
- Multiple-comparison correction has to be used when several dependent or independent statistical tests are performed simultaneously to avoid a spurious positives
  - Storey's FDR: most liberal p-value correction method that controls number of false findings among all accepted ones (i.e. 5% FDR and 200 findings implies that 10 are wrong)
  - Benjamini-Hochberg FDR: liberal p-value correction method that controls number of false findings among all accepted ones (i.e. 5% FDR and 200 findings implies that 10 are wrong)
  - Bonferroni: most conservative p-value correction method that controls odds of making at least one false findings among all accepted ones (i.e. 0.05 and 200 findings implies that there is 5% chance of getting one false finding)

Q-value >> BH-FDR >> Bonferroni

Liberal … Conservative

α/β diversity

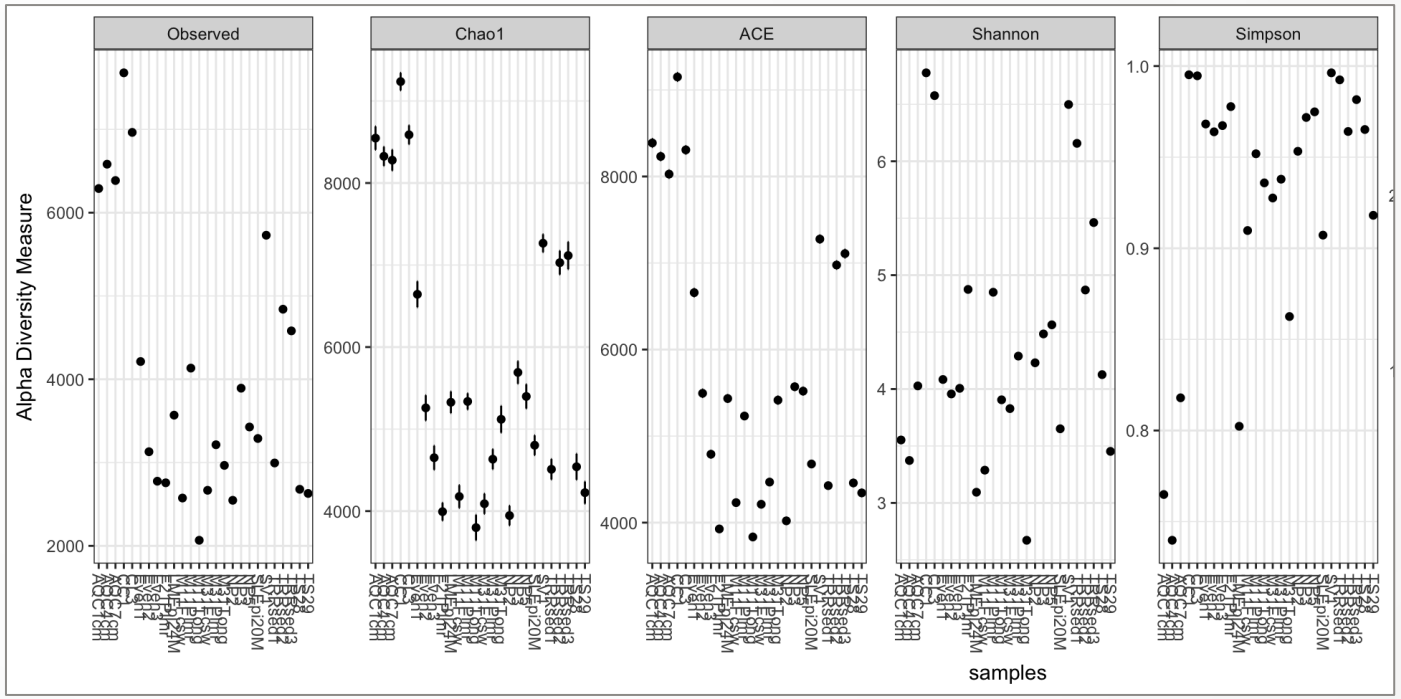
- α: diversity within a community (i.e. number of species)
- β: diversity between communities (i.e. how different are samples)
- γ: global diversity, γ = α × β



α/β diversity

- Various distance metrics generated for measuring α, β, and/or γ diversity. Common one include Unifrac (i.e. fraction of total unshared branch length), Jaccard (i.e. intersection over union), and Bray-Curtis ($1-(2C_{ij}/S_i+S_j)$).

