Exercises in Marine Ecological Genetics

11. Metabarcoding: microbiome analysis

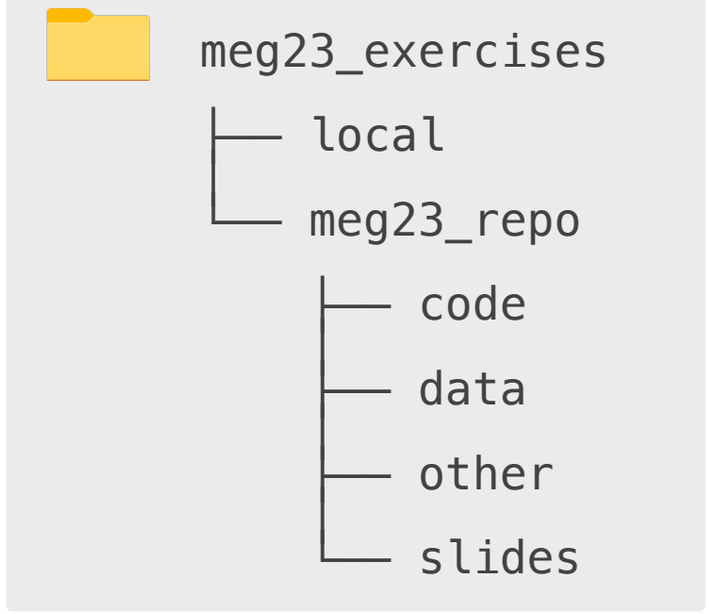
- Prepare metabarcoding raw data
- “Denoise” reads and identify ASVs (Amplicon Sequence Variants)
- Analyze the microbial community composition
- Assess alpha and beta diversity

Martin Helmkampf

Download course materials using git

Go to project directory

```
cd dir          # e.g. Documents/meg23_exercises
ls -l    # view directory contents, long format
```



Update course repository

```
cd meg23_repo
git pull
```

In case of an error message

```
cd ..                                          # go back to project directory
rm -rf meg23_repo                                    # delete old repository
git clone https://github.com/mhelmkampf/meg23_repo.git
```

Avoiding version conflict

Please do not save over files in the course repository. Instead, save your own scripts to the local subdirectory (including copies of course scripts you would like to edit), e.g with

```
cp code/11_meta.sh ../local/11_meta_lc.sh                # cp [source] [destination]
```

Example dataset

Sabrin Abdelghany

Do sea cucumbers of the species *Actinopyga crassa* in the Red Sea:

- feed selectively?
- change the microbial community of the sediment?

