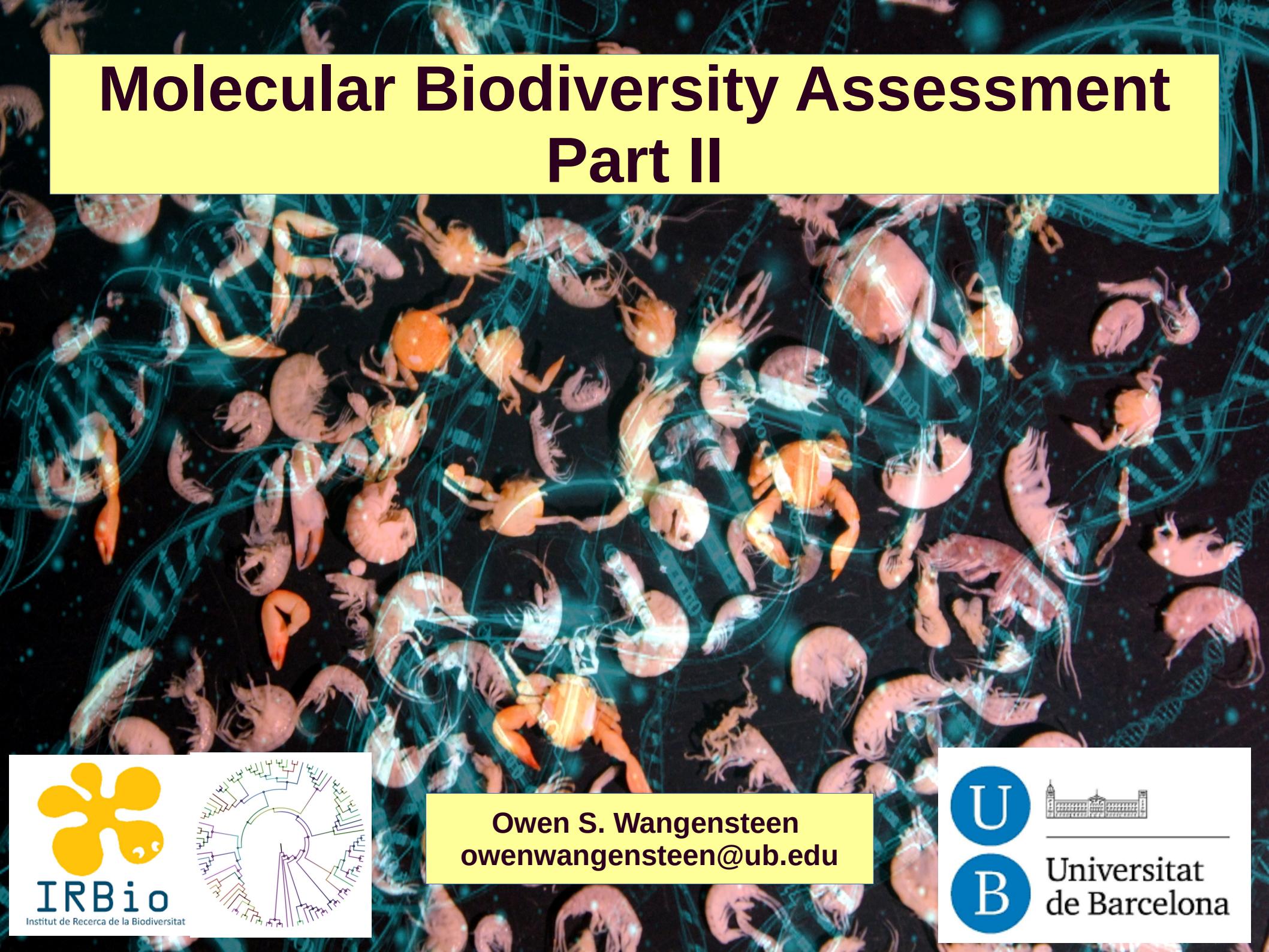
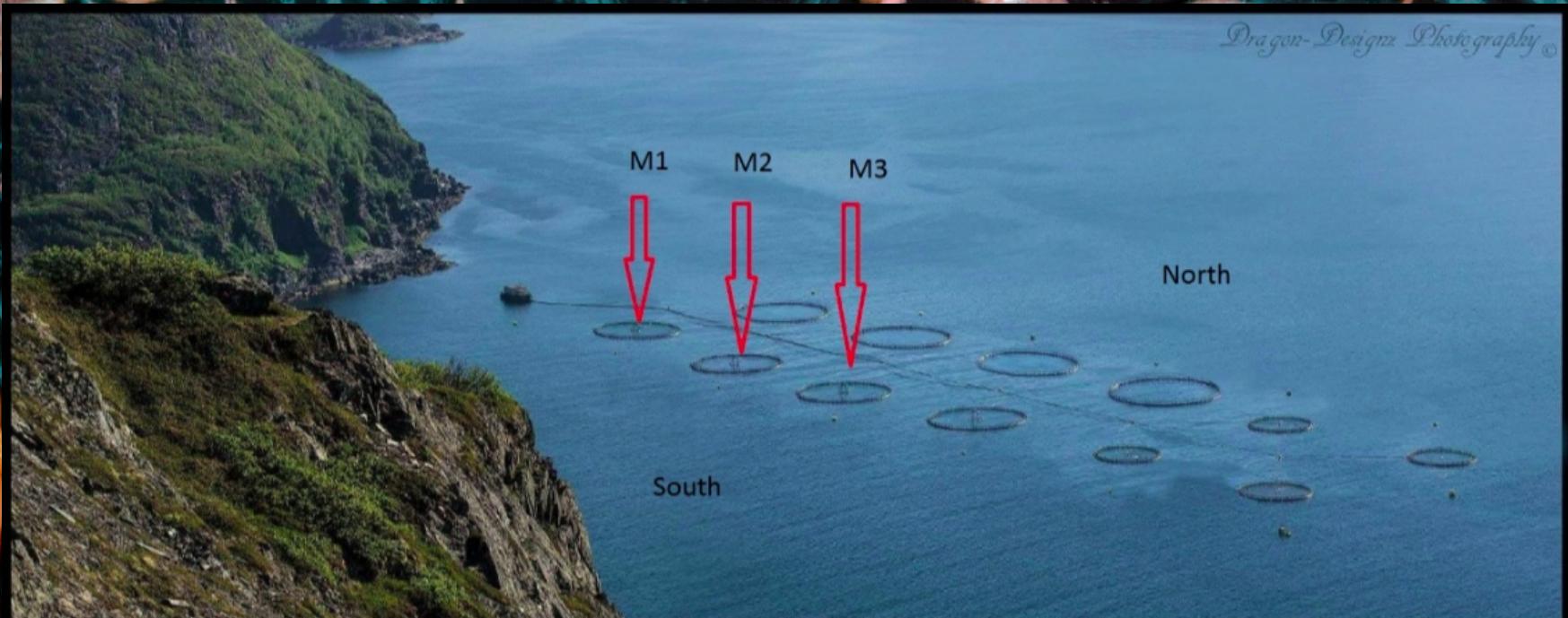


Molecular Biodiversity Assessment

Part II



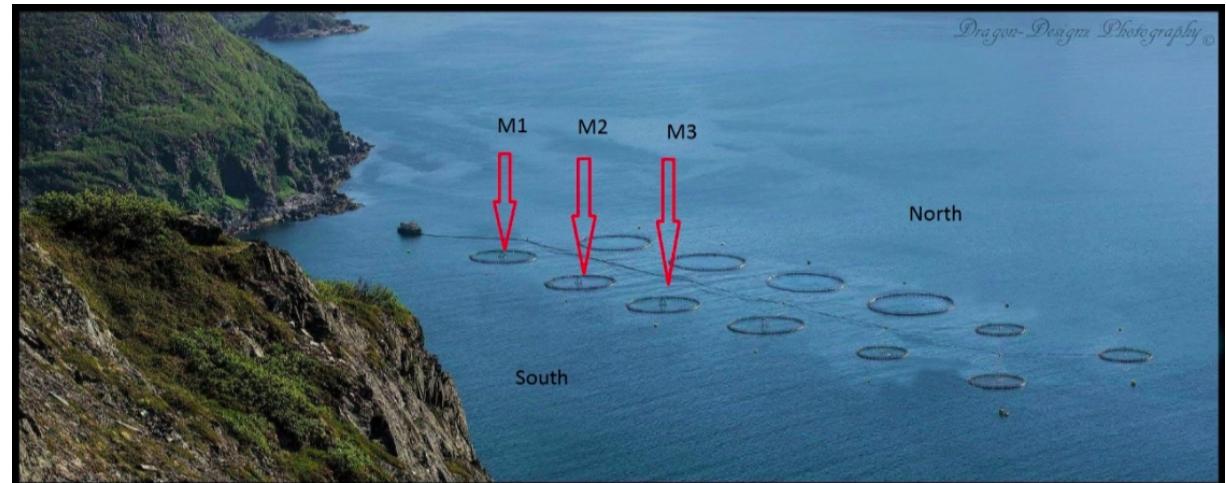


**AND EXAMPLE USING THE
MJOLNIR PIPELINE**

The ULØY Dataset

4 multiplexed libraries:

- DNA ULO1: 30 samples
- DNA ULO2: 30 samples
- DNA ULO3: 16 samples
- DNA ULO4: 33 samples



7 time points: Jul20 – Sep03 – Sep12 – Oct01 – Oct15 – Oct25 – Nov07

15 samples from each sampling day:

- 3 stations sampled inside salmon cages x 3 water replicates
- 2 stations outside the cages (North & South) x 3 water replicates

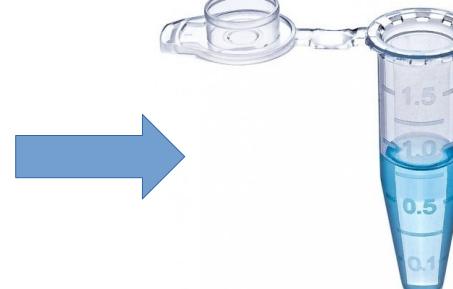
Plus 4 extraction blanks

Total: 109 samples

Turon et al. (2022) Fine-scale differences in eukaryotic communities inside and outside salmon aquaculture cages revealed by eDNA metabarcoding. *Front. Genet.* 13, 957251.

doi: [10.3389/fgene.2022.957251](https://doi.org/10.3389/fgene.2022.957251)

The ULØY Dataset



water sampled
from 2 m depth

1L filtered through
Sterivex 0.22 µm

DNA extracted by modified
QIAGEN Blood & Tissue kit

COI Leray-XT amplification (tagged primers)
1 PCR + ligation (PCR-free) library prep



Illumina MiSeq 2x250 bp sequencing

only a small portion (1/20) of total output
was selected for this workshop example

ULO1_R1.fastq
ULO1_R2.fastq

ULO2_R1.fastq
ULO2_R2.fastq

ULO3_R1.fastq
ULO3_R2.fastq

ULO4_R1.fastq
ULO4_R2.fastq

The ULØY Dataset

All the files we need to run the MJOLNIR pipeline are here:

<https://drive.google.com/drive/folders/17IX3z7DuWr3eSGePykGcgmV1XvfylQAh?usp=sharing>

1. The FASTQ paired files for all four libraries:

ULO1_R1.fastq

ULO1_R2.fastq

ULO2_R1.fastq

ULO2_R2.fastq

ULO3_R1.fastq

ULO3_R2.fastq

ULO4_R1.fastq

ULO4_R2.fastq

2. The ngsfilter.tsv tables for all four libraries:

ngsfilter_ULO1.tsv

ngsfilter_ULO2.tsv

ngsfilter_ULO3.tsv

ngsfilter_ULO4.tsv

3. The metadata file with the original sample names and the mjolnir agnomens we used in the ngsfilter tables:

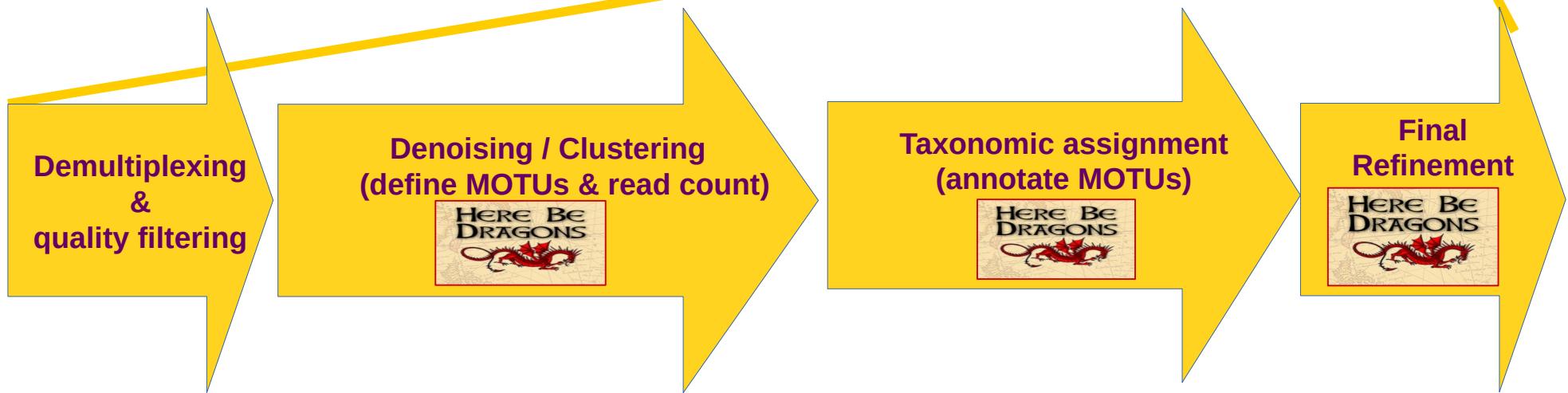
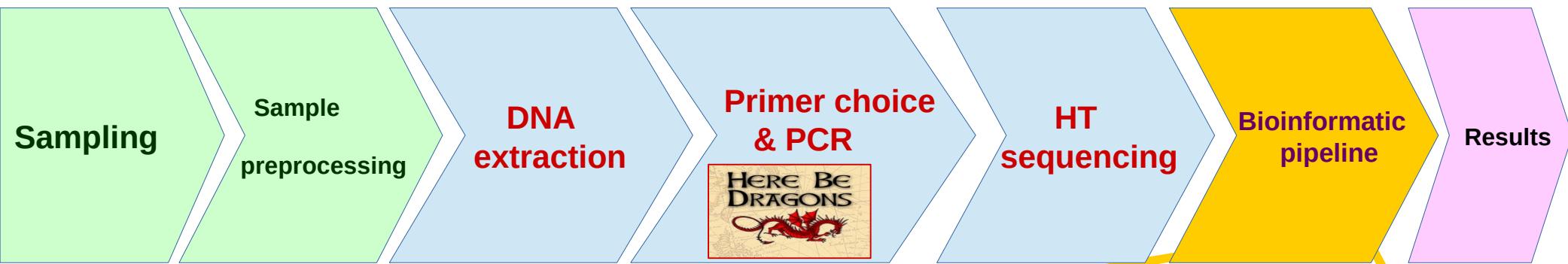
ULOY_metadata.tsv

4. The pipeline itself:
mjolnir_ULOY.R

And the results are here:

https://drive.google.com/drive/folders/1eglz3_U1pAT1UIA6MAq1hBgxO-3DjFdI?usp=sharing

METABARCODING WORKFLOW





The MJOLNIR Pipeline (2020)

<https://github.com/uit-metabarcoding/MJOLNIR>

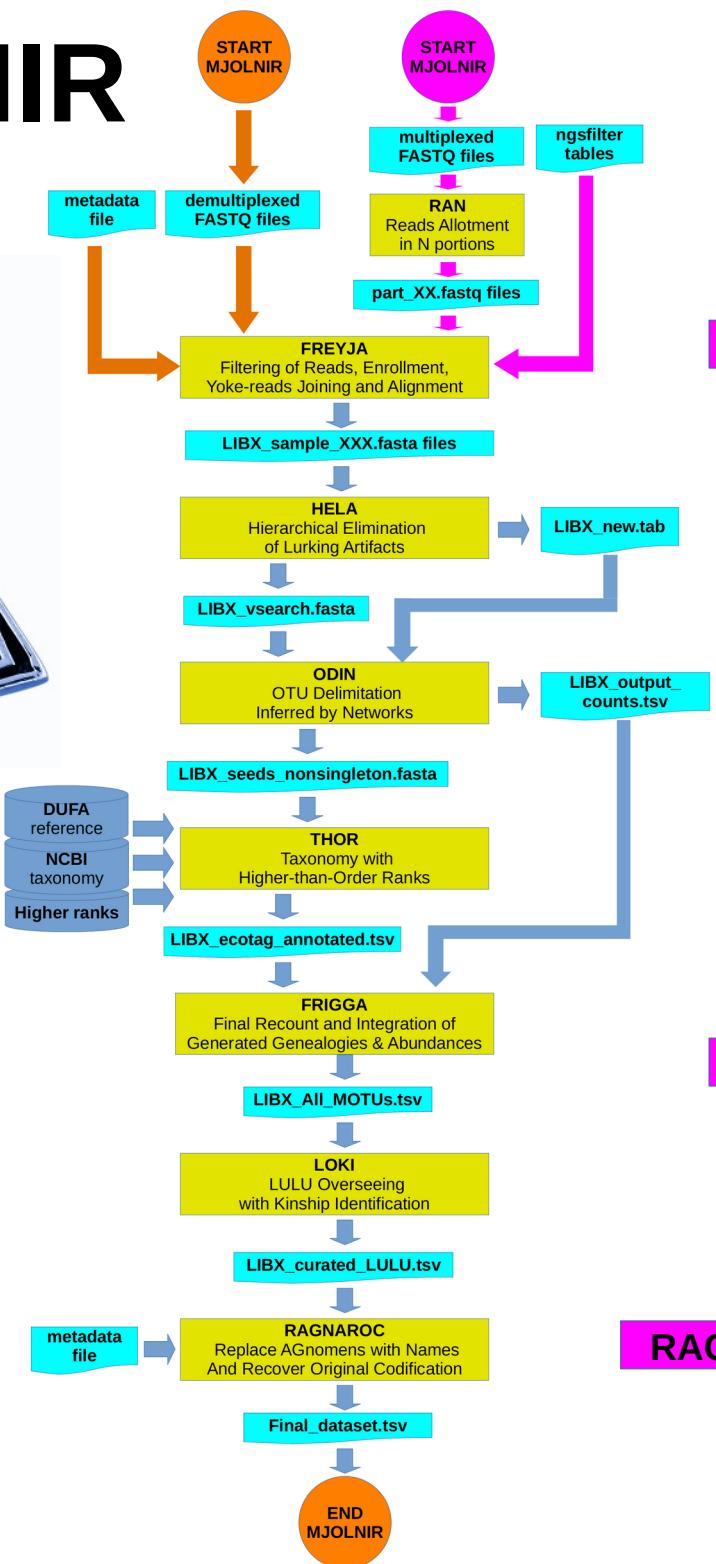
Metabarcoding Joining OBITools & Linkage Networks In R



- 1. RAN:** Reads Allotment in N portions
- 2. FREYJA:** Filtering of Reads, Enrollment, Yoke-reads Joining and Alignment
- 3. HELA:** Hierarchical Elimination of Lurking Artifacts
- 4. ODIN:** OTU Delimitation Inferred by Networks
- 5. THOR:** Taxonomy with Higher-than-Order Ranks
- 6. FRIGGA:** Final Recount and Integration of Generated Genealogies and Abundances
- 7. LOKI:** LULU Overseeing with Kinship Identification
- 8. RAGNAROC:** Replace AGnomens with Names And Recover Original Codification



MJOLNIR



OBITools / vsearch / swarm equivalents

RAN: obidistribute: prepares for parallel processing

FREYJA: illuminapairedend: aligns paired-end reads
ngsfilter: demultiplexes reads into each sample
obigrep: filters by length and retain just ACGT

HELA: obiuniq: dereplicates sequences intra sample
vsearch uchime_denovo: removes chimeras
obiuniq: dereplicates sequences across samples
abitab: generates a table of read abundances

ODIN: swarm: clusters sequences into MOTUs
owi_recount: recounts MOTU abundances

THOR: ecotag: assigns taxonomy to each MOTU seed
owi_add_taxonomy: adds higher taxonomic ranks

FRIGGA: owi_combine: combines taxonomy with abundances

LOKI: LULU: removes pseudogenes

- Recovers original sample names from metadata
- Filters by relative abundance intra sample
- Filters by minimum total abundance across samples
- Removes bacterial reads
- Removes known contaminants

A MJOLNIR pipeline is an R script:

```
library(mjolnir)

R1_filenames <-
c("ULO1_R1.fastq.gz", "ULO2_R1.fastq.gz", "ULO3_R1.fastq.gz", "ULO4_R1.fastq.gz")

lib_prefixes <- c("ULO1", "ULO2", "ULO3", "ULO4")

lib <- "ULOY"

cores <- 20

mjolnir1_RAN(R1_filenames, cores, lib_prefixes, R1_motif="_R1.", R2_motif="_R2.")

mjolnir2_FREYJA(lib_prefixes, cores, Lmin=299, Lmax=320)

mjolnir3_HELA(lib, cores)

mjolnir4_ODIN(lib, cores, d=13)

mjolnir5_THOR(lib, cores, taxo_dir="~/taxo", ref_db="DUFA_COLR_20210723.fasta",
taxo_db="taxo_NCBI_20210720")

mjolnir6_FRIGGA(lib)

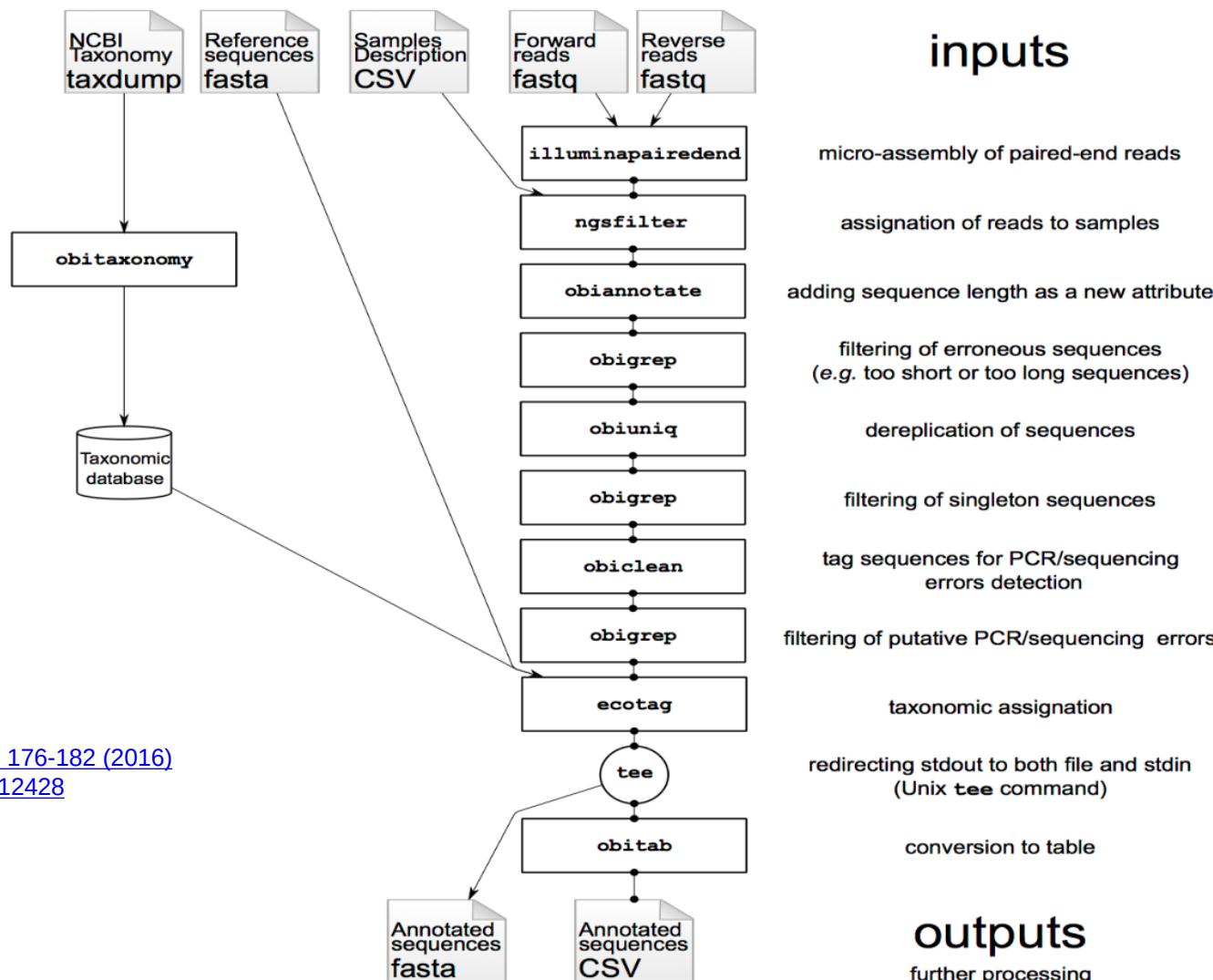
mjolnir7_LOKI(lib, min_id=.84)

mjolnir8_RAGNAROC(lib, "ULOY_metadata.tsv")
```

OBITOOLS: a UNIX-inspired software package for DNA metabarcoding

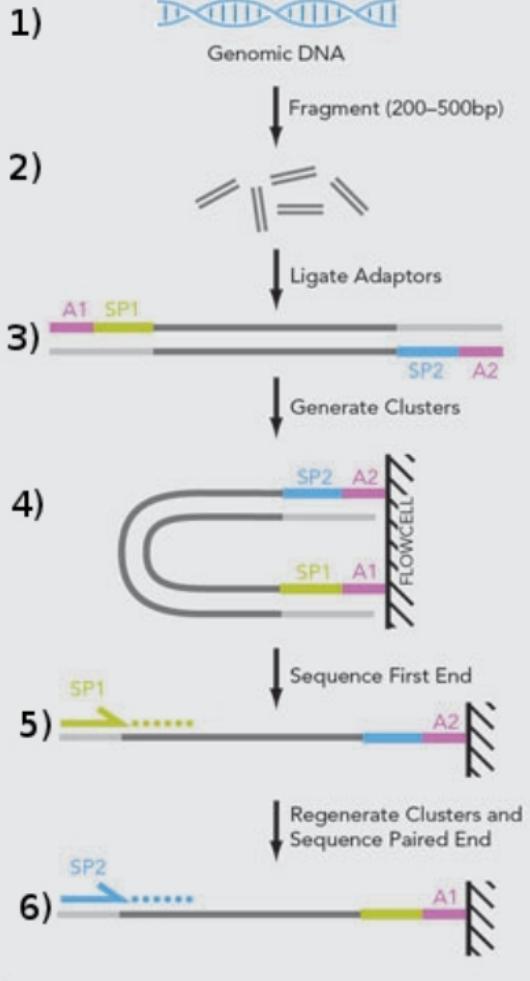
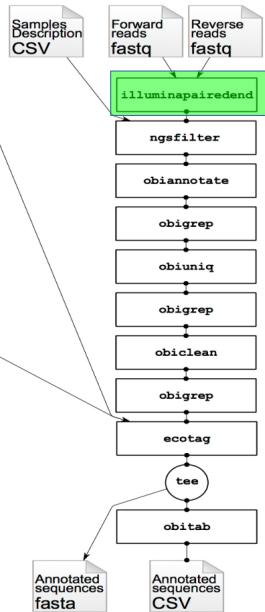
Frédéric Boyer, Céline Mercier, Aurélie Bonin, Yvan Le Bras, Pierre Taberlet, Eric Coissac✉

OBITools BASIC WORKFLOW



Molecular Ecology Resources, 16, 176-182 (2016)
<https://doi.org/10.1111/1755-0998.12428>

illuminapairedend (FREYJA)



>Read_1_Forward
TAAT**CTGACCTT****TAGACCTCGGGGTGGCCGAAAAATCA**AAAGAGATGTTGGTAAAGGACGGGATGCC
GCCTCCTACAGGATGGAAGAATGGAGTATTAAAGTCCGGTCAGTGAGAAAGTATAGTAATTGCCGG
CTAAGACTGGGAGTGAGAGAGAAGTAGGAGCACTGTTGAACAATAAGGGATCATACAAATAGAGGCACA
CGCTCTCCTGTTATTCCCTGCTACGTGTGGTTCGTGTGTTACTA

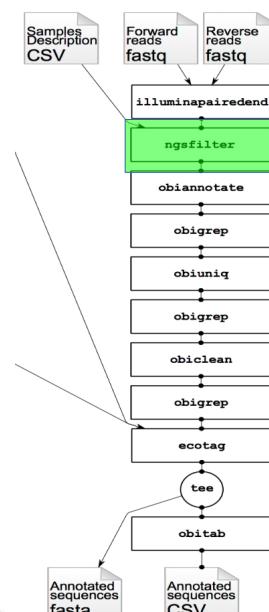
>Read_1_Reverse
TCCTGACCTT**GGTACTAGATGGACTGTGTATCCTCC**TCTCGCAGGAAATTCAATTCCACAGCGGCCCG
CAGTAGATATAGCCATCTTCTCCCTCCATCTTGCCGGGCCAGATCCATTCTAGGCTCTATCAATTTC
TTCGTAACAGCACGAACCACTCGTAAGCAGGGATTAACATGAGATCGTGTCCCCTTTTATTCTC
CCTTATTTCACCTTCTCCTCTCCCTCCCCGTCTTAT

>Read_1_Reverse_complement
ATAAGACGGGGAGGGAGAGAAGAAGGAGAAAGGTGGAAAATAAGGGAGAATAAAAGAGGGAAACA
CGATCTCATGTTAACCCCTGCTTACGAGTGGTTCTGCTGTACGAAGAAATTGATAGAGCCTAGAAT
GGATCTGGCCCGGCAAGATGGAGGGAGAAGATGGCTATATCTACTGCGGGGCCGTGTGGAATGAAT
TTCCTGCGAGA**GGAGGATACACAGTCCATCTAGTACCAAGGTCAAGGTCAAGGA**

>Read_1_aligned
taat**CTGACCTT****TAGACCTCGGGGTGGCCGAAAAATCA**aaagagatgttggtaaaggacgggatgcc
gcctcctacaggatggaagaatggagtattaaagtccggtcagtgagaagtataattgcgcgg
ctaagactggagtgagagaagtaggagcactgttgaacaataagggatcatacaaatagaggcaca
cgctctcctgttattccctgcttacgagtggttcgtgttactaagaattgatagagcctagaat
ggatctggccccggcaagatggagggagaagatggctatatctactgccccccgtgtggaatgaat
ttcctgcaga**GGAGGATACACAGTCCATCTAGTACCAAGGTCAAGGTCAAGGA**

OBITOOLS:

ngsfilter (demultiplexing) (FREYJA)



nntagccacttagactcggggatggccaaagaatcaaaaaagatgtgataaagaactgggtctccacccctgctgggtcaaaaatgtgtatcgattcgatcagttaataatatacgtaattgcacctgctaaaacgggtaatgaggtaaaagtaaaattgtgtataaaaaactgctatacaaataatggtatacggtctatagttattncaacaggtcgcatgttaataactgttgtataaaaattactgctcctaaaatagaggatacccctgctagatgaagagaaaaattcctatatcaactgatgctcctgcgtgagcaattctggctgataaaaggggggtaaacagttcatccatccagtggctannnn

ngsfilter table:

lib	sample	tags	forward_primer	reverse_primer
UL01	UL01_sample_006	ggatgatc:ggatgatc	GGWACWRGTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
UL01	UL01_sample_007	ggattcga:ggattcga	GGWACWRGTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
UL01	UL01_sample_008	tagcaagg:tagcaagg	GGWACWRGTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
UL01	UL01_sample_011	tgaggaca:tgaggaca	GGWACWRGTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
UL01	UL01_sample_012	cgtacac:cgtacac	GGWACWRGTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
UL01	UL01_sample_013	tagccact:tagccact	GGWACWRGTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA

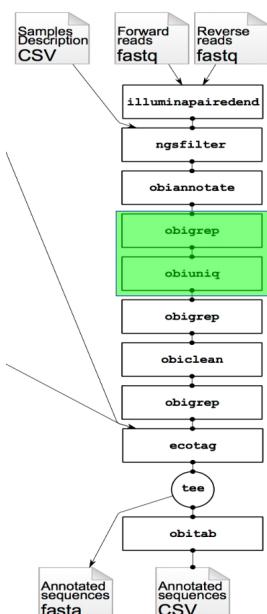
obigrep (FREYJA) & obiuniq (HELA)

Length selection

```
obigrep -p 'seq_length>303' -p 'seq_length<323' -s '^ACGT+$'
```

`-p 'seq_length>303' -p 'seq_length<323'` : keep only sequences between 303 and 323 bp long

`-s '^ACGT+$'` : keeps only sequences composed of A, C, G, and T.
Removes sequences with Ns or other wobbly bases



Dereplication

```
obiuniq -m sample
```

`-m`: keeps information of abundance in each sample

```
>M03292:11:000000000-AEF91:1:1101:20836:1098_SUB_CONS_SUB_SUB_CMP experiment=UL01; count=16103;
merged_sample={'UL01_sample_013': 2309, 'UL01_sample_027': 3224, 'UL01_sample_001': 2316,
'UL01_sample_005': 3375, 'UL01_sample_012': 307, 'UL01_sample_020': 1995, 'Empty_UL01': 3};
aaaaagatgttataaagaactgggtctccacctcctgctgggtcaaaaatgatgtattcagattcgatcagttaataatatagttaattgcacctgctaa
aacgggtaatgaggtaaaagtaaaattgctgtataaaaaactgctcatacaataatggtatacggtgtatagttattncaacaggtcgcattgttaataac
tggttataaaatttactgctcctaaaatagaggataccctgctagatgaagagaaaaattcctatatcaactgatgctcctgcgtgagcaattctggc
tgataaa
```

VSEARCH: uchime_denovo (HELA)

Chimera removal

UCHIME algorithm

A	81	CCTTGGTAGGCCGtTGCCCTGCCAACTAGCTAATCAGACGCgggtCCATCtcaCACCAccggAgtTTTtcTCaCTgTacc	160								
Q	81	CCTTGGTAGGCCGCTGCCCTGCCAACTAGCTAATCAGACGCATCCCCATCCATCACCGATAAAATCTTAATCTCTTCAG	160								
B	81	TCTTGGTgGGCCGtTaCCCCcGCCAACaAGCTAATCAGACGCATCCCCATCCATCACCGATAAAATCTTAAaCTCTTCAG	160								
Diff	A	A	p A	A	A	BBBB	BBB	BBBBB BB	BBa B	B BBB	
Votes	+	+	0	+	+	+	+++	+++	+++++ ++	++! + + + +	
Model	AAAAAAAAAAAAAAAAAAAAAxxxxxxxxxxxxxBB										

Region from a chimeric alignment generated by UCHIME.