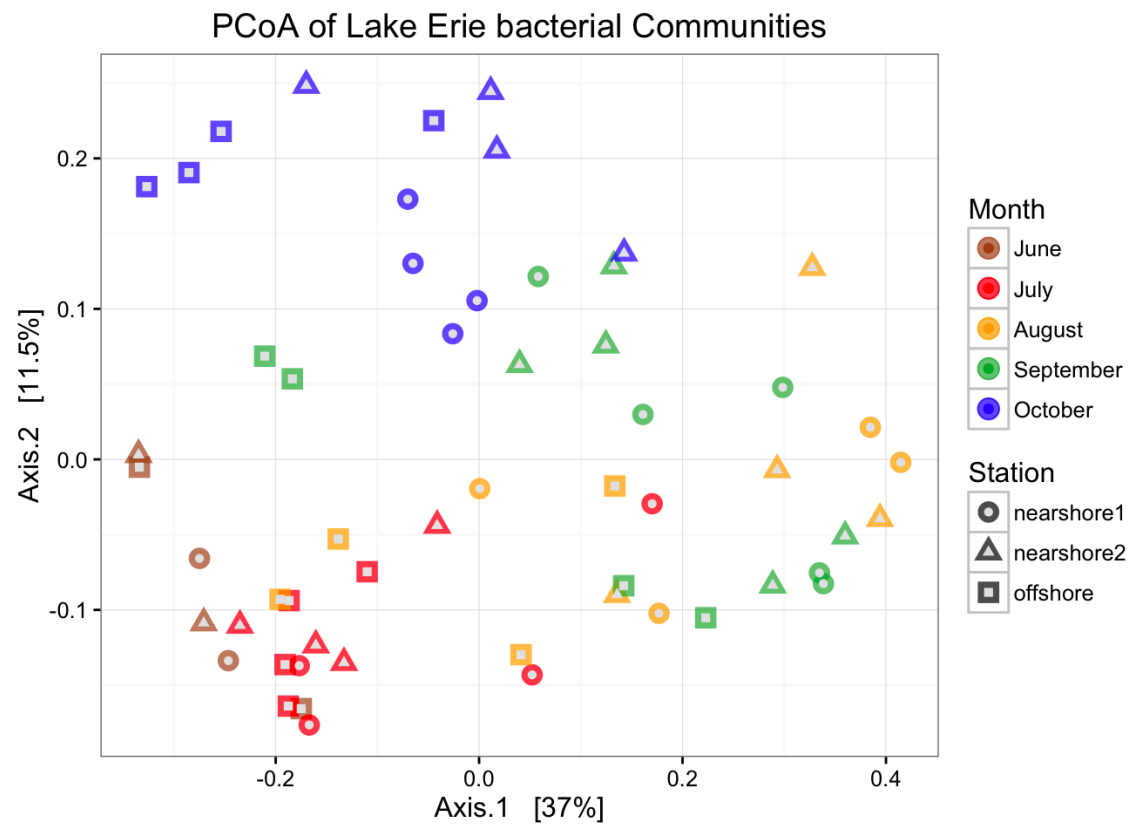
	Categorical	Phylogenetic
Presence/ Absence	Jaccard	Unifrac
Quantitative Abundance	Bray-Curtis	Weighted Unifrac