Sample	N
Seawater	3
Seagrass	4
Sediment	4
Foregut	5
Midgut	5
Hindgut	5

Approach: Amplicon sequencing of 16S region V3-V4 (~ 460 bp) on Illumina MiSeq

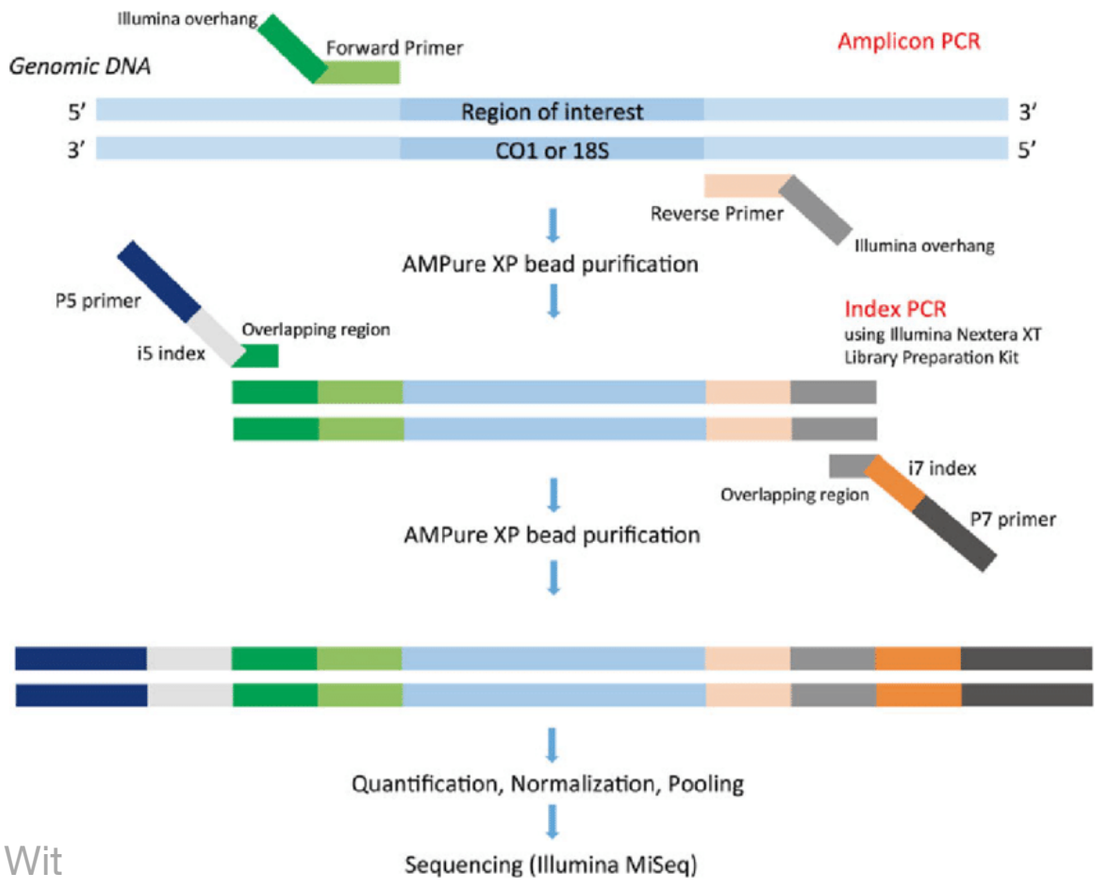
Microbiome analysis with QIIME 2

Amplicon sequencing

Next-gen sequencing of specific genomic regions (e.g. 16S)

Many samples in parallel by multiplexing (using unique molecular barcode for each sample)