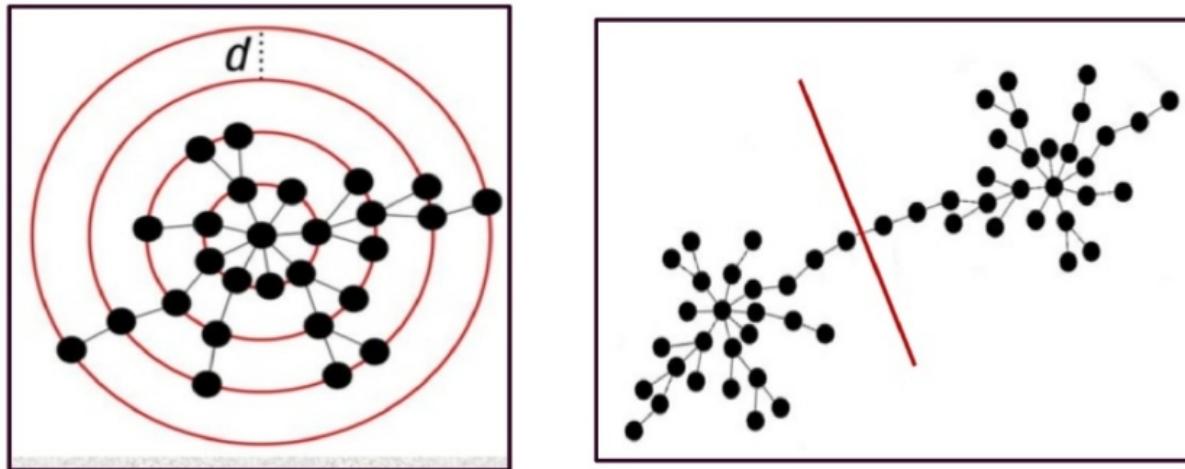
Diffs and votes are annotated. The 'Model' row indicates the three segments of the alignment which are closer to A, the crossover (X), and closer to B, respectively. Diffs are 'A'=diff with Q closer to A in the A segment, 'a'=diff with Q closer to A in the B segment, and similarly for 'B' and 'b'. A 'p' diff indicates that the parents agree but are different from Q. Votes are '+' (yes), '!' (no) and '0' (abstain), indicating whether the corresponding diff supports or contradicts the model.

uchime_ref : parent sequences are provided by the user. "Knowing" which sequences should appear.

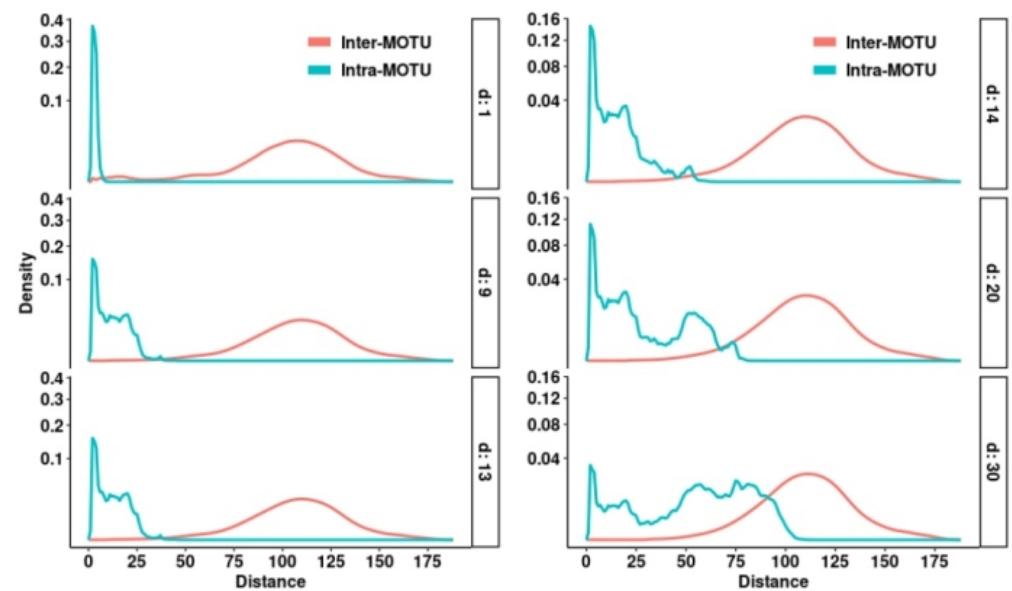
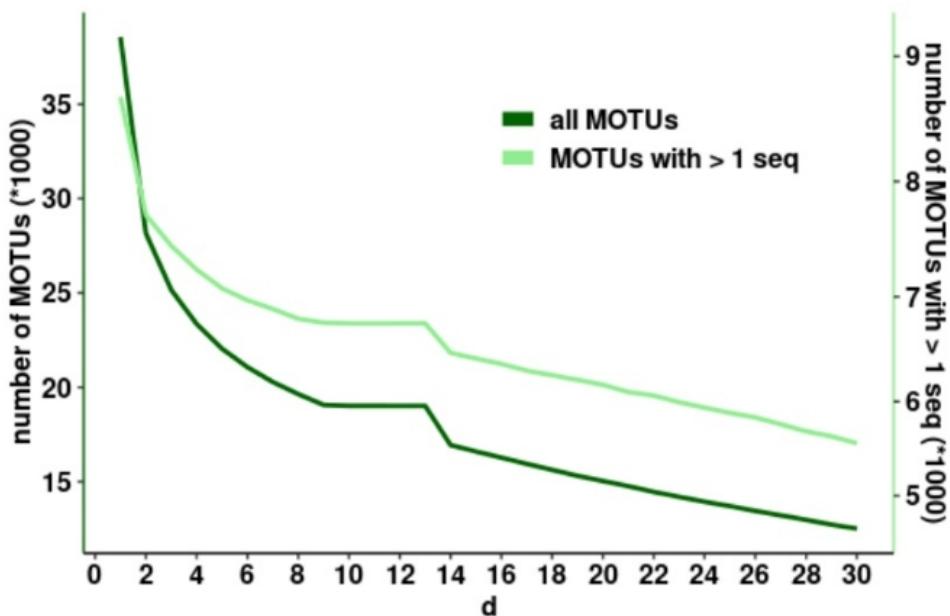
uchime_denovo : parent sequences are calculated in base of the abundance, as they are assumed to be more abundant than the chimeras.

FLEXIBLE CUTOFF CLUSTERING: SWARM (ODIN)

SWARM: Step-by-step Aggregation Algorithm (Mahé et al. 2014)

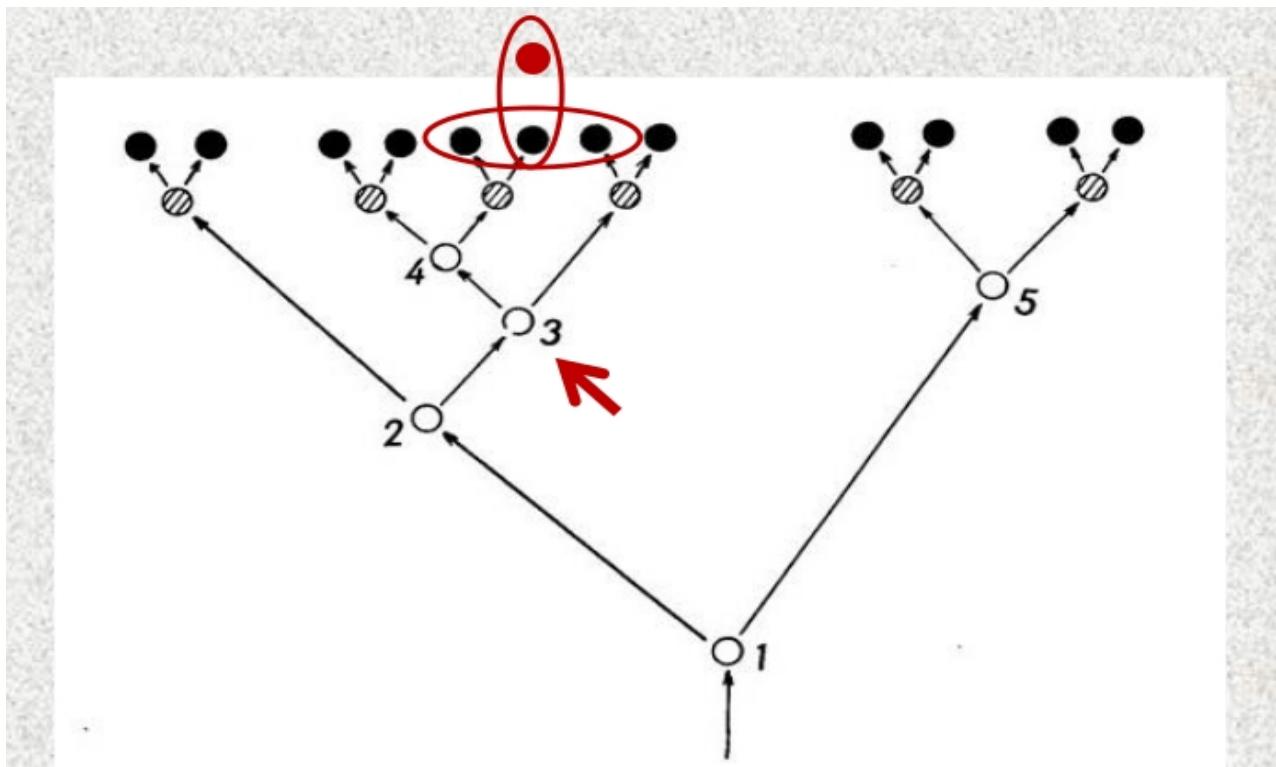
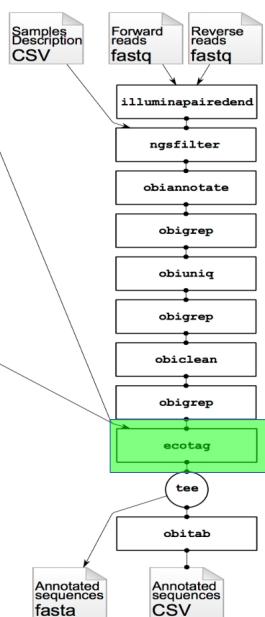


- Deterministic
- Fast!
- Behaves well for COI at $d=9 - 13$



ecotag (taxonomic assignment) (THOR)

- THE "LAST COMMON ANCESTOR" (LCA) ALGORITHM
- - The best match to the **query sequence** in the reference database is found (with a given ID %)
- - Other sequences in the reference database that are closer or equally close to the best match than the query sequence are also considered.
- - The query sequence is assigned to the most inclusive taxon that includes all these reference sequences.



FINAL REFINEMENT: PSEUDOGENE REMOVAL

LULU algorithm (Frøslev et al. 2017) (LOKI)

<https://github.com/tobiasgf/lulu>

Removal (or clustering) of MOTUs that are putative pseudogenes and NUMTs, considering:

- Patterns of co-occurrence and abundance
- Sequence similarity

Needs:

- Table with abundances in every sample
- List of pairwise matches (identity %) of MOTU sequences

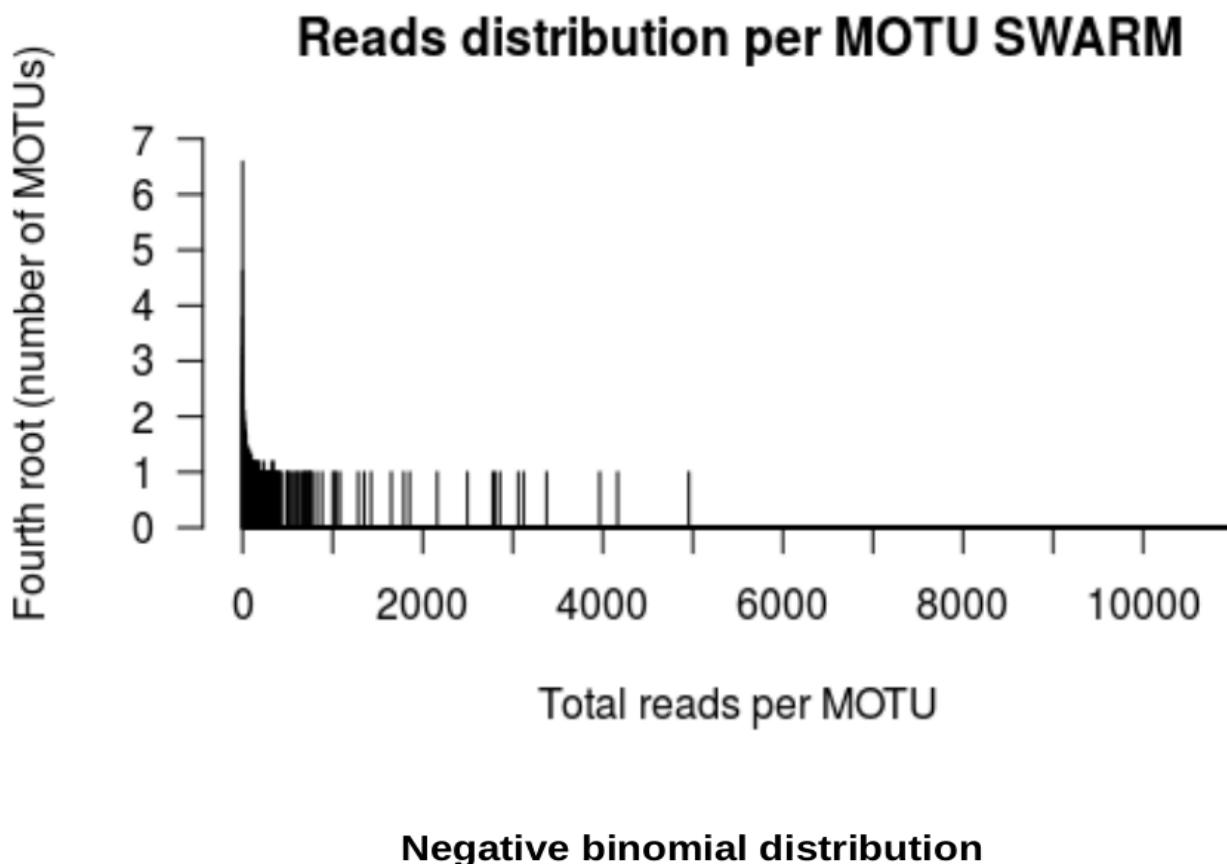
Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications* 8(1), 1188. list

FINAL REFINEMENT: ABUNDANCE FILTERING

- MINIMAL ABSOLUTE READ ABUNDANCE PER MOTU
- MINIMAL RELATIVE READ ABUNDANCE PER SAMPLE

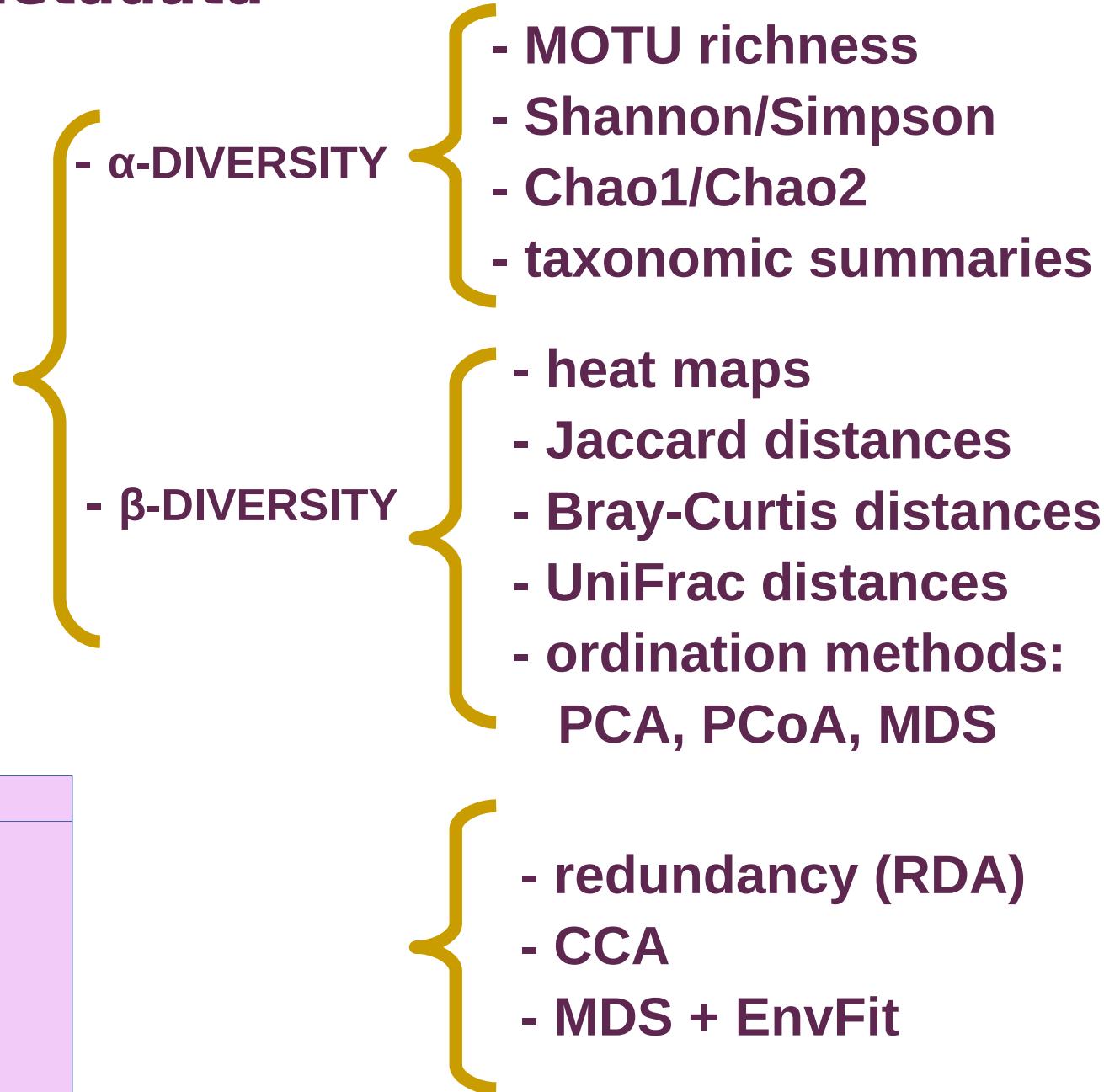
(RAGNAROC)

MINIMAL ABUNDANCE FILTERING

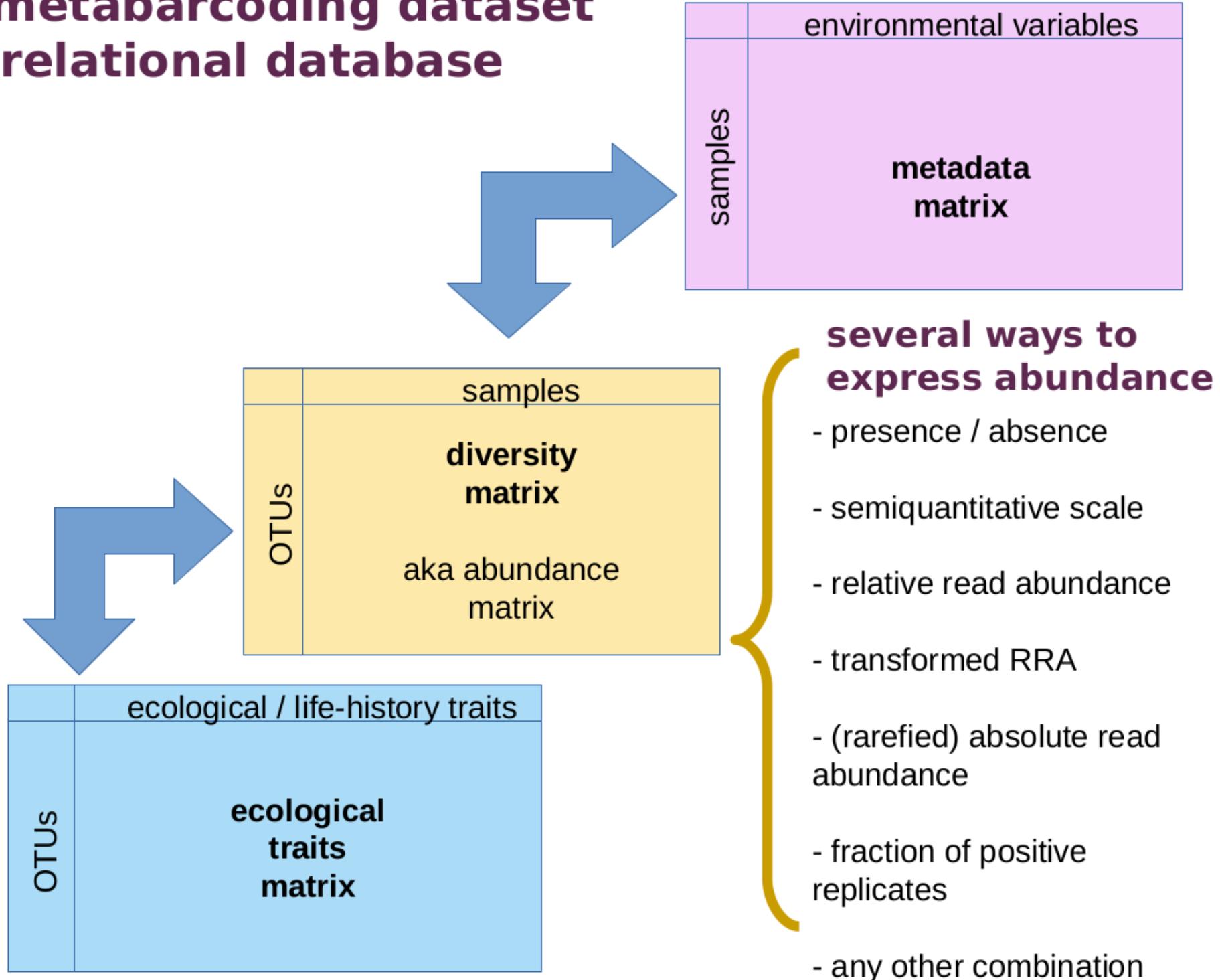


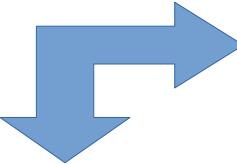
Ecological analyses based on the biodiversity matrix and on ecological metadata

	samples
	diversity matrix (aka abundance matrix)
OTUs	



The metabarcoding dataset as a relational database



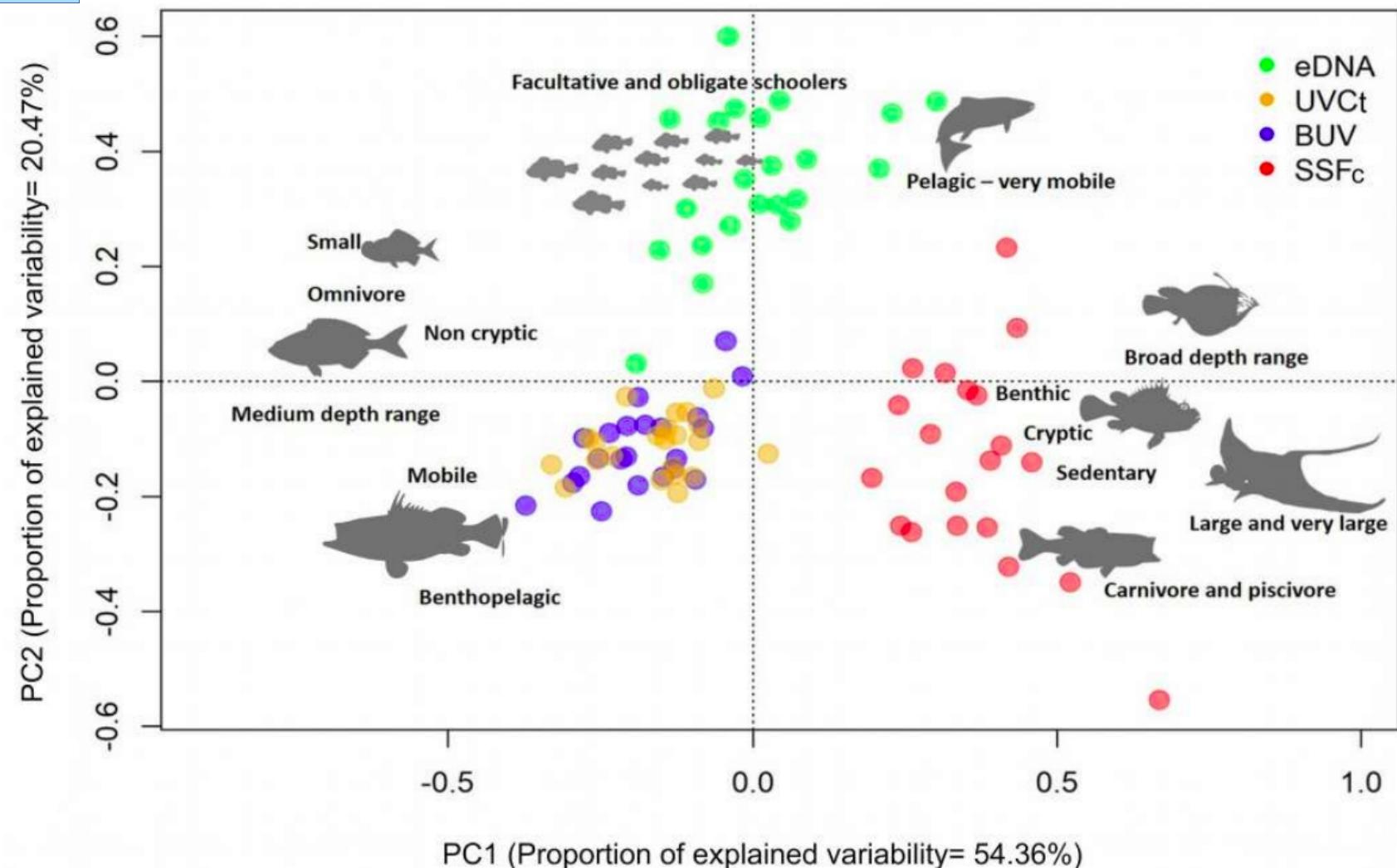


	OTUS	ecological traits
samples	OTUS	ecological traits matrix

	samples
OTUS	diversity matrix

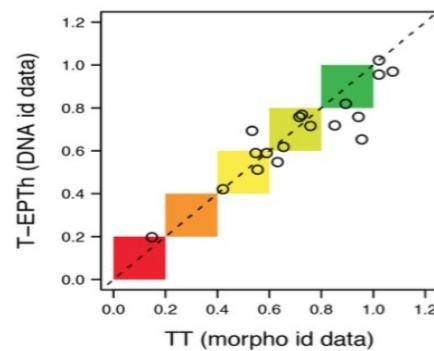
Environmental DNA effectively captures functional diversity of coastal fish communities

Giorgio Aglieri^{1,2} | Charles Baillie³ | Stefano Mariani⁴ | Carlo Cattano^{1,2,5} | Antonio Calò^{1,6} | Gabriele Turco^{1,2} | Davide Spatafora^{1,2} | Antonio Di Franco^{5,6} | Manfredi Di Lorenzo⁷ | Paolo Guidetti^{2,6,8} | Marco Milazzo^{1,2}

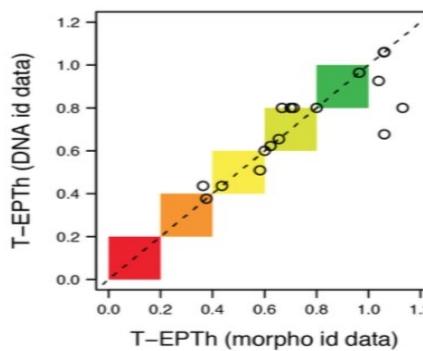


ECOLOGICAL INDICES CALCULATIONS

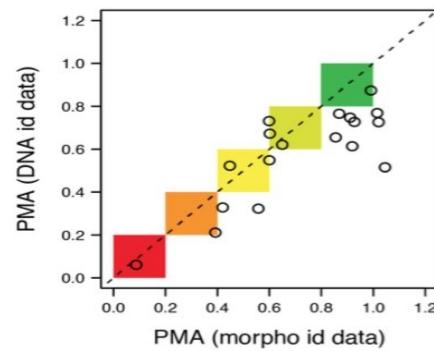
A) TT (p/a data)



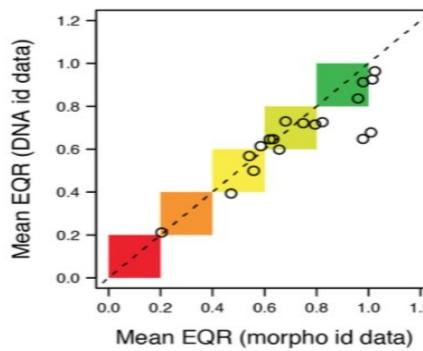
B) T-EPTh (p/a data)



C) PMA (rel. abund. data)

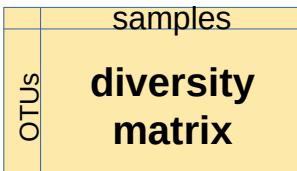


D) Mean EQR (mean of A, B & C)



Assessment: Bad Poor Moderate Good High

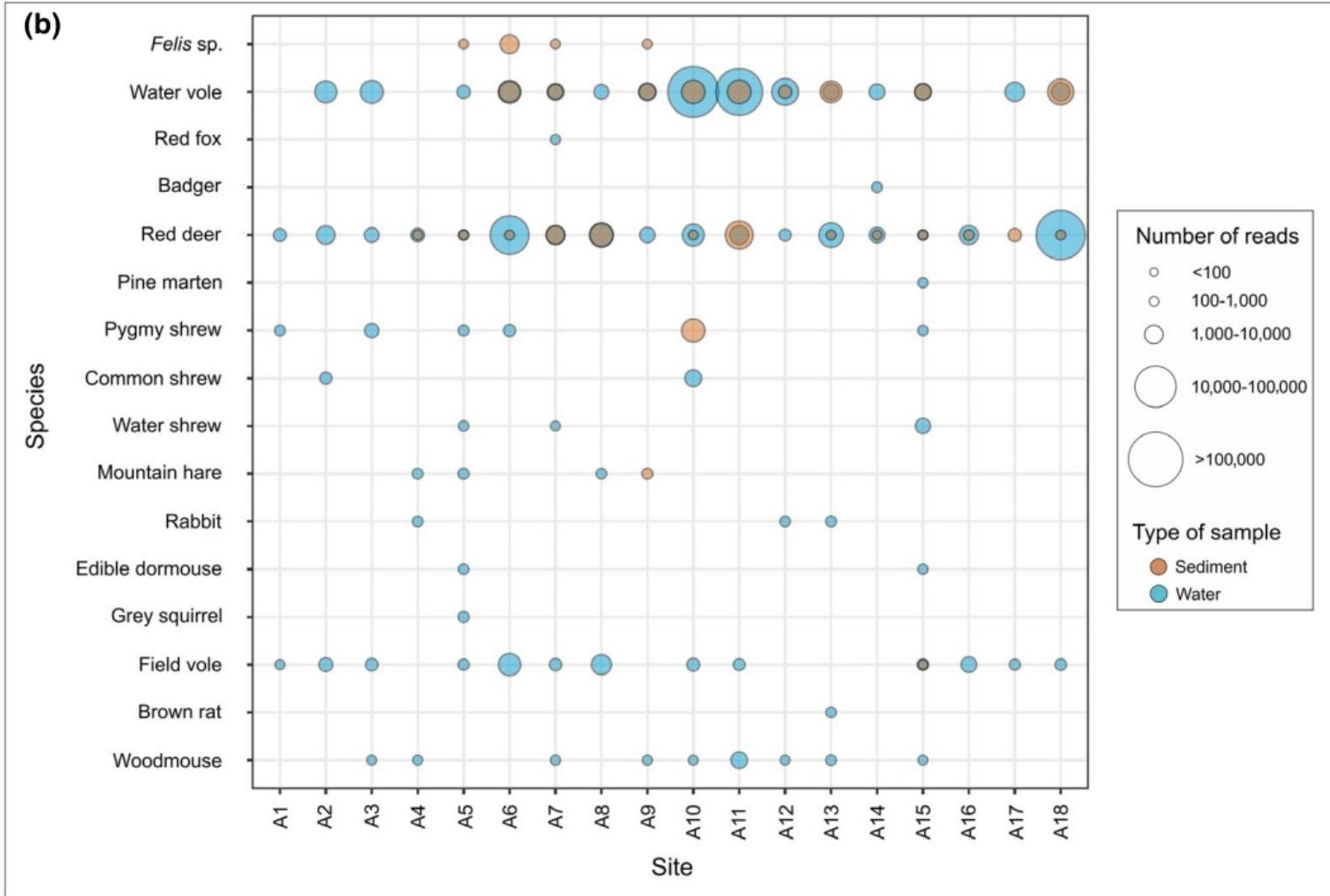
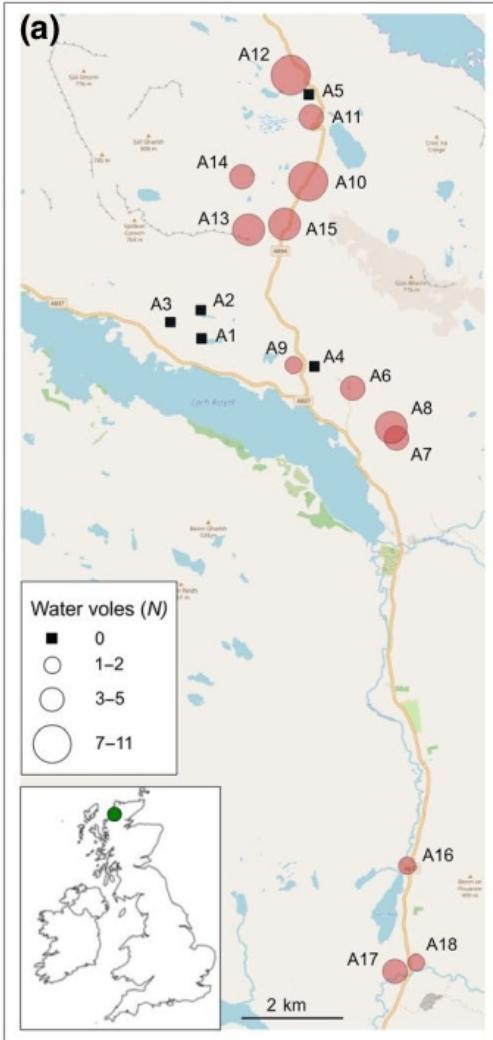
Figure 3: Comparison of Finnish macroinvertebrate WFD assessment indices calculated with taxa lists based on morphological- and DNA-based (BF2+BR2 primer) identification. The three indices are shown as normalized Ecological Quality Ratios (EQR) ranging from 0 (Bad status) to 1 (High status with no anthropogenic alteration). For all four indices, there was a significant correlation between morphological- and DNA-based assessments (Pearson correlation, $p > 0.0001$). A) Occurrence of river Type-Specific Taxa (TT, based on p/a data). B) Occurrence of river Type-Specific EPT-families (EPTh, based on p/a data). C) Percent model affinity (PMA, based on relative abundance data). D) Mean EQR of the three indices.



Elbrecht et al. (2017) <https://doi.org/10.1111/2041-210X.12789>

OTU	samples
	diversity matrix

PRESENCE/ABSENCE/READ ABUNDANCE BUBBLE PLOTS



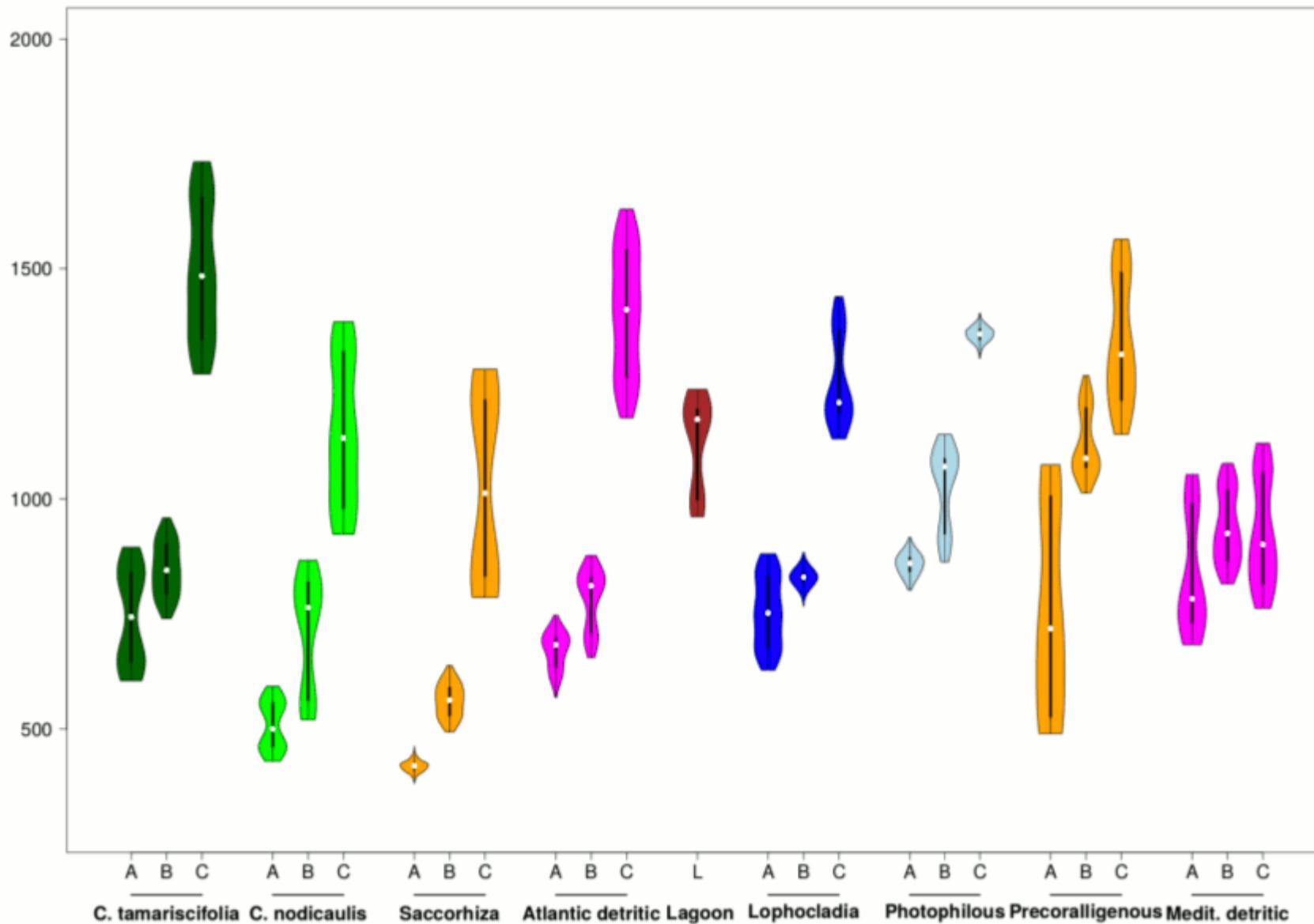
samples

OTUS

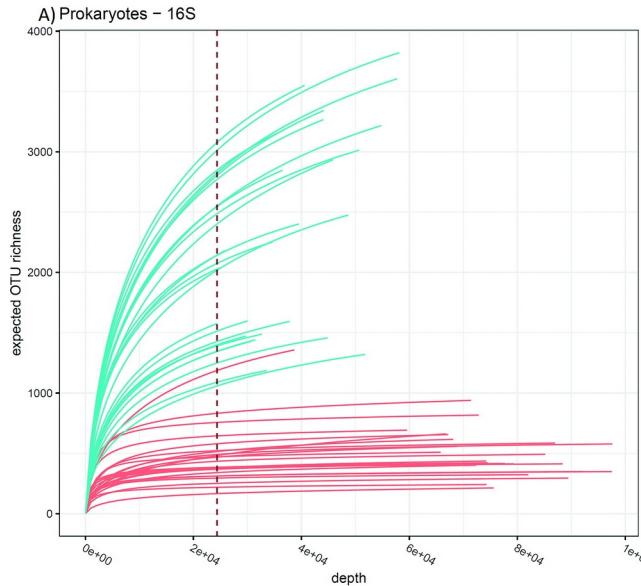
diversity
matrix

α -DIVERSITY: MOTU RICHNESS

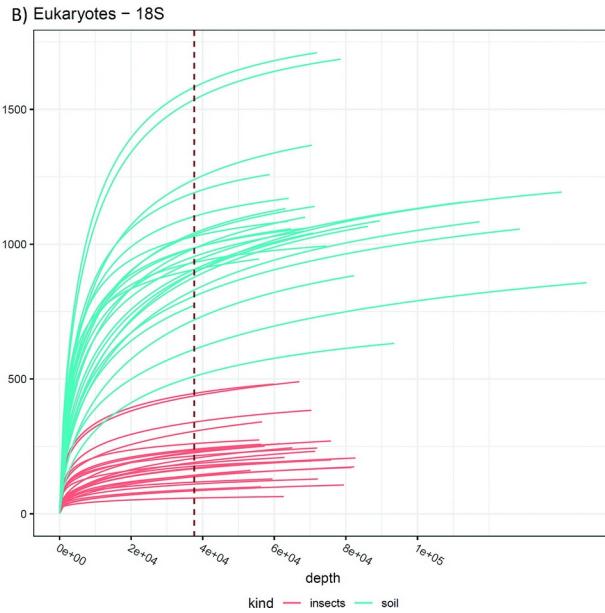
MOTU richness (rarefied to 11,000 reads)



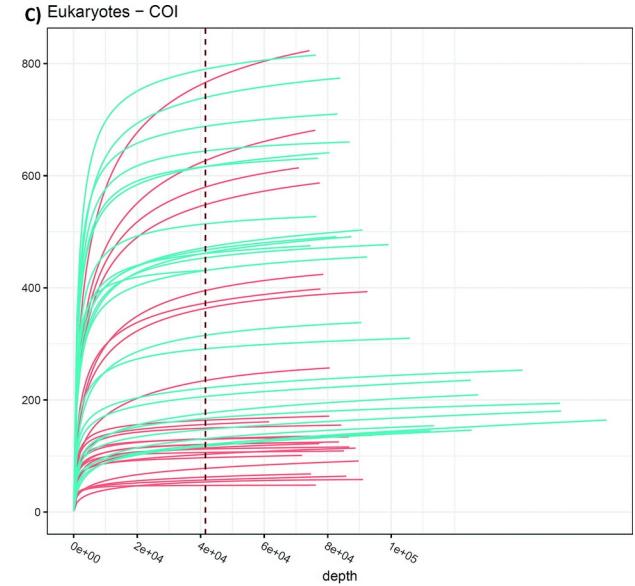
α -DIVERSITY: RAREFACTION CURVES



NUMBER OF READS

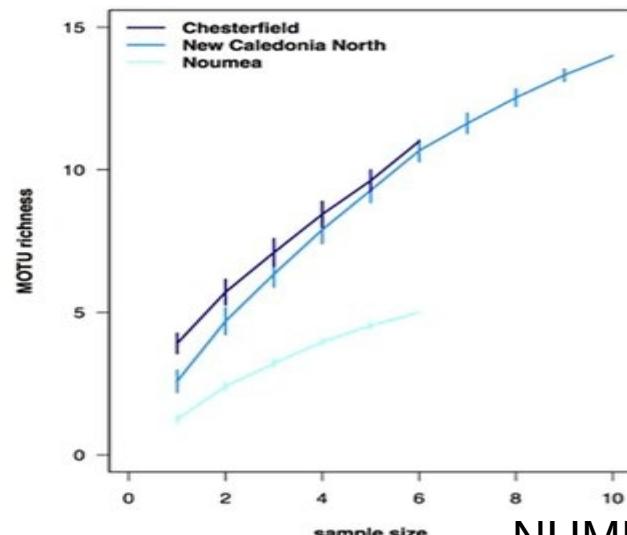
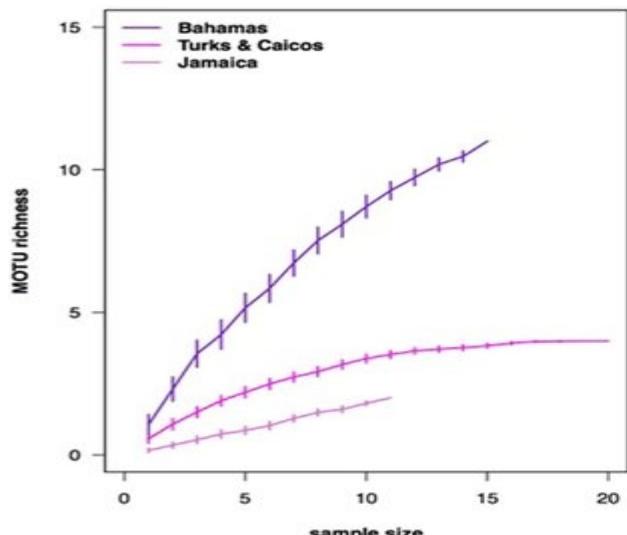


NUMBER OF READS

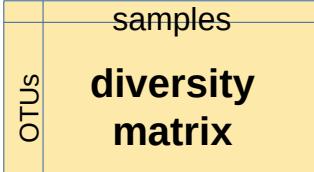


NUMBER OF READS

MOTU ACCUMULATION CURVES

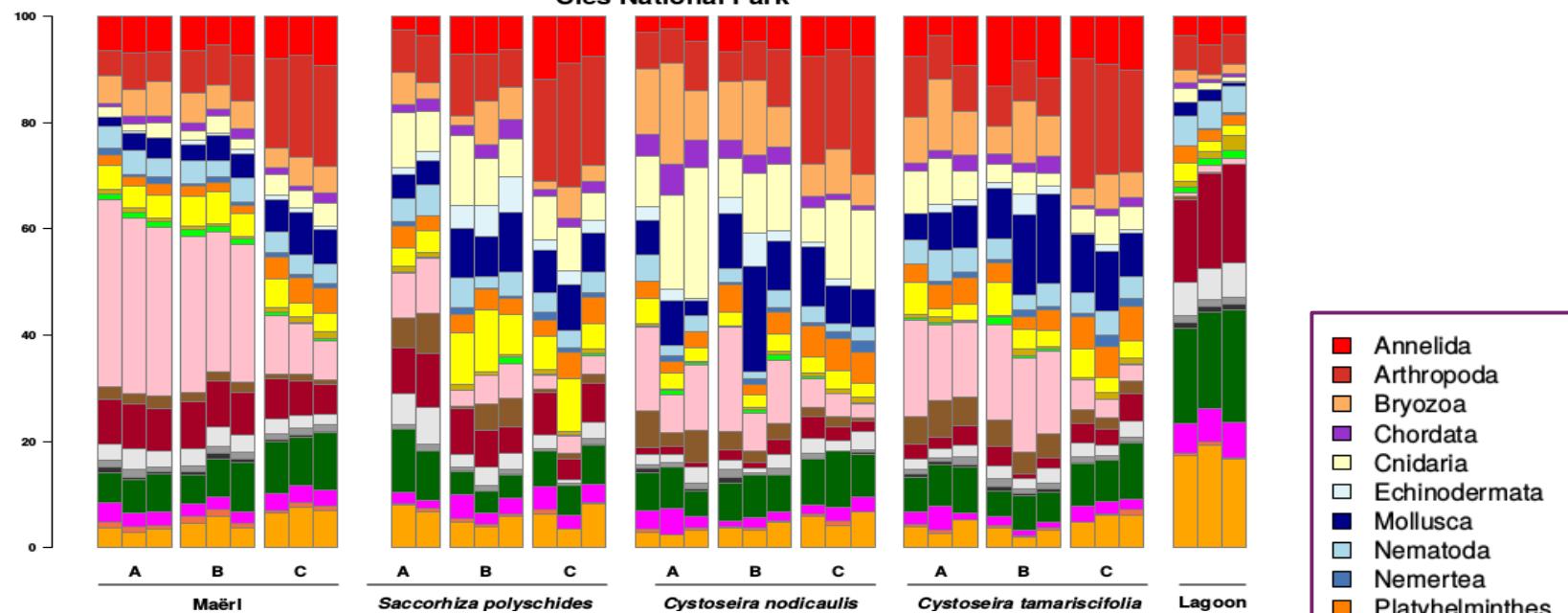


NUMBER OF SAMPLES

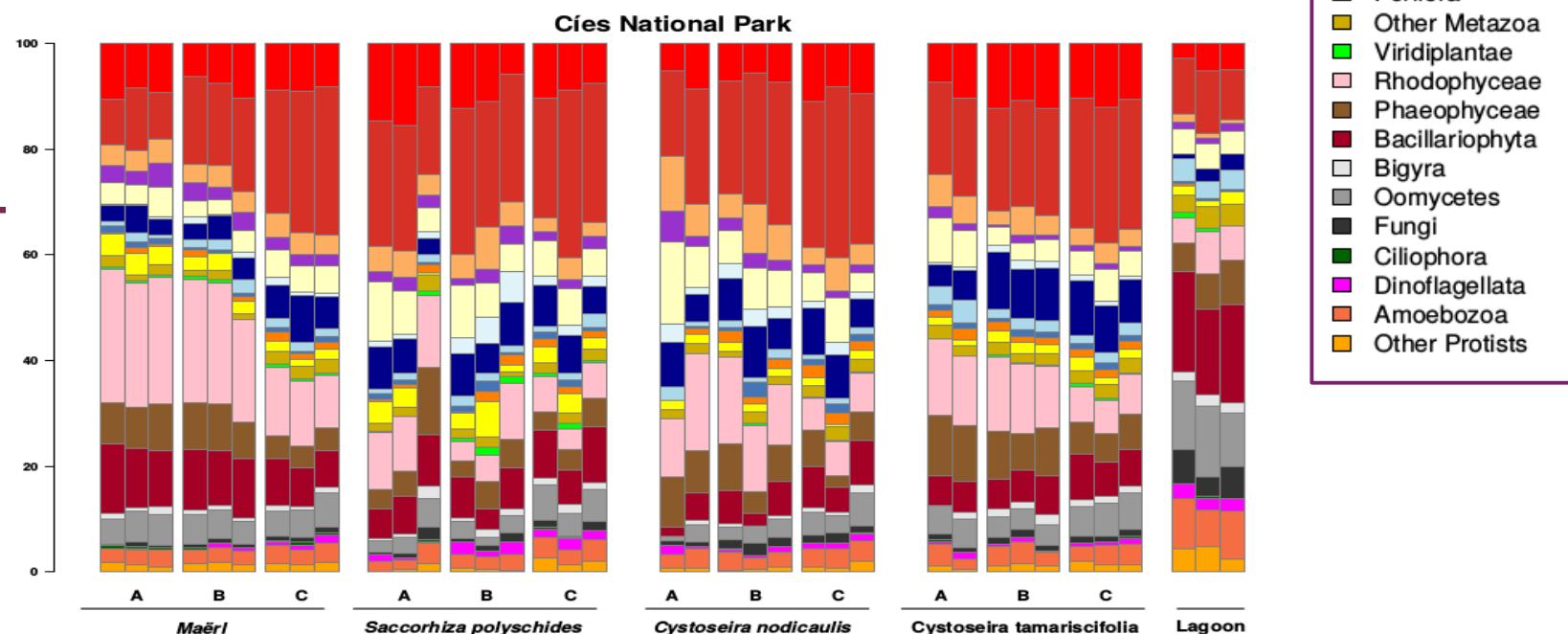


18S V7 % Assigned MOTUs

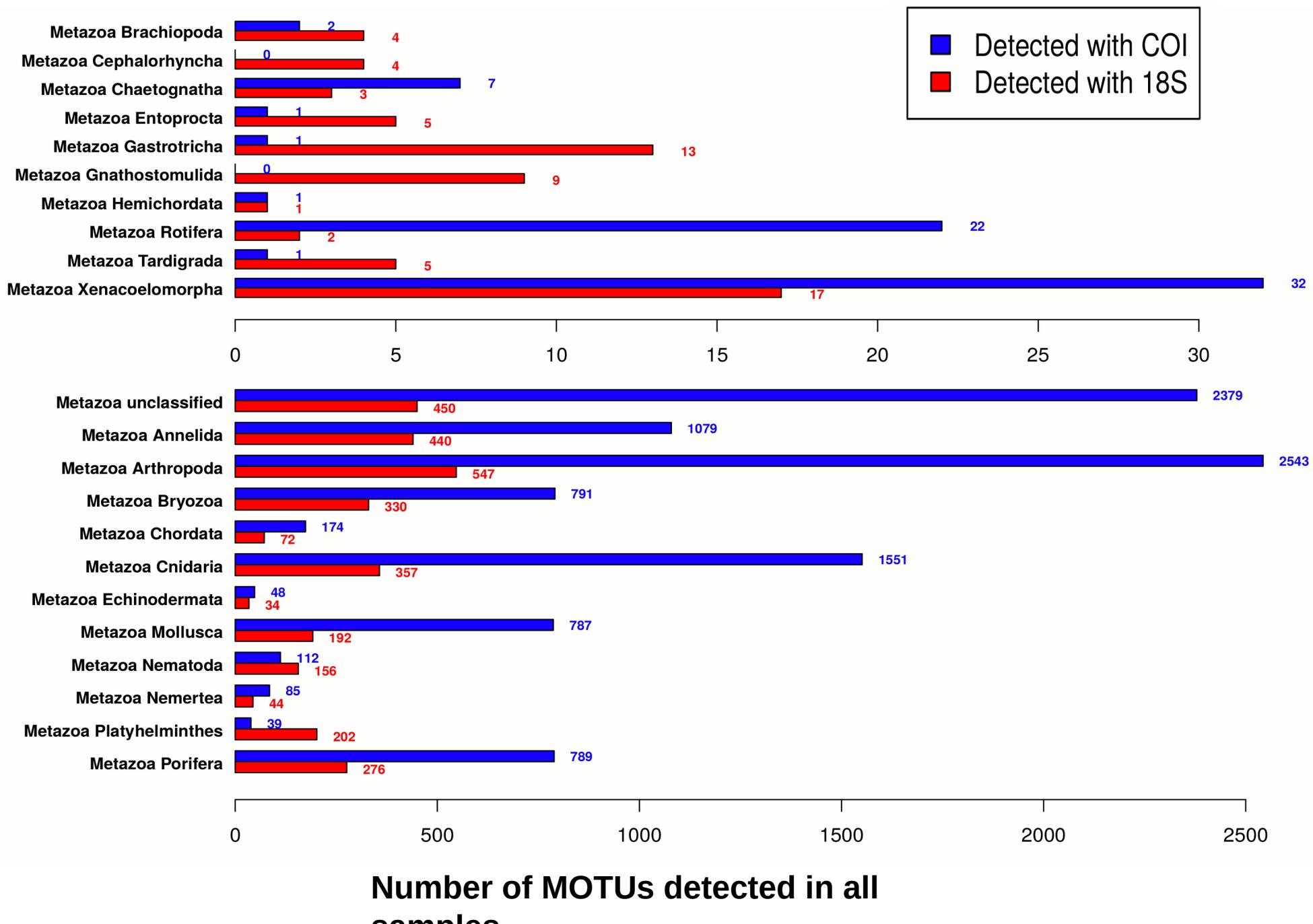
RELATIVE DIVERSITY (% MOTUS)



COI Leray-XT % Assigned MOTUs



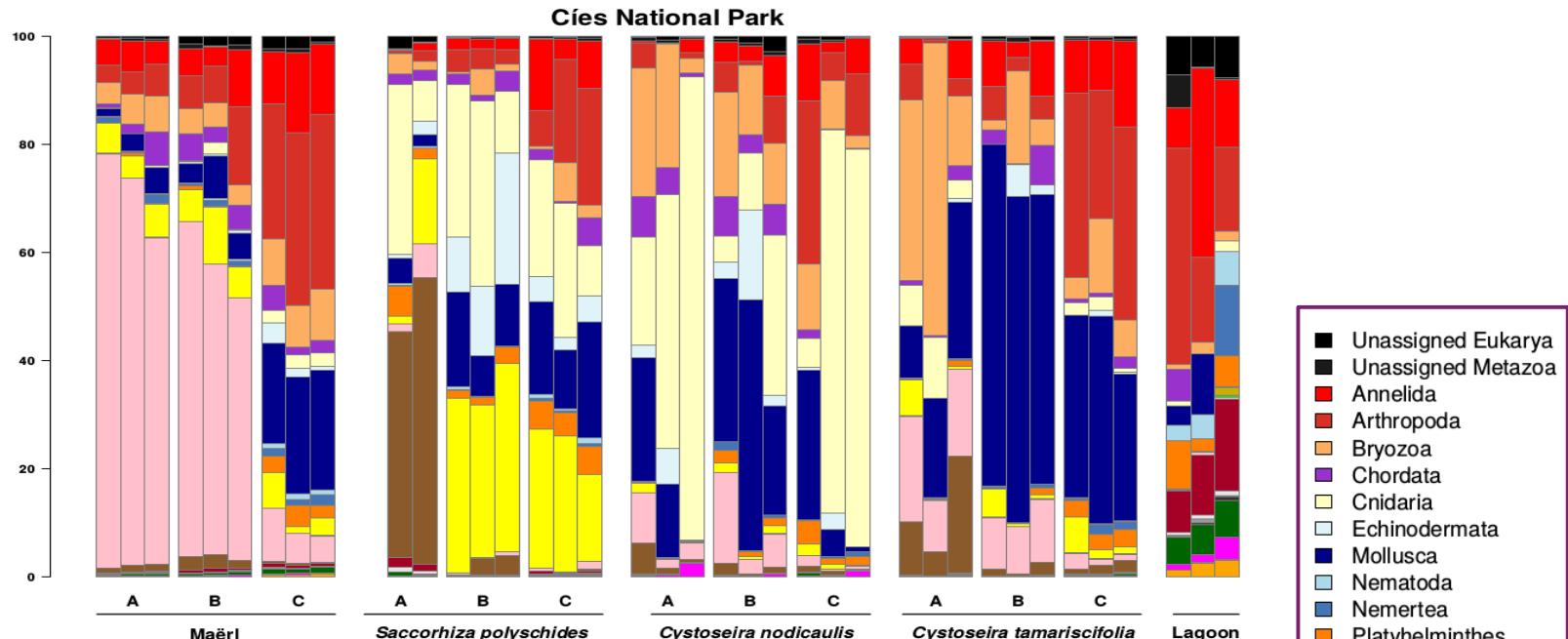
α -DIVERSITY: COMPARING TWO MARKERS



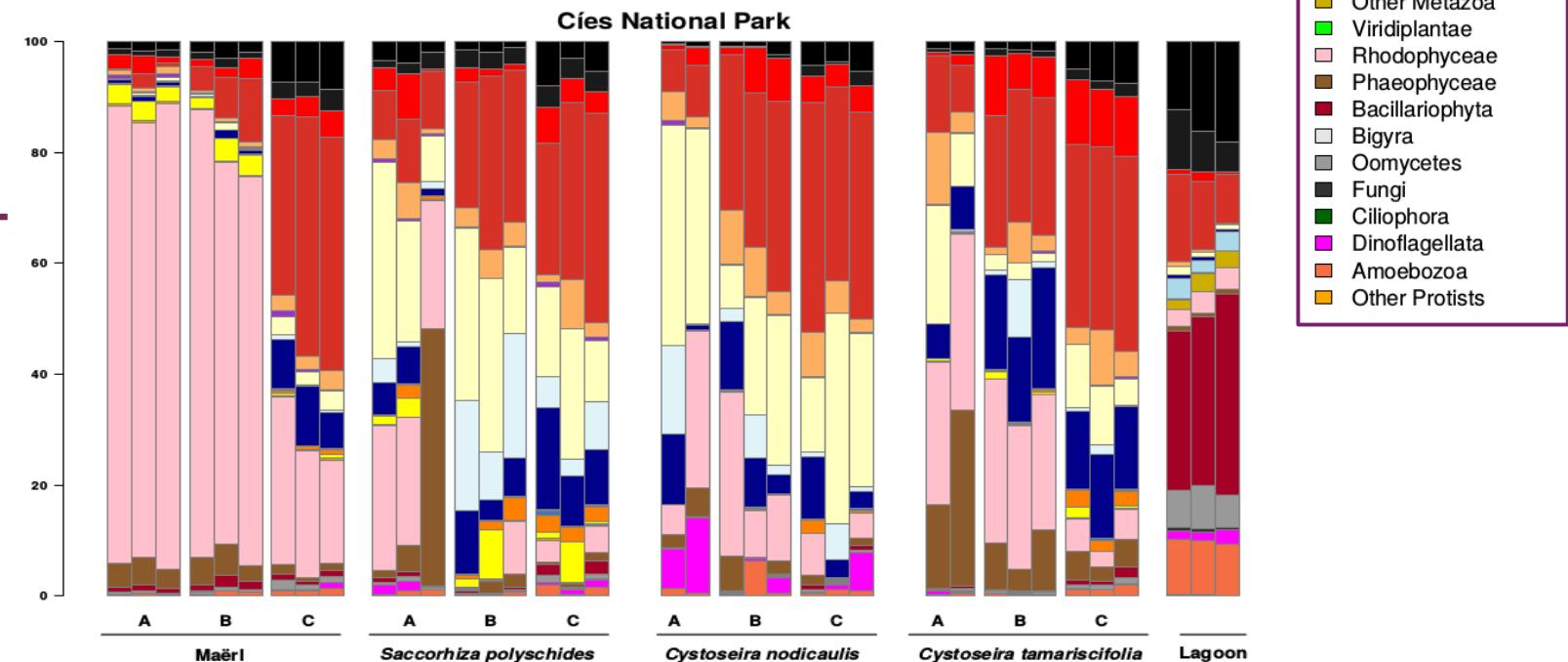
	samples
OTUS	diversity matrix

RELATIVE READ ABUNDANCE

18S V7
% Reads



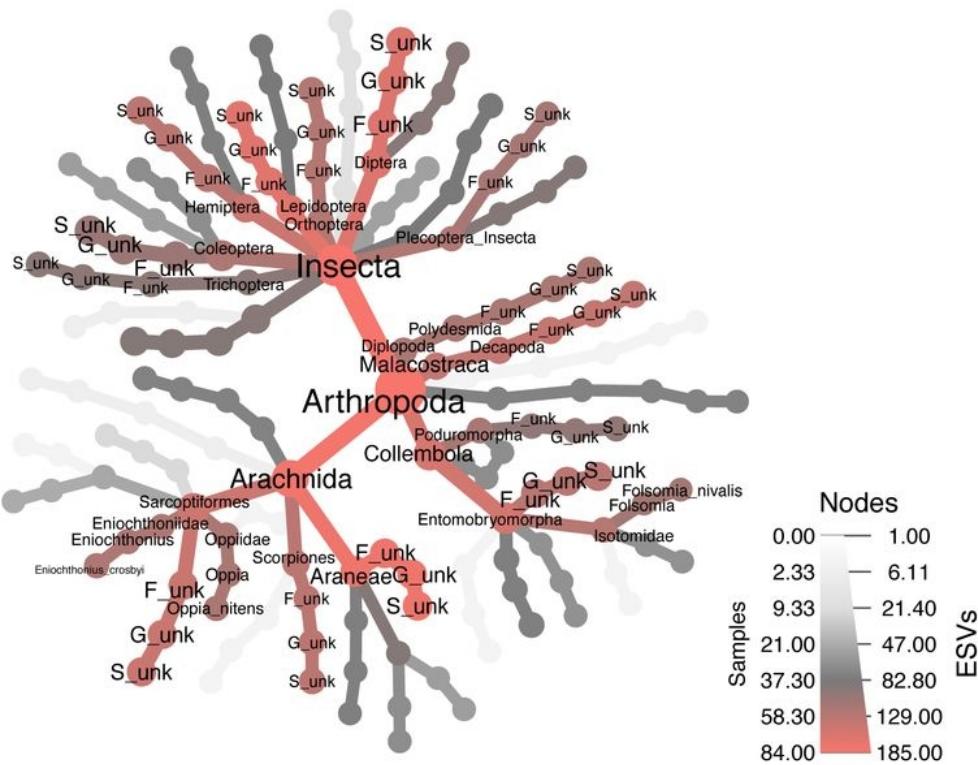
COI Leray-XT
% Reads



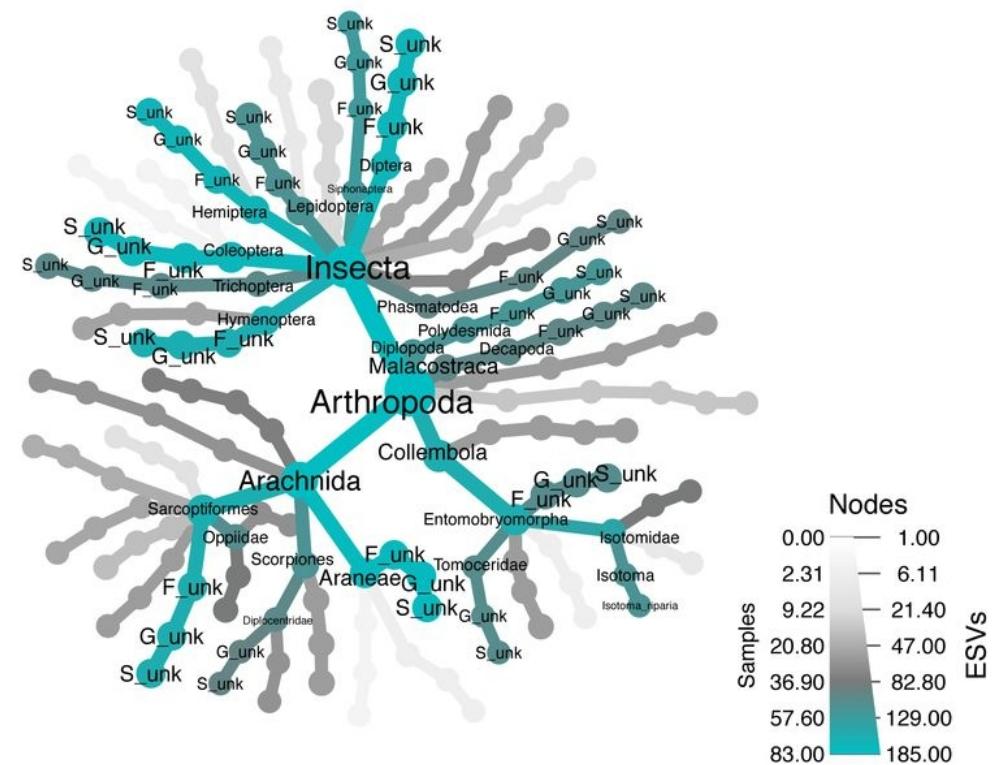
TAXONOMIC SUMMARY: HEAT TREES

(package MetaCoder)

Island Lake



Nimitz

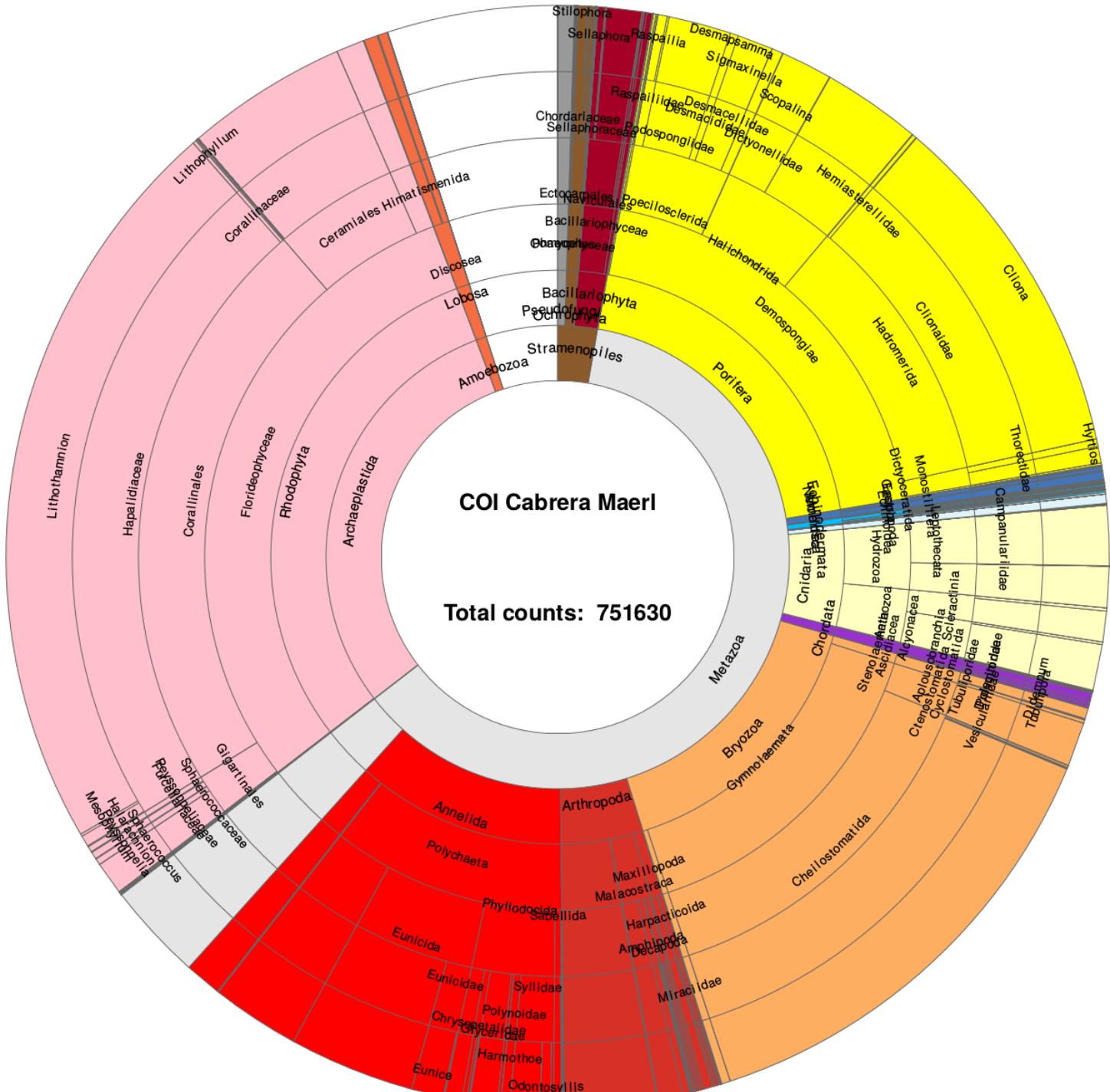


samples

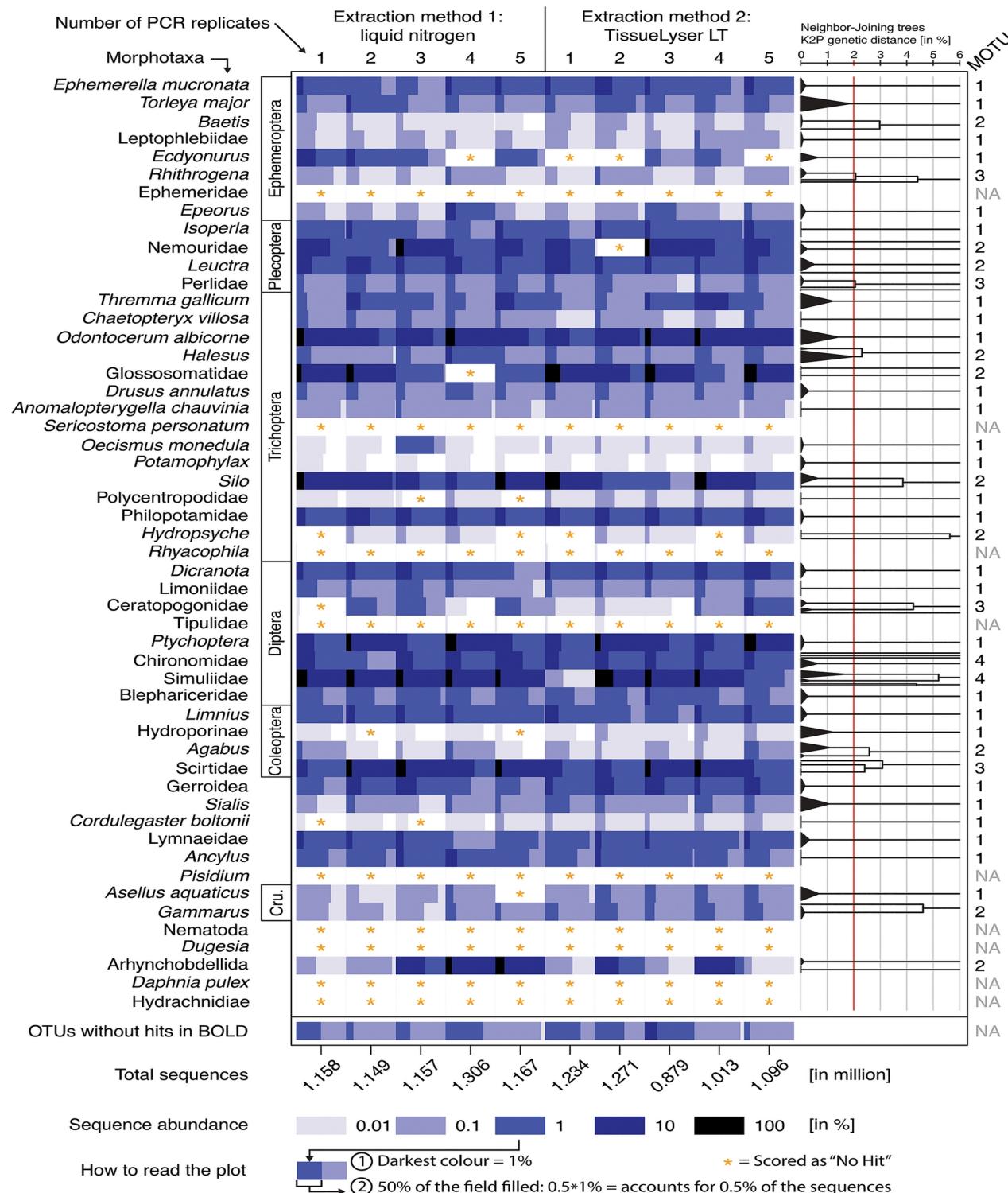
diversity matrix

TAXONOMIC SUMMARY: KRONA PLOT

Cabrera Maërl COL



β-DIVERSITY (HEAT MAPS)



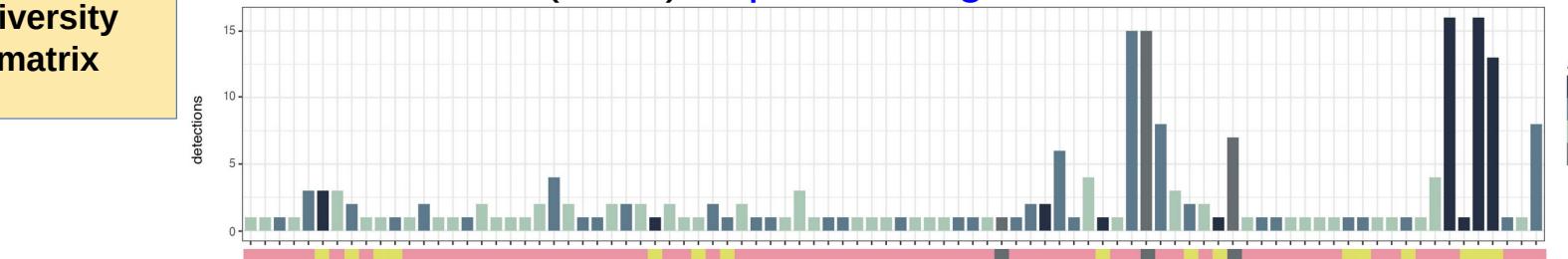
Elbrecht & Leese 2015.
<http://www.doi.org/10.1371/journal.pone.0130324>

samples
OTUS
diversity
matrix

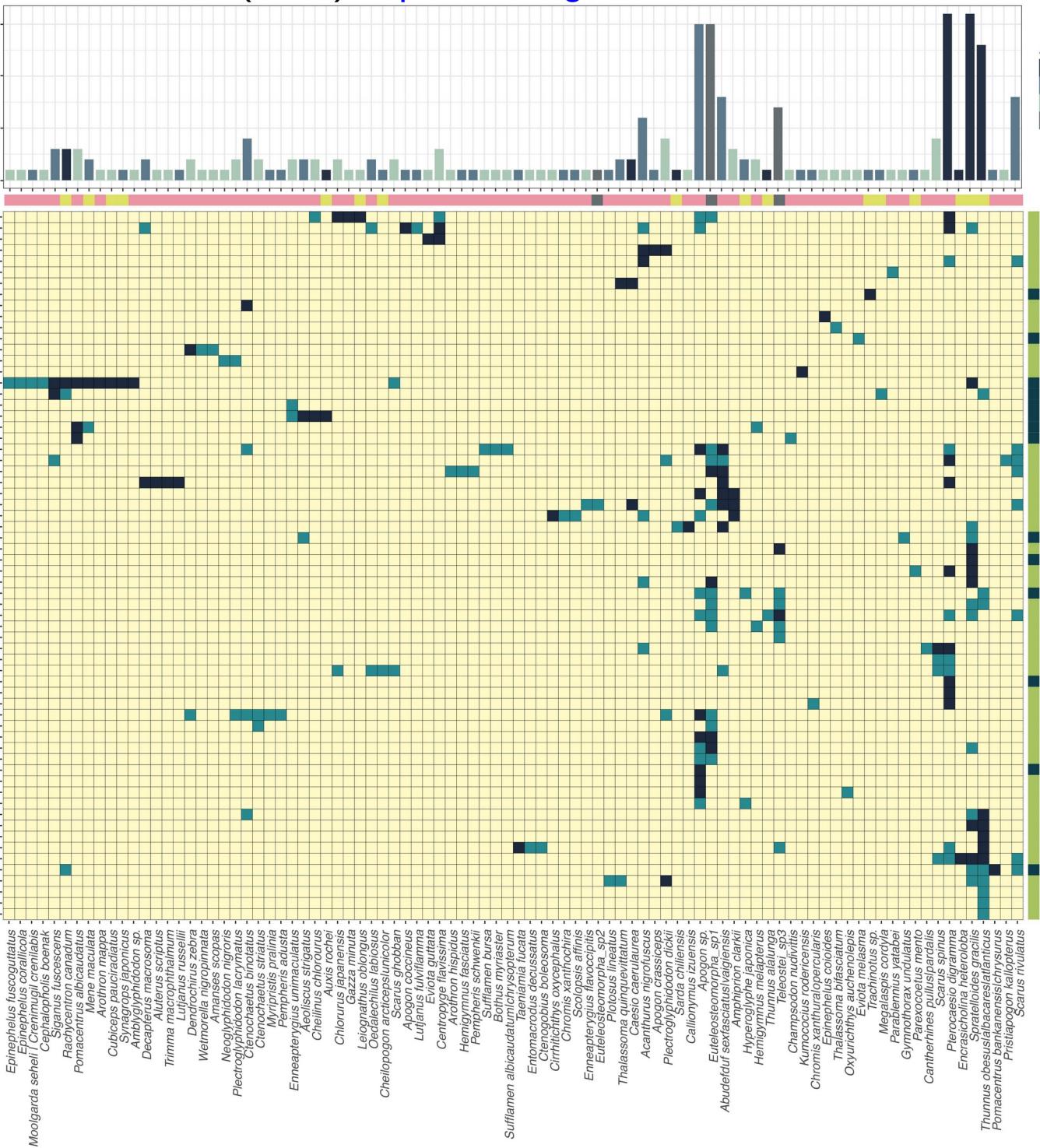
Turon et al. (2020) <https://doi.org/10.3389/fmars.2020.605148>

eco-traits
matrix

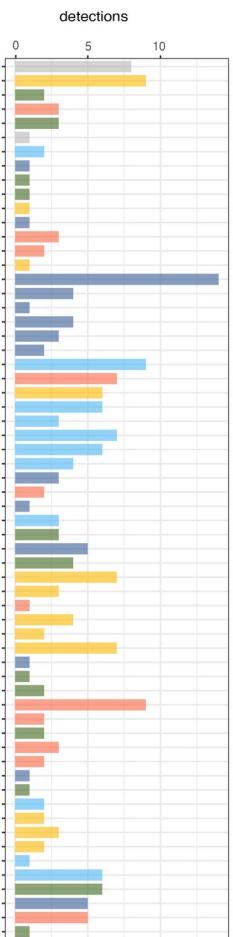
Sponge Samples



PCR detections
0 1 2



Aggregation
Habitat
Large
Small
Solitary
Unknown
Reef
Pelagic
Unknown

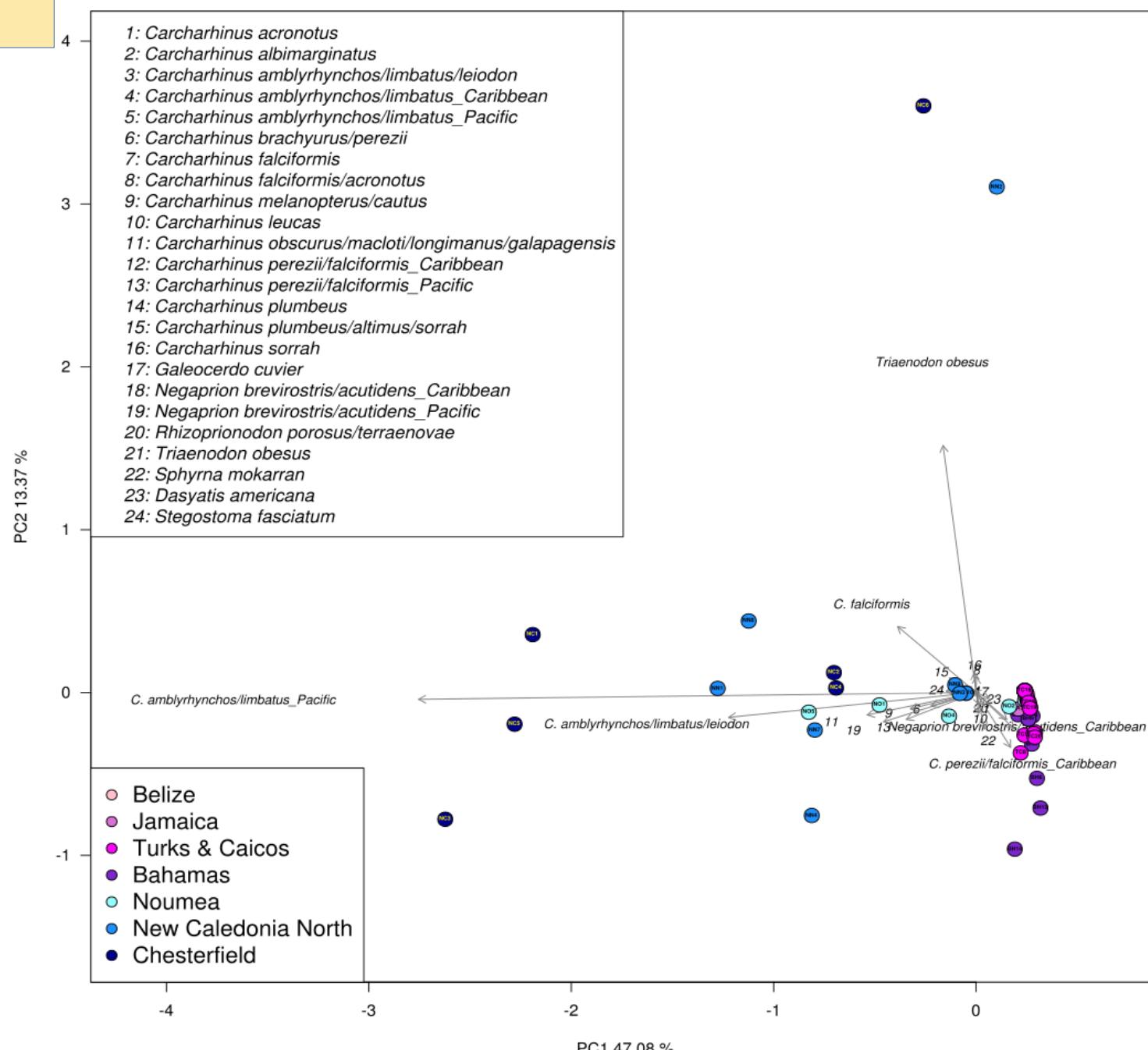


Env.
Eutrophic
Well-preserved
Location
Dambay bay
Dambay proximity
Hun Mun crack
Hun Mun North
Hun Mun South
Nock

metadata matrix

OTUS	samples
	diversity matrix

ORDINATION METHODS: PCA



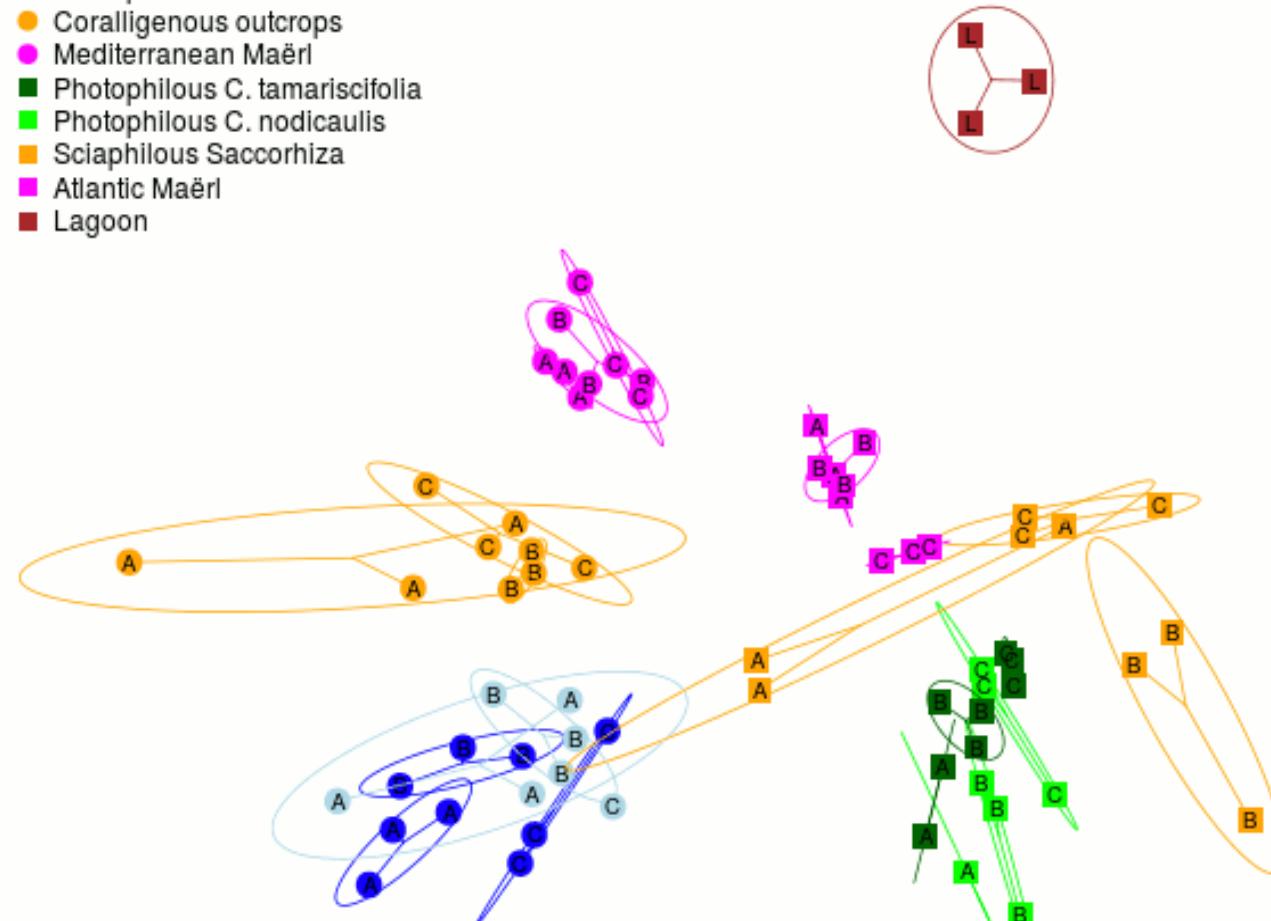
	samples
OTUS	diversity
	matrix

ORDINATION METHODS

NON-METRIC MULTIDIMENSIONAL SCALING

- Photophilous Lophocladia
- Photophilous Mediterranean
- Coralligenous outcrops
- Mediterranean Maërl
- Photophilous C. tamariscifolia
- Photophilous C. nodicaulis
- Sciaphilous Saccorhiza
- Atlantic Maërl
- Lagoon

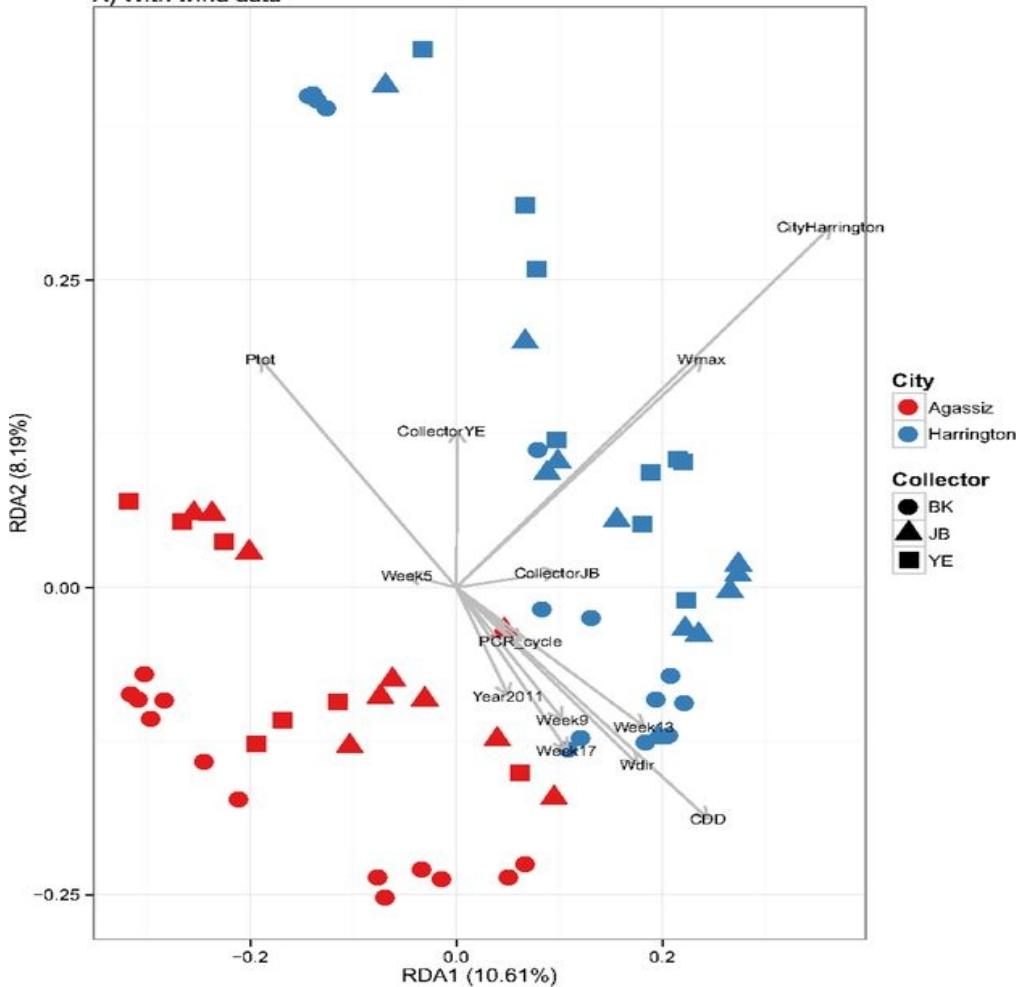
stress = 17.539 %



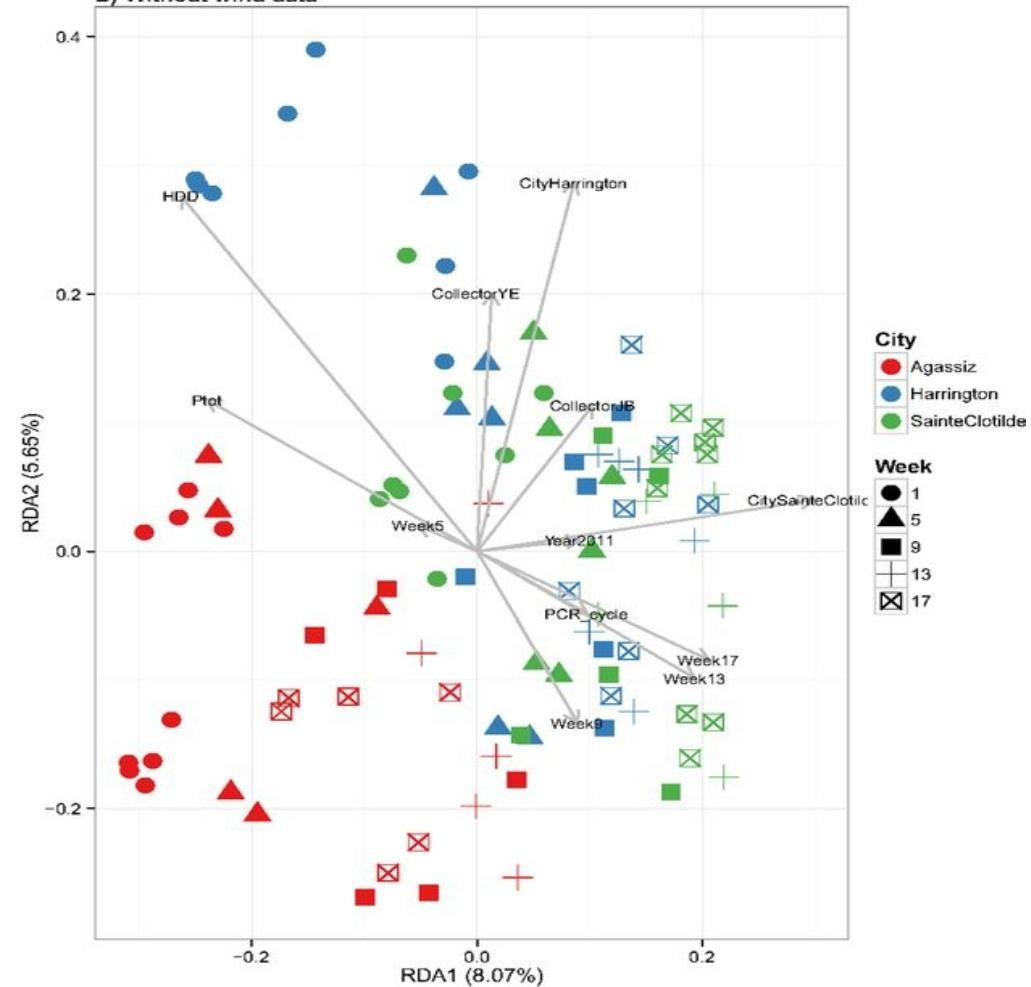
CORRELATION WITH ENVIRONMENTAL VARIABLES

REDUNDANCY ANALYSIS (RDA)

A) With wind data



B) Without wind data



Chen et al. (2018) <http://doi.org/10.1128/AEM.02601-17>

OTUS

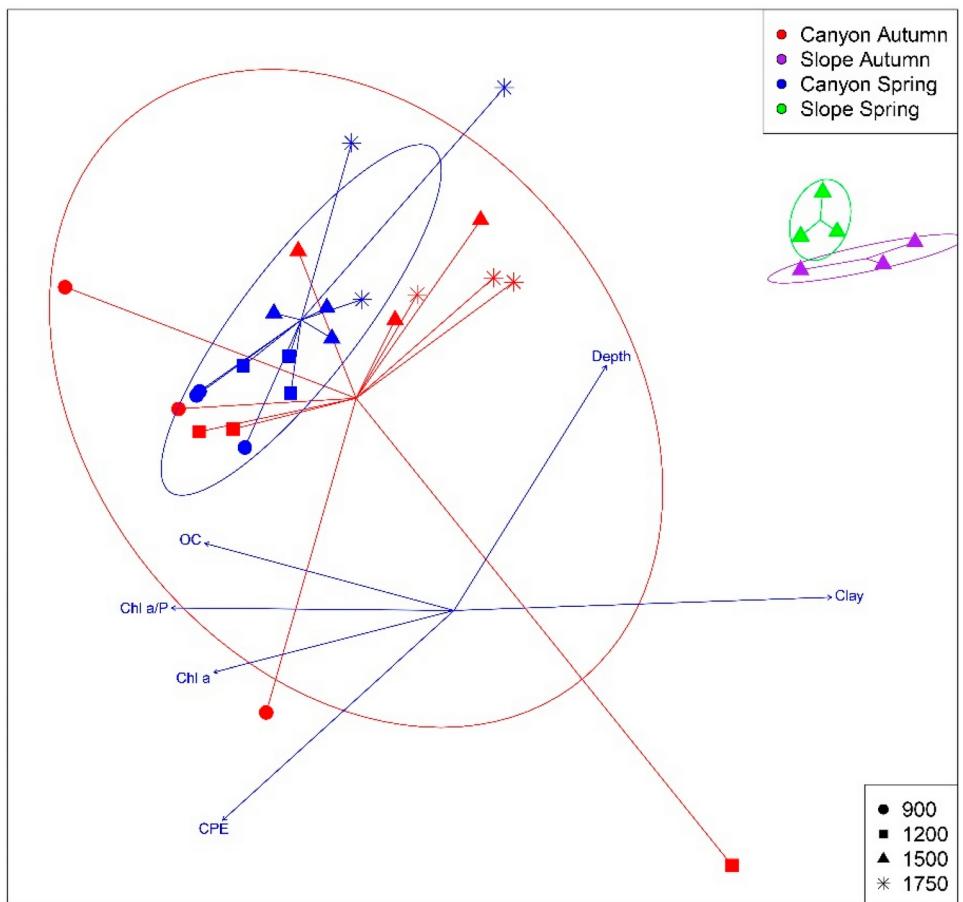
samples

diversity matrix

+

metadata matrix

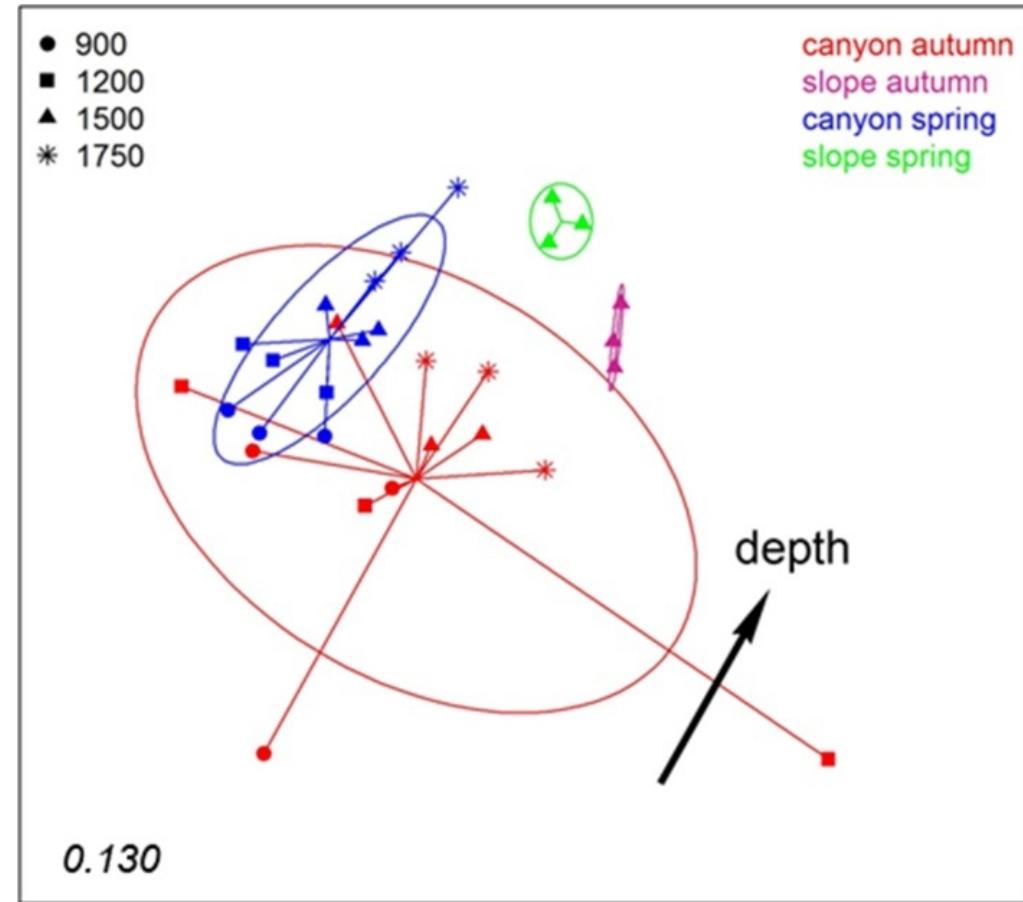
CORRELATION WITH ENVIRONMENTAL VARIABLES USING nMDS (ENVFIT)



COI

Atienza et al. (2020)

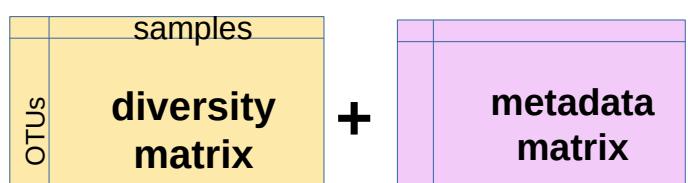
<https://doi.org/10.3390/d12040123>



18S

Guardiola et al. (2016)

<http://doi.org/10.7717/peerj.2807>





THANK YOU!