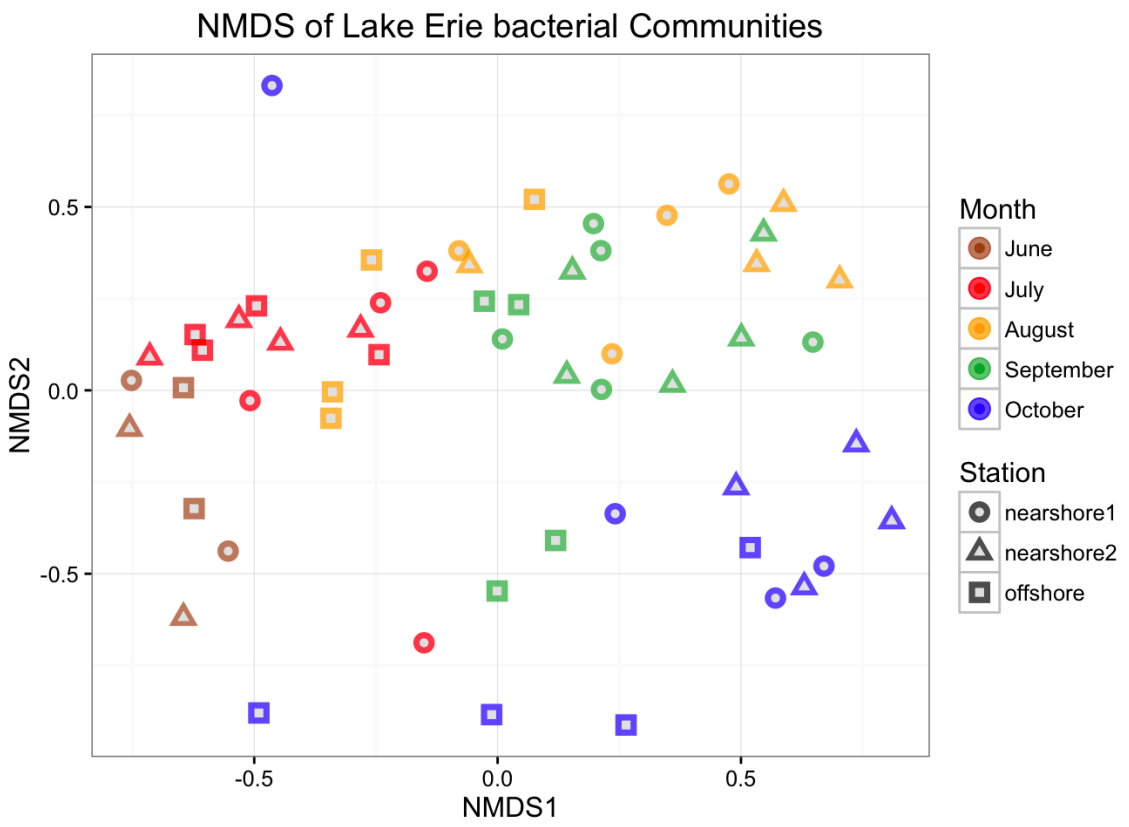


Ordination methods

- Project high-dimensional data onto lower dimensions. Most common include nMDS > PCoA > PCA
  - Principle coordinate analysis (PCoA) and non-metric multidimensional scaling (nMDS) not affected by zeros and can use different distance metrics. nMDS use ranks and non-linear mapping. nMDS should obtain stress value < 0.2
  - Principle component analysis (PCA) assumes linearity and uses Euclidean distance. Can be good if data has no zeros