Dual-PCR library preparation



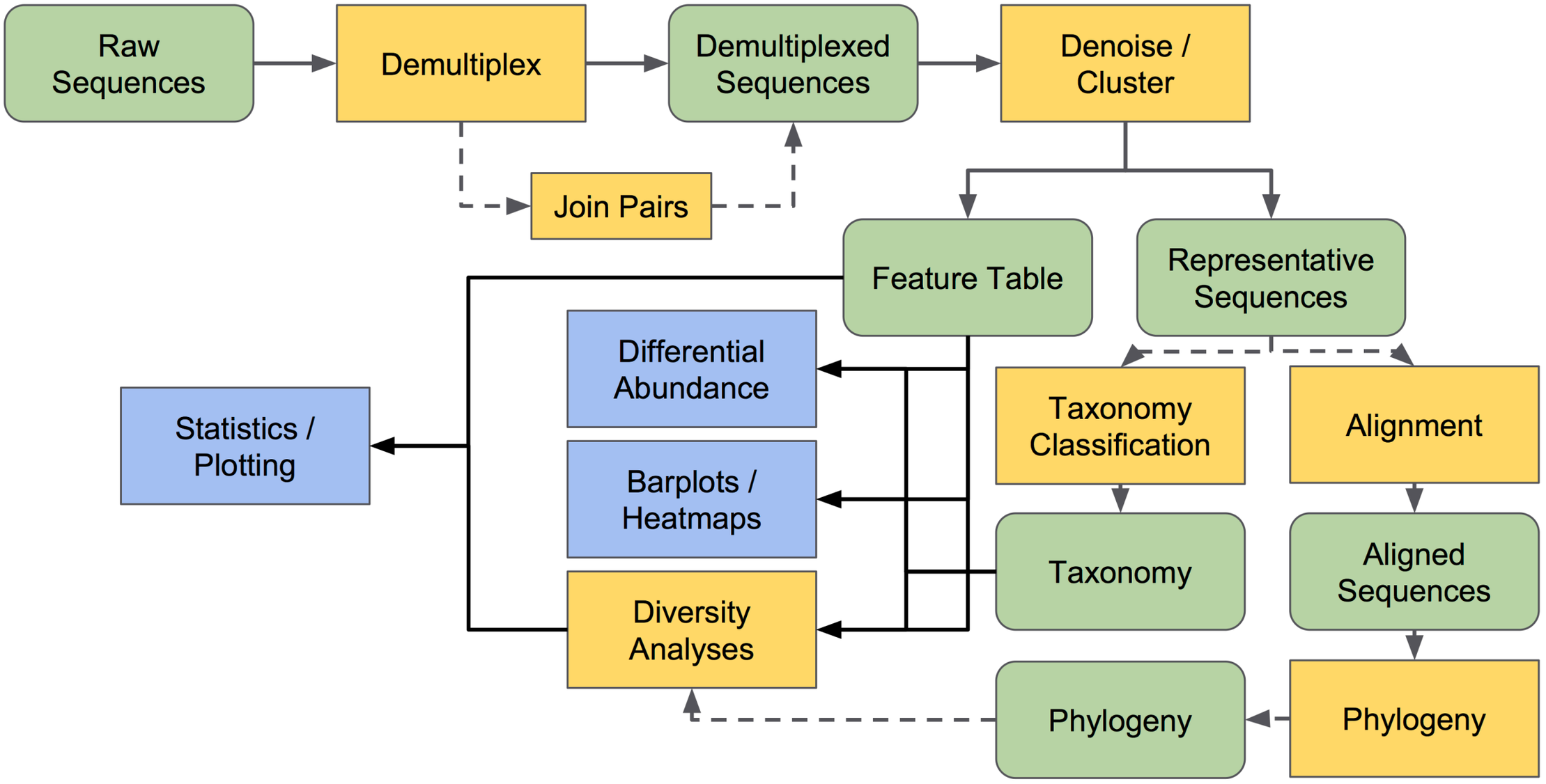
Illumina MiSeq bench top sequencer

7–25 million reads, up to 2 × 300 bp

Pierre De Wit

Illumina Inc.

QIIME 2 microbiome analysis workflow



qiime2.org

Demultiplexing

Exercise 1

- Import raw (e.g. FASTQ) and metadata — `qiime tools import`
- Demultiplex, remove primers — `qiime demux-paired` `qiime trim-paired`
- Join pairs — `qiime vsearch join-pairs`
- Quality check — `qiime demux summarize`

at each step

↓

output.qza (data) — Input for next step

output.qzv (visualization) — View at https://view.qiime2.org

Denoising

`qiime deblur denoise-16S`

Exercise 2

- Quality filter

  Filter out low-quality sequences, trim reads

- Dereplicate

  Count sequences, remove singletons and PCR artifacts

- Deblur

  Statistically infer error-free sequences, remove chimeras

Alternative:

Cluster sequences into Operational Taxonomic Units (OTUs), e.g. with 97% threshold

↓

Amplicon Sequence Variants (ASVs)

ASV sequences	ASV frequencies	Denoising stats
Representative sequences	Feature table	

Matching sequences to database with BLAST

Exercise 3

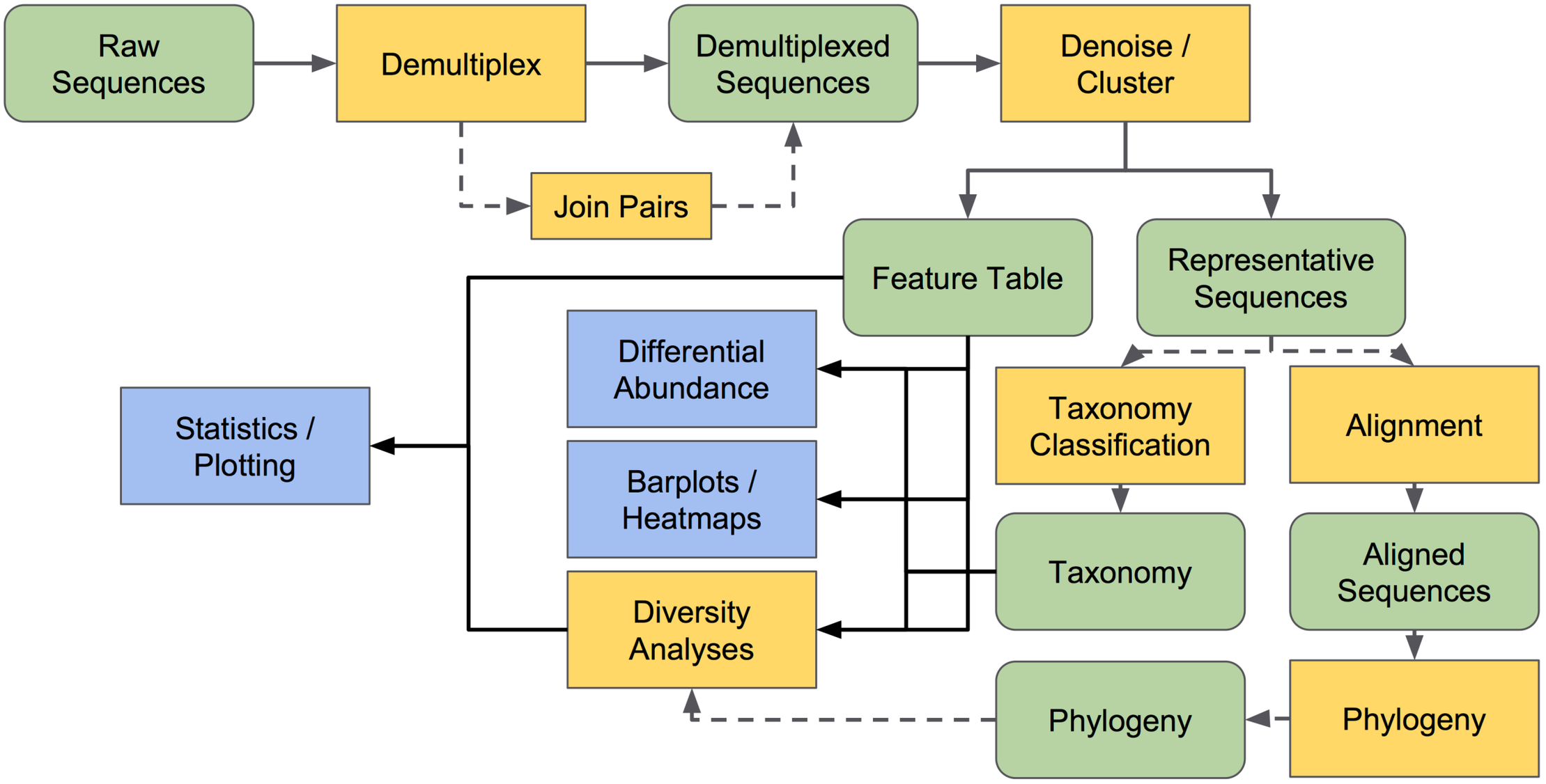
Algorithm overview

- Split query into very short segments (*k*-mers or words)
- Find exact matches between words and sequences in database (seeds)
- Extend matches to local alignments (HSP; stop once too many mismatches occur)
- Evaluate statistical significance of each HSP (e-value)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Oncorhynchus keta mitochondrial COX1 gene for cytochrome c oxidase subunit 1, partial cds, isolate: OK_M08F	Oncorhynchus keta	1029	1029	100%	0.0	99.64%	772	LC094471.1
☑	Oncorhynchus keta mitochondrial COX1 gene for cytochrome c oxidase subunit 1, partial cds, isolate: OK_M01F	Oncorhynchus keta	1029	1029	100%	0.0	99.64%	772	LC094464.1
☑	Oncorhynchus keta isolate 10_Narva cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	Oncorhynchus keta	1029	1029	100%	0.0	99.64%	655	KR778851.1

...

QIIME 2 microbiome analysis workflow



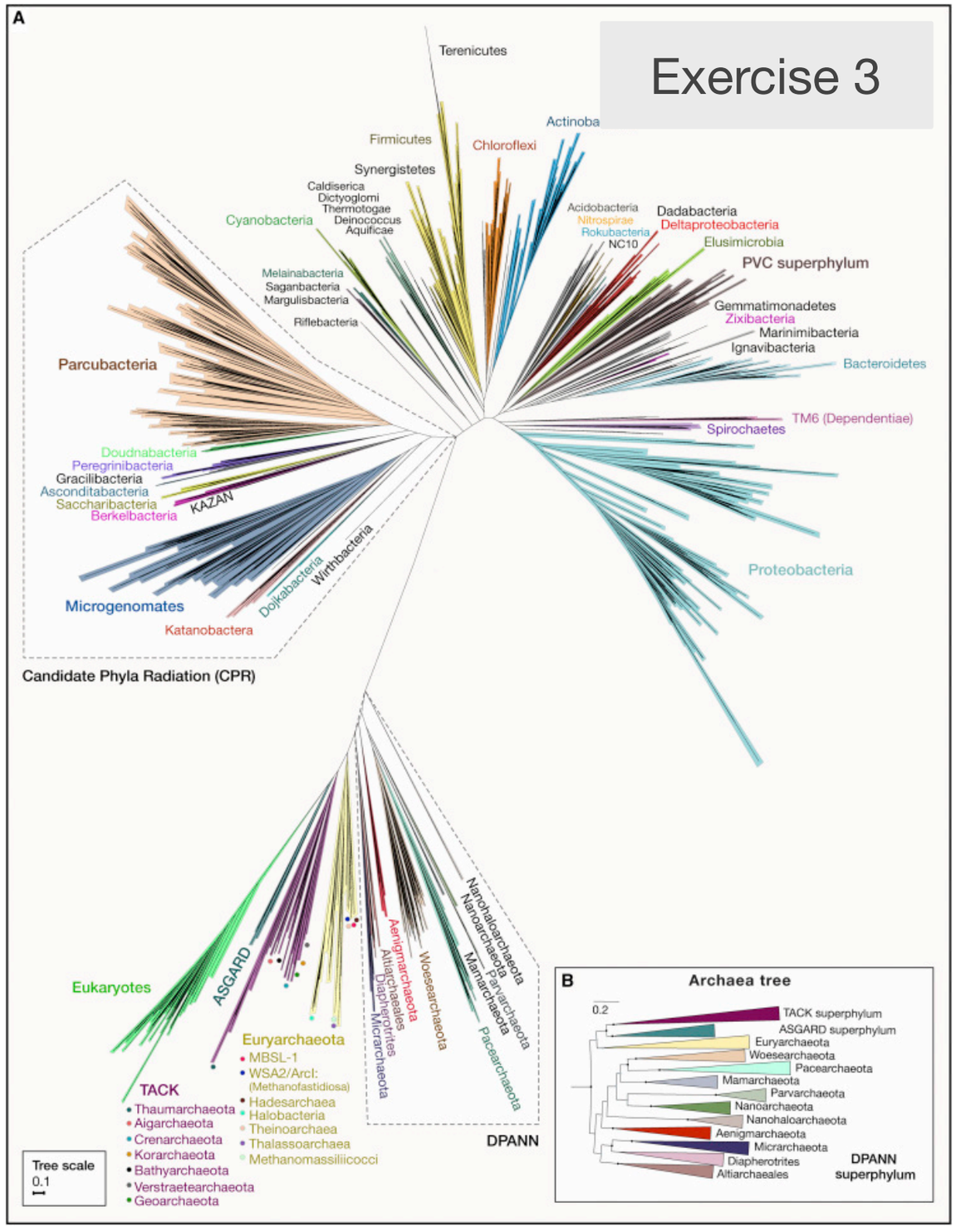
qiime2.org

Taxonomic classification

- What species are present in each sample?
- Achieved by comparing query sequences (ASVs) to reference database
- Alignment-based (e.g. BLAST) or using machine-learning

```
qiime feature-classifier
```

silva
high quality ribosomal RNA databases



Castelle & Banfield 2018, Cell

Diversity analysis

Exercise 4

How diverse is the community?

How similar / dissimilar are the communities?

`qiime diversity`

Alpha diversity

Within-community diversity, e.g. species richness, Shannon's index, — $H' = -\sum_{i=1}^{R} p_i \ln p_i$

Faith's Phylogenetic Diversity, Pielou's Evenness — $J' = \frac{H'}{H'_{\max}}$

Beta diversity

Dissimilarity between communities, e.g. Bray-Curtis index (taxonomic dissimilarity), — $BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$

Jaccard distance, UniFrac distance (phylogenetic dissimilarity) — $d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$

0 (same community)

– 1 (no similarity)

Diversity analysis continued

Exercise 4

- Diversity estimates are sensitive to sample size (sequencing depth)

  → subsampling to the same depth is required (rarefaction)

- Statistical testing of differences possible

`qiime feature-table rarefy`

- Visualize by ordination (Principal *Coordinate* Analysis)

Further analyses

- Differential abundance testing
- Phylogenetic tree building
- Time-series data

Course outline

May be subject to change

Class	Date	Topics	Script
01	Apr 14	Introduction, software installation	01_intro.R
02	Apr 21	Hardy-Weinberg equilibrium	02_hwe.R
03	Apr 28	Genetic drift and effective population size	03_drift.R
04	May 05	Population structure and gene flow	04_structure.R
05	May 12	Isolation by distance (lecture online, exercises in person)	05_ibd.R
–	May 19	Himmelfahrt break	–
06	May 26	Genome sequencing and assembly	06_genseq.sh
07	Jun 02	Genotyping, SNPs and population genomics	07_snps.sh
08	Jun 09	Recombination and linkage disequilibrium	08_recomb.R
–	Jun 16	Student presentations	–
09	Jun 23	Selection and mutation	09_sel.R
10	Jun 30	DNA barcoding	10_barcode.sh
11	Jul 07	Metabarcoding: microbiome analysis	11_meta.sh
12	Jul 14	Metabarcoding: eDNA	