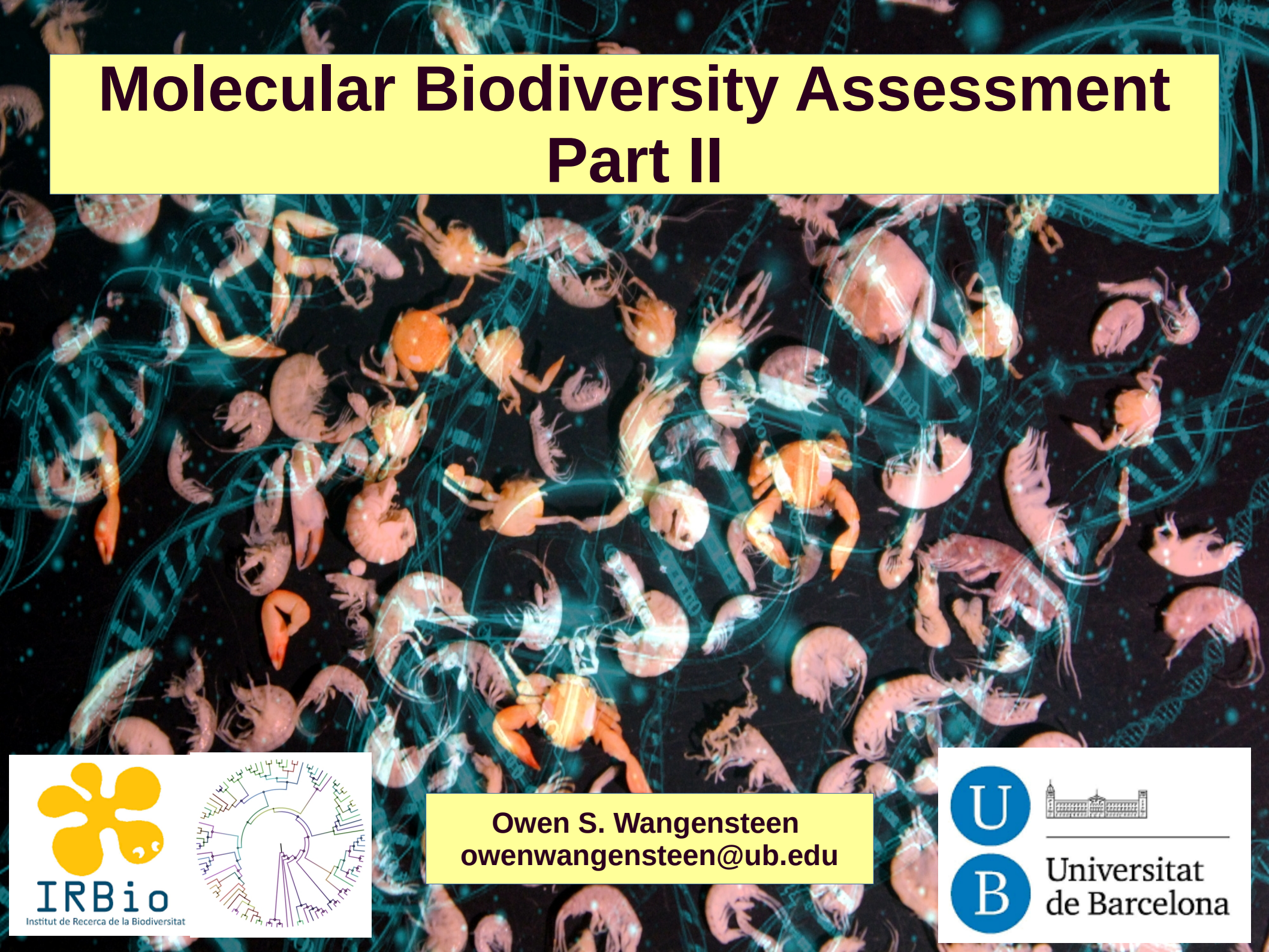
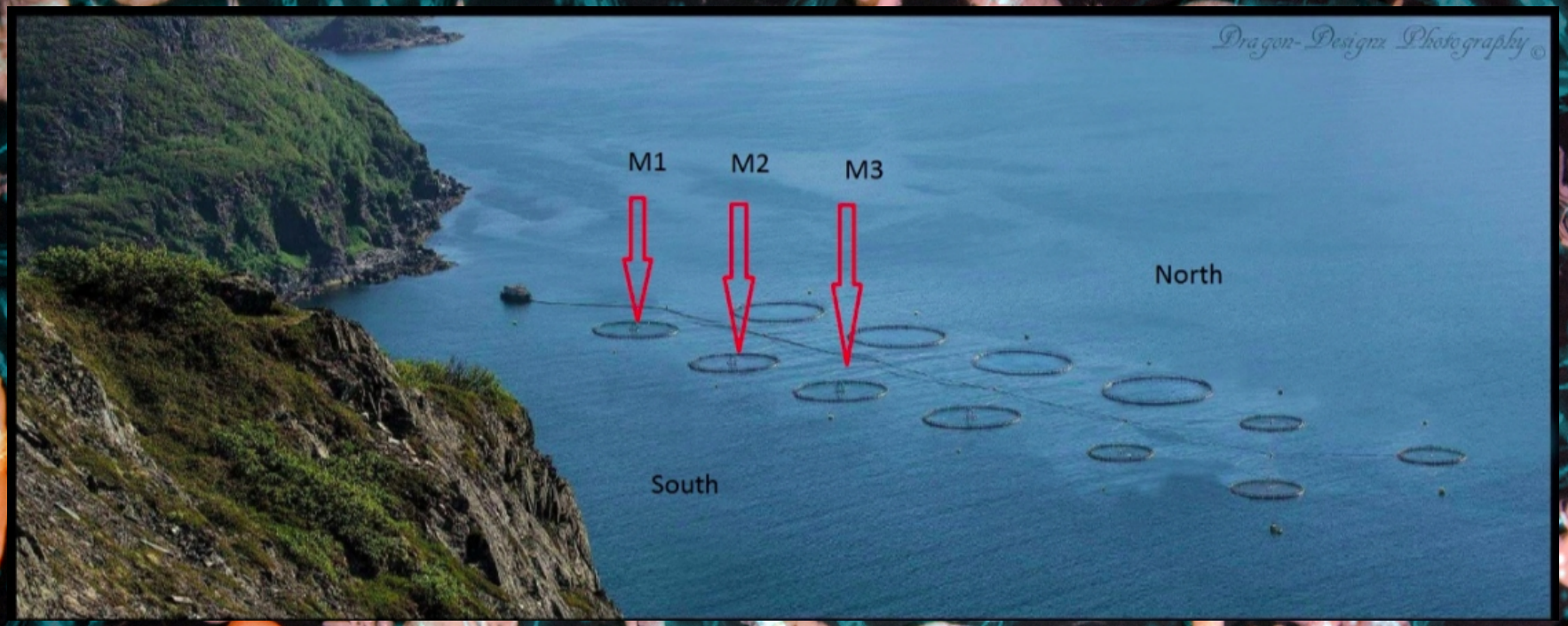
Molecular Biodiversity Assessment
Part II



Owen S. Wangensteen
owenwangensteen@ub.edu

IRBio
Institut de Recerca de la Biodiversitat

Universitat de Barcelona

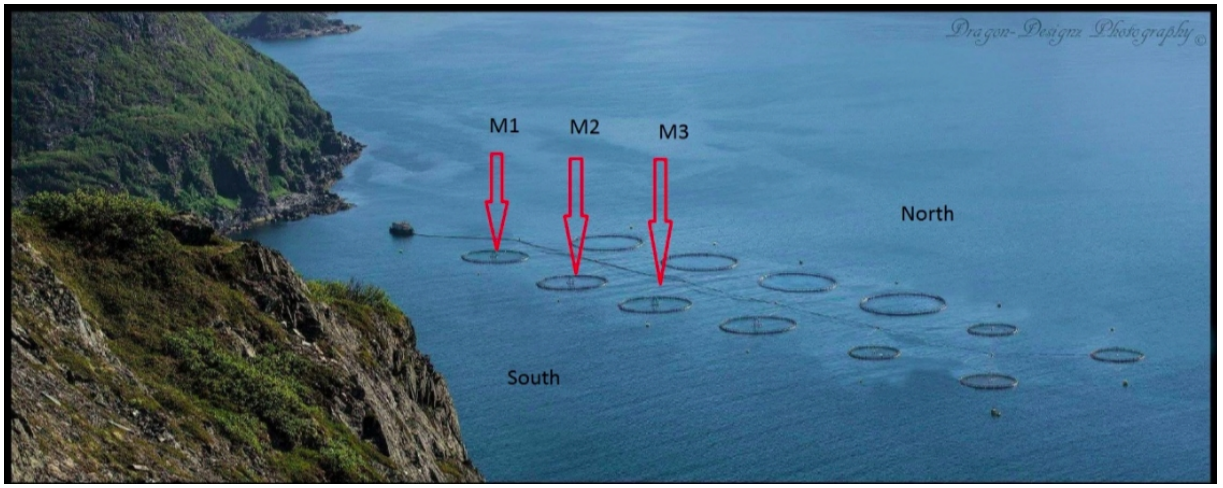


AND EXAMPLE USING THE MJOLNIR PIPELINE

The ULØY Dataset

4 multiplexed libraries:

- ULO1: 30 samples
- ULO2: 30 samples
- ULO3: 16 samples
- 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

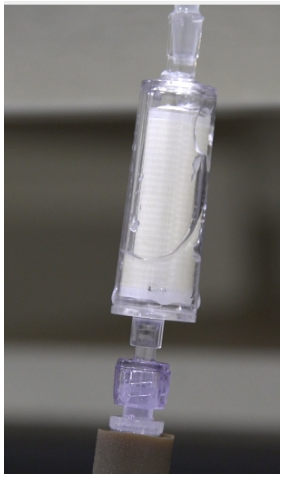
Total: 109 samples

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

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/17lX3z7DuWr3eSGePykGcgmV1XvfylQAh?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

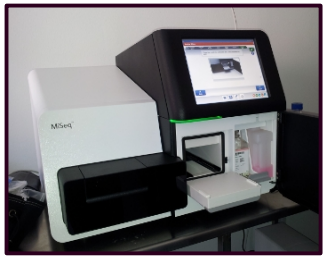
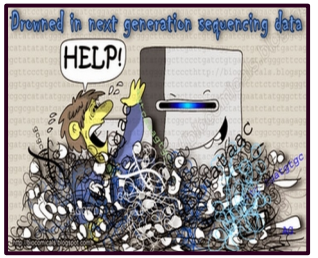
4. The pipeline itself:
mjolnir_ULOY.R

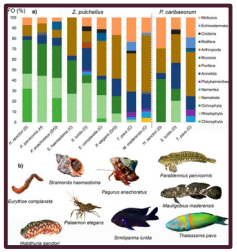
And the results are here: https://drive.google.com/drive/folders/1egIz3_U1pAT1UIA6MAq1hBgxO-3DjFdI?usp=sharing

METABARCODING WORKFLOW







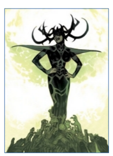
Sampling → Sample preprocessing → DNA extraction → Primer choice & PCR → HT sequencing → Bioinformatic pipeline → Results

Demultiplexing & quality filtering → Denoising / Clustering (define MOTUs & read count) → Taxonomic assignment (annotate MOTUs) → Final Refinement

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

START MJOLNIR

metadata file

demultiplexed FASTQ files

START MJOLNIR

multiplexed FASTQ files

ngsfilter tables

RAN
Reads Allotment in N portions

part_XX.fastq files

FREYJA
Filtering of Reads, Enrollment, Yoke-reads Joining and Alignment

LIBX_sample_XXX.fasta files

HELA
Hierarchical Elimination of Lurking Artifacts

LIBX_new.tab

LIBX_vsearch.fasta

ODIN
OTU Delimitation Inferred by Networks

LIBX_output_counts.tsv

LIBX_seeds_nonsingleton.fasta

DUFA reference

NCBI taxonomy

Higher ranks

THOR
Taxonomy with Higher-than-Order Ranks

LIBX_ecotag_annotated.tsv

FRIGGA
Final Recount and Integration of Generated Genealogies & Abundances

LIBX_All_MOTUs.tsv

LOKI
LULU Overseeing with Kinship Identification

LIBX_curated_LULU.tsv

metadata file

RAGNAROC
Replace AGnomens with Names And Recover Original Codification

Final_dataset.tsv

END 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
obitab: 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

RAGNAROC

- 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

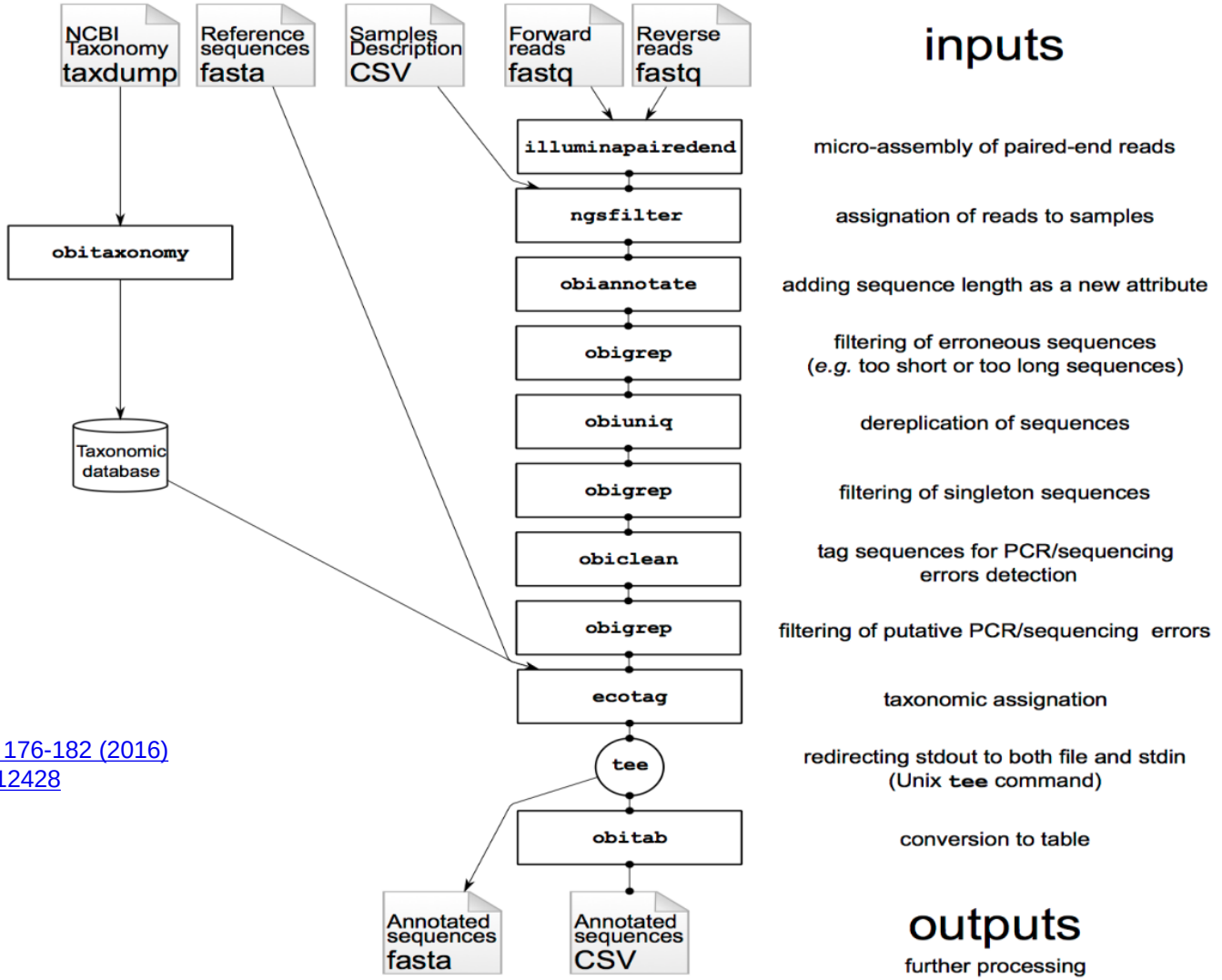
A MJOLNIR pipeline is an R script:

```
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,tax_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

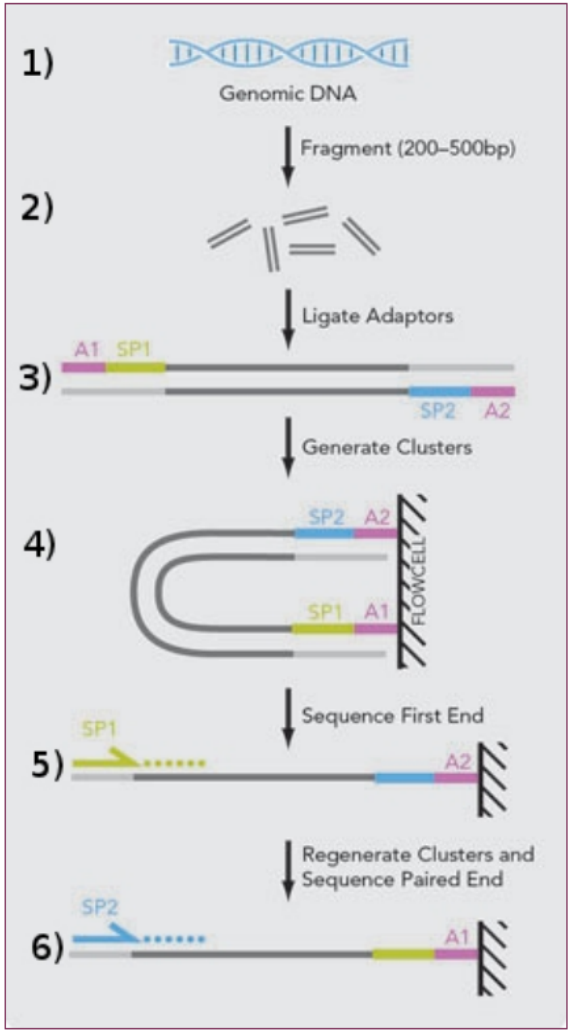
Frédéric Boyer, Céline Mercier, Aurélie Bonin, Yvan Le Bras, Pierre Taberlet, Eric Coissac

OBITools BASIC WORKFLOW



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

illuminapairedend (FREYJA)



>Read_1_Forward

TAATCTGACCTTTAGACCTCGGGGTGGCCGAAAAATCAAAAGAGATGTTGGTAAAGGACGGGATCGCC
GCCTCCTACAGGATGGAAGAATGGAGTATTAAAGTTCCGGTCAGTGAGAAGTATAGTAATTGCGCCGG
CTAAGACTGGGAGTGAGAGAAGTAGGAGCACTGTTGTAACAATAAGGGATCATACAAATAGAGGCACA
CGCTCTCCTGTTATTCCCTGCTTACGTGTGGTTCGTGCTGTTACTA

>Read_1_Reverse

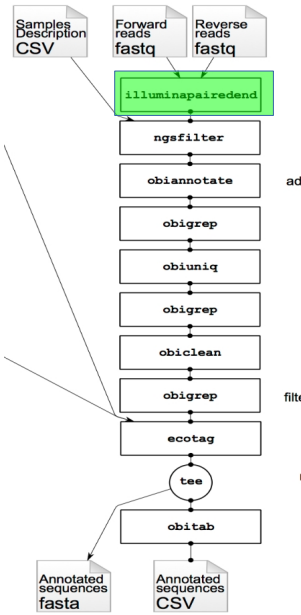
TCCTGACCTTGGTACTAGATGGACTGTGTATCCTCCTCTCGCAGGAAATTCATTCCACAGCGGCCCCG
CAGTAGATATAGCCATCTTCTCCCTCCATCTTGCCGGGGCCAGATCCATTCTAGGCTCTATCAATTTC
TTCGTAACAGCACGAACCACTCGTAAGCAGGGATTAACATGAGATCGTGTTCCCCTCTTTTTATTCTC
CCTTATTTTTCCCACCTTTCTCCTTCTTCTCTCCCTCCCCGTCTTAT

>Read_1_Reverse_complement

ATAAGACGGGGAGGGAGAGAAGAAGGAGAAAGGTGGGAAAAATAAGGGAGAATAAAAAGAGGGGAACA
CGATCTCATGTTAATCCCTGCTTACGAGTGGTTCGTGCTGTTACGAAGAAATTGATAGAGCCTAGAAT
GGATCTGGCCCCGGCAAGATGGAGGGAGAAGATGGCTATATCTACTGCGGGGCCGCTGTGGAATGAAT
TTCCTGCGAGAGGAGGATACACAGTCCATCTAGTACCAAGGTCAGGA

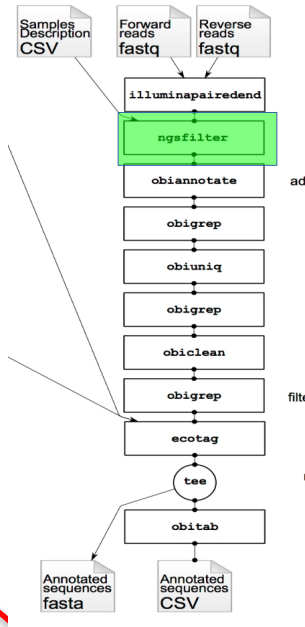
>Read_1_aligned

taatCTGACCTTTAGACCTCGGGGTGGCCGAAAAATCAaaagagatgttggtaaaggacgggatcgcc
gcctcctacaggatggaagaatggagtattaaagttccggtcagtgagaagtatagtaattgcgccgg
ctaagactgggagtgagagaagtaggagcactgttgtaacaataagggatcatacaaatagaggcaca
cgctctcctgttattccctgcttacgagtggttcgtgctgttactaagaaattgatagagcctagaat
ggatctggccccggcaagatggagggagaagatggctatatctactgcggggccgctgtggaatgaat
ttcctgcgagaGGAGGATACACAGTCCATCTAGTACCAAGGTCAGga



OBITOOLS:

ngsfilter (demultiplexing) **(FREYJA)**

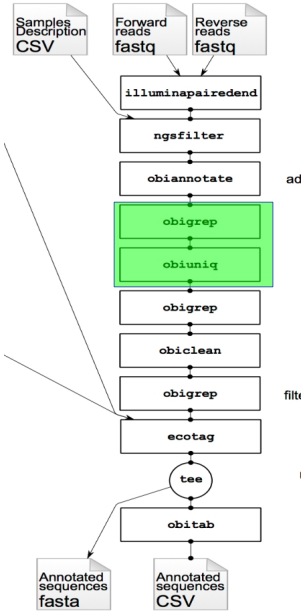


nntagccacttagacttcggggatggccaaagaatcaaaaaagatgttgataaagaactgggtctccacctcctgctgggtcaaaaaatgatgta
ttcagatttcgatcagttaataatatagtaattgcacctgctaaaacgggtaatgagggtaaaagtaaaattgctgtaataaaaact**gctcatacaa**
ataatggtatacggtctatagttattncaacaggtcgcatgttaataactgttgttataaaatttactgctcctaaaatagaggatacccctgctag
atgaagagaaaaaattcctatatcaactgatgctcctgcgtgagcaattctggctgataaagggggtaaacagttcatccagttccagtggctan
nnn

ngsfilter table:

lib	sample	tags	forward_primer	reverse_primer
ULO1	ULO1_sample_006	ggatgatc:ggatgatc	GGWACWRGWTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
ULO1	ULO1_sample_007	ggattcga:ggattcga	GGWACWRGWTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
ULO1	ULO1_sample_008	tagcaagg:tagcaagg	GGWACWRGWTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
ULO1	ULO1_sample_011	tgaggaca:tgaggaca	GGWACWRGWTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
ULO1	ULO1_sample_012	cgatacac:cgatacac	GGWACWRGWTGRACWNTNTAYCCYCC	TANACYTCNGGRTGNCCRAARAAYCA
ULO1	ULO1_sample_013	tagccact:tagccact	GGWACWRGWTGRACWNTNTAYCCYCC	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=ULO1; count=16103;
merged_sample={'ULO1_sample_013': 2309, 'ULO1_sample_027': 3224, 'ULO1_sample_001': 2316,
'ULO1_sample_005': 3375, 'ULO1_sample_012': 307, 'ULO1_sample_020': 1995, 'Empty_ULO1': 3};
aaaaagatgttgataaagaactgggtctccacctcctgctgggtcaaaaaatgatgtattcagatttcgatcagttaataatatagtaattgcacctgctaa
aacgggtaatgagggtaaaagtaaaattgctgtaataaaaactgctcatacaaataatggtatacggtctatagttattncaacaggtcgcatgttaataac
tgttgttataaaatttactgctcctaaaatagaggataccctgctagatgaagagaaaaaattcctatatcaactgatgctcctgcgtgagcaattctggc
tgataaa
```

VSEARCH: uchime_denovo (HELA)

Chimera removal

UCHIME algorithm

```
A     81 CCTTGGTAGGCCGtTGCCCTGCCAACTAGCTAATCAGACGCgggtCCATCtcaCACCaccggAgtTTTtcTCaCTgTacc 160
Q     81 CCTTGGTAGGCCGCTGCCCTGCCAACTAGCTAATCAGACGCATCCCCATCCATCACCGATAAATCTTTAATCTCTTTCAG 160
B     81 TCTTGGTgGGCCGtTaCCCcGCCAACaAGCTAATCAGACGCATCCCCATCCATCACCGATAAATCTTTAAaCTCTTTCAG 160
Diffs    A       A      p A    A       A               BBBB      BBB     BBBBB BB    BBa B   B BBB
Votes    +       +      0 +    +       +               ++++      +++     +++++ ++    ++! +   + +++
Model    AAAAAAAAAAAAAAAAAAAAAAAAAAxxxxxxxxxxxxxxxBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

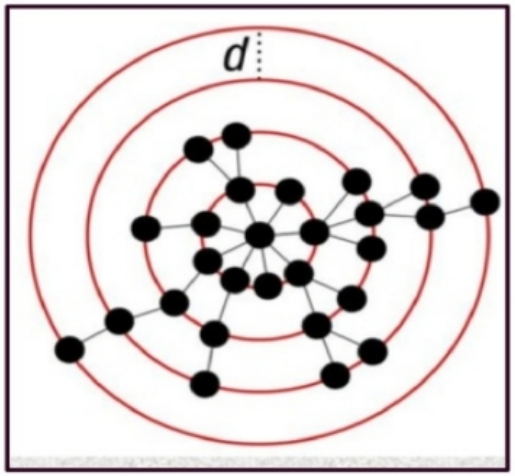
Region from a chimeric alignment generated by UCHIME.
Diffs and votes are annotated. The 'Model' row indicates the three segments of the alignment which are closer to A, the crossover (X), and closer to B, respectively. Diffs are 'A'=diff with Q closer to A in the A segment, 'a'=diff with Q closer to 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.

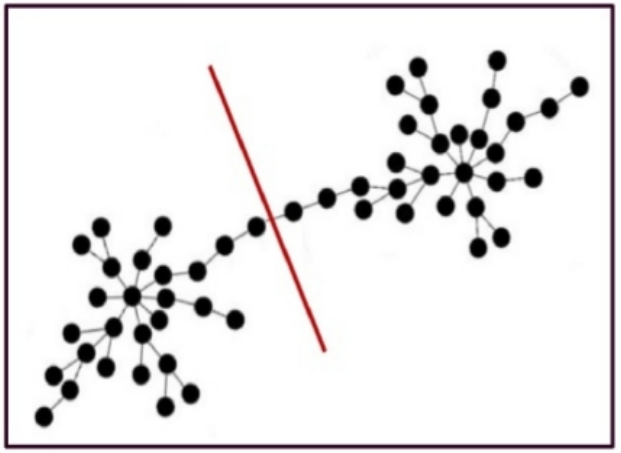
uchime_ref : parent sequences are provided by the user. "Knowing" which sequcenes should appear.
uchime_denovo : parent sequences are calculated in base of the abundance, as they are assumed to be more abundant than the chimeras.

FLEXIBLE CUTOFF CLUSTERING: SWARM (ODIN)

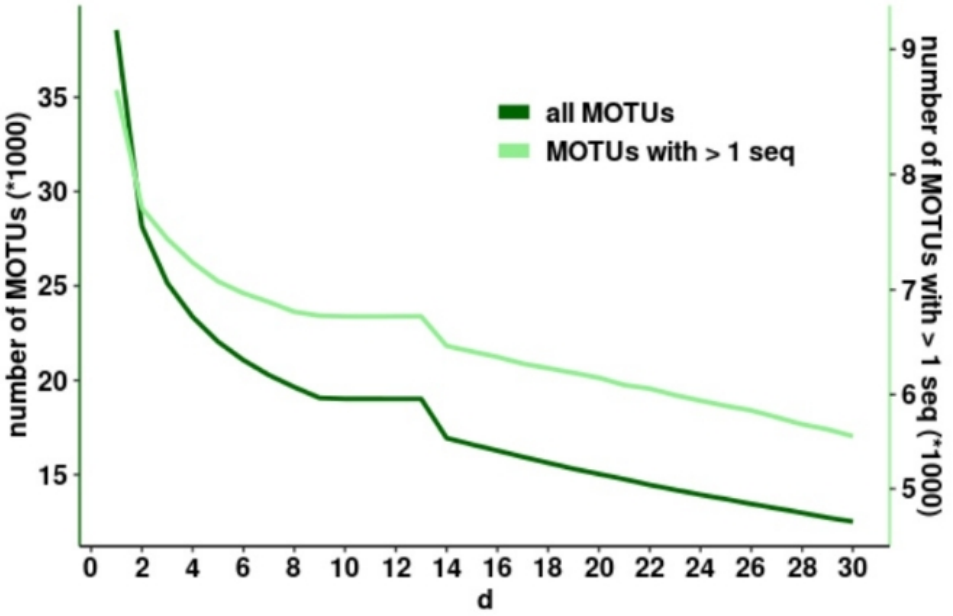
SWARM: Step-by-step Aggregation Algorithm

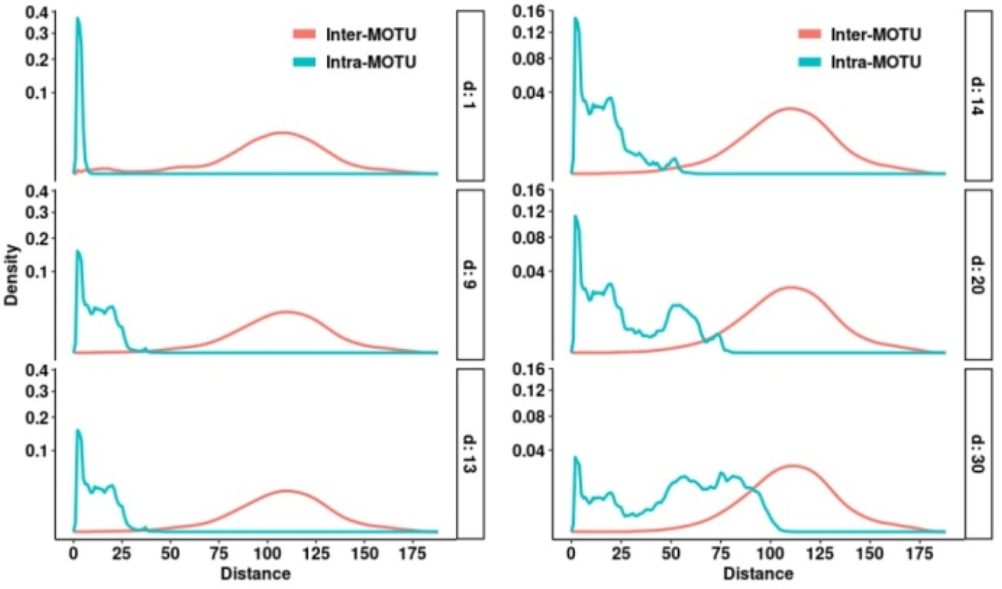
(Mahé et al. 2014)



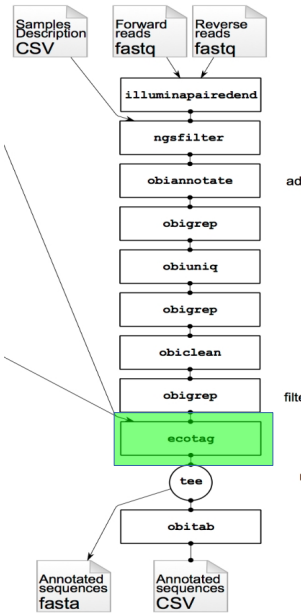


- Deterministic
- Fast!
- Behaves well for COI at d= 9 - 13



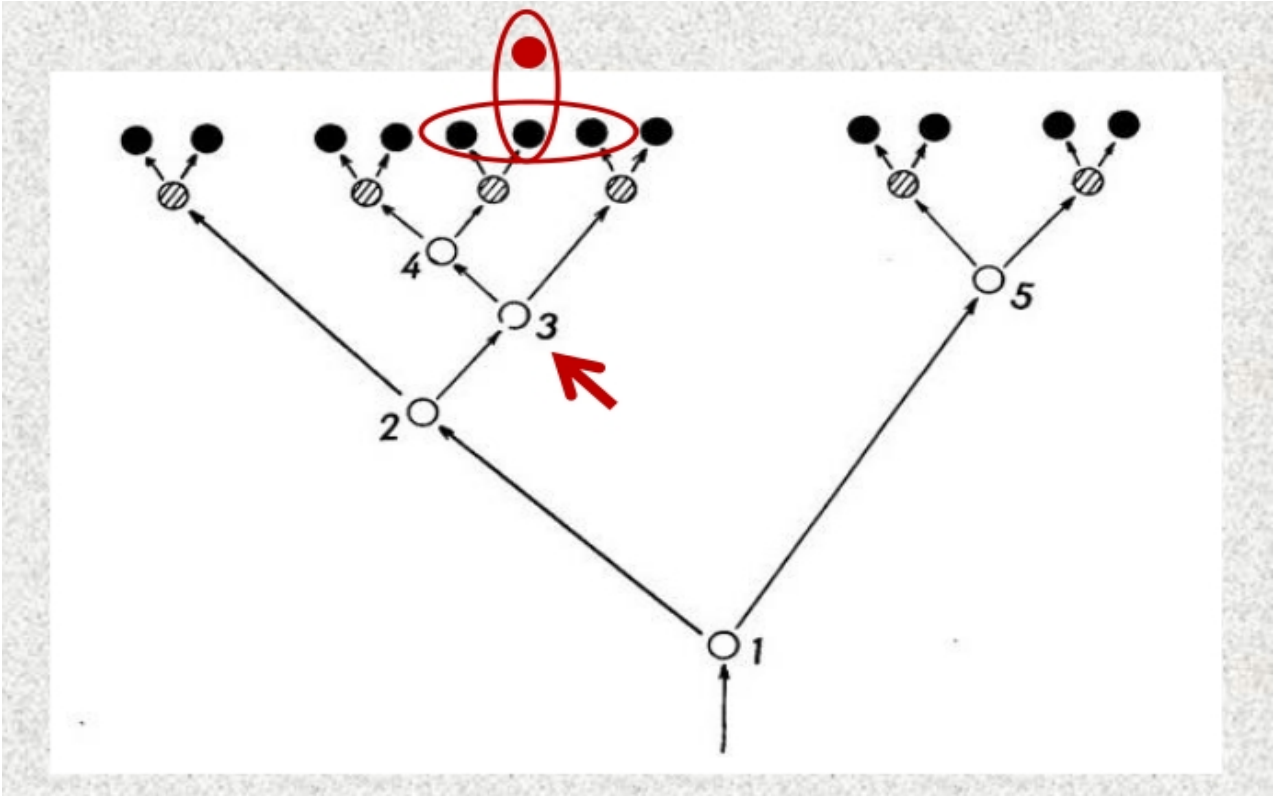


ecotag (taxonomic assignment) **(THOR)**



- **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-occurence and abundance

- Sequence similarity

Needs:

- Table with abundances in every sample

- List of pairwise matches (identity %) of MOTU sequences

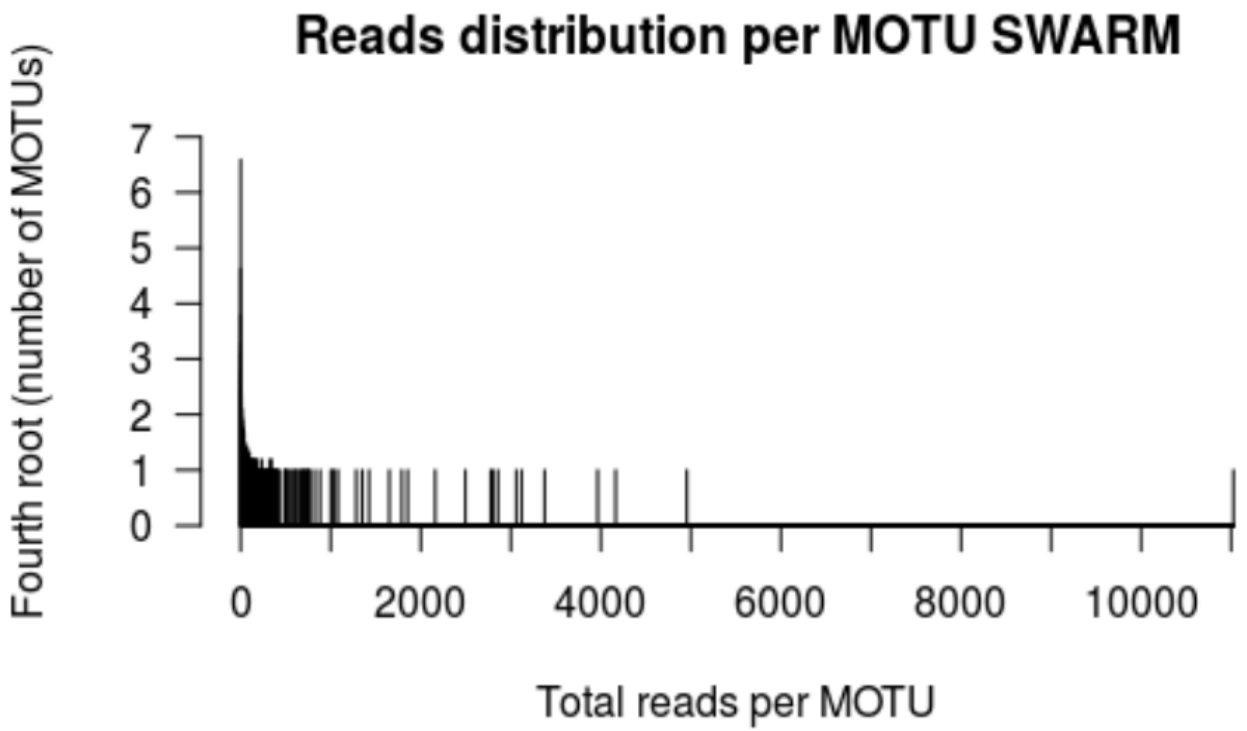
Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications* 8(1), 1188. list

FINAL REFINEMENT: ABUNDANCE FILTERING

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

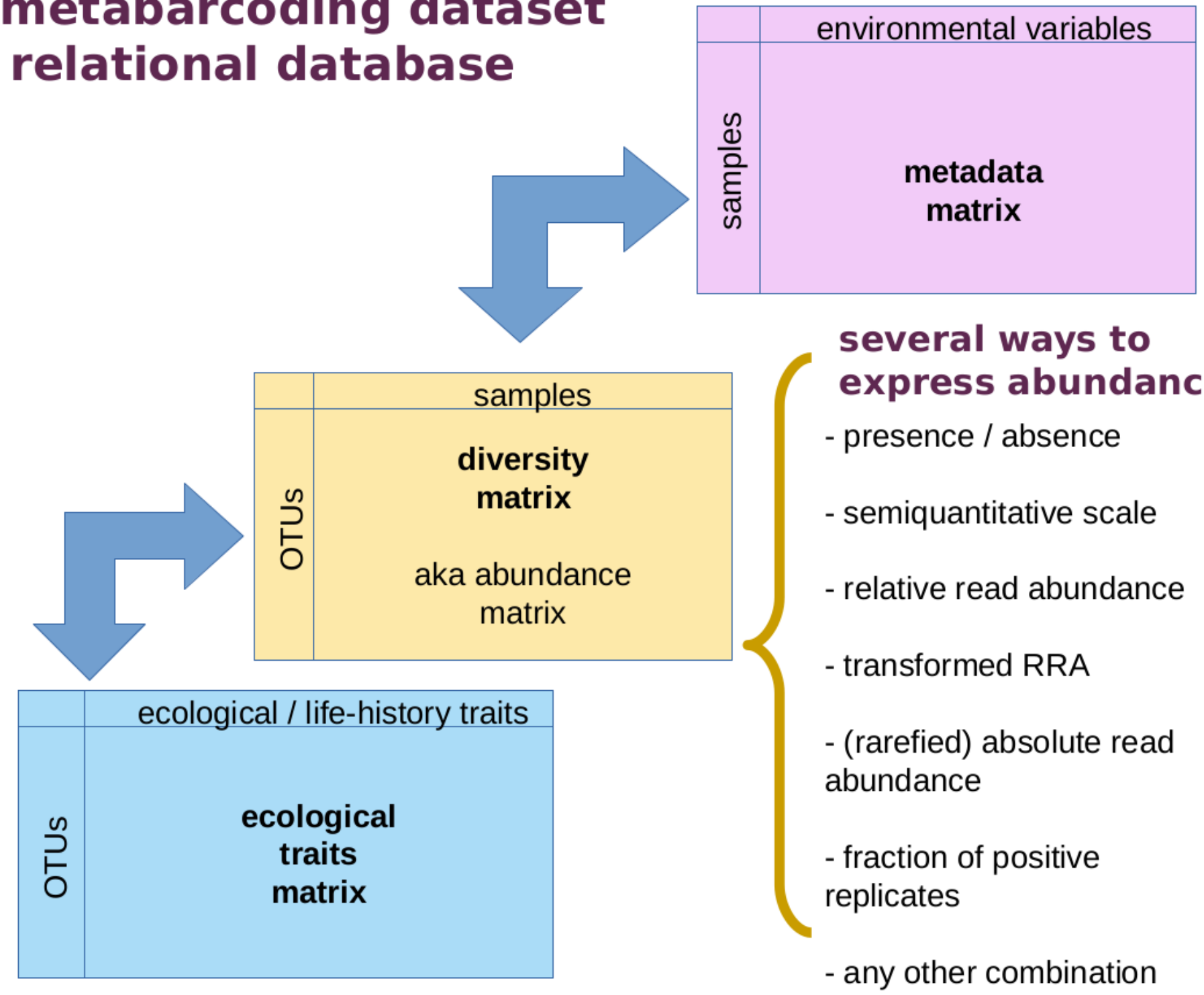
(RAGNAROC)

MINIMAL ABUNDANCE FILTERING