Ordination methods



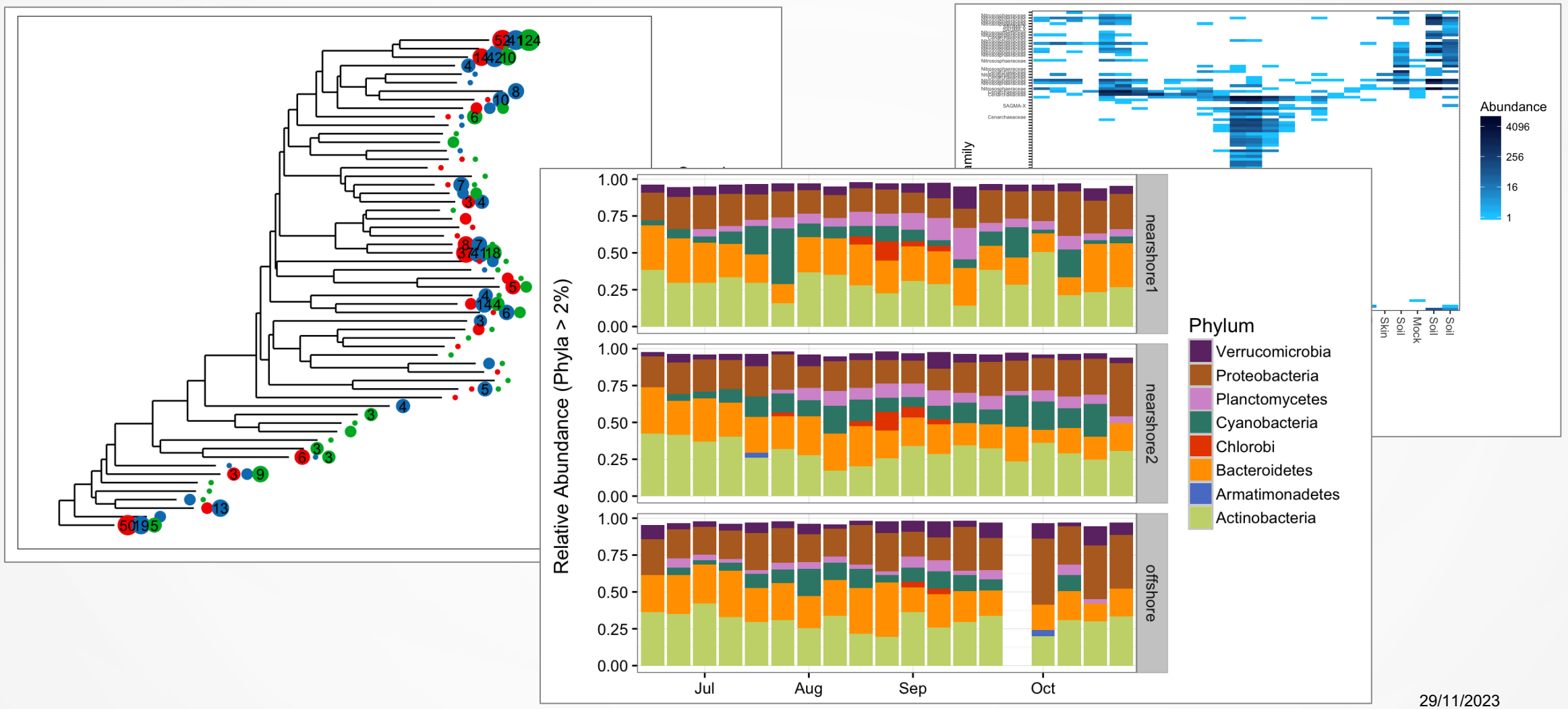


Ordination significance testing

- Permanova test followed by homogeneity of dispersion (variance) test can be used to find our whether group centroids between groups differ or not and/or to examine level of dysbiosis etc between individuals from different treatment groups

- ANOSIM can be used to test whether distances between groups are greater than within groups

Visualizations using phyloseq



Discuss with you neighbour(s)

You are about to study human gut microbiota before and after use of antimicrobial drugs using 16S amplicon sequencing. What types of secondary analyses would you do for your amplicon data

Content

- Microbiome (and host) sequencing
- Design of 16S experiments
- Primary analysis of 16S sequencing data
- Secondary analysis of 16S sequencing data
- Tool suggestions

Tool suggestions

Category	Tool / Method
Read trimming	**Trimmomatic**, Mothur, QIIME
Paired-end assmebly	FLASH, **PEAR**, PANDA-seq
Artefact removal	Mothur, QIIME
Chimera Removal	**VSEARCH**, UCHIME, **UCHIME2**, DECIPHER, ChimeraSlayer
OTU clustering	**UPARSE**, USEARCH, VSEARCH, opticlust, CD-HIT
AVS	**DADA2**, Deblur, MED, **UNOISE**
Taxonomy classification	**UPARSE**, USEARCH, **VSEARCH**, opticlust, CD-HIT
Taxonomy databases	**SILVA**, **EzBio**, GG, NCBI
Contamination removal	**Decontam**, SourceTracker
Differential abundance testing	**DESeq2**, edgeR, MWU, metagenomeSeq, ANCOM
16S statistical testing	**vegan**
Multivariate analysis	**PERMANOVA**
p-value adjustment	Q-value, **FDR**, Bonferroni
Oordination	**nMDS**, **PCoA**, PCA
Visualization	**phyloseq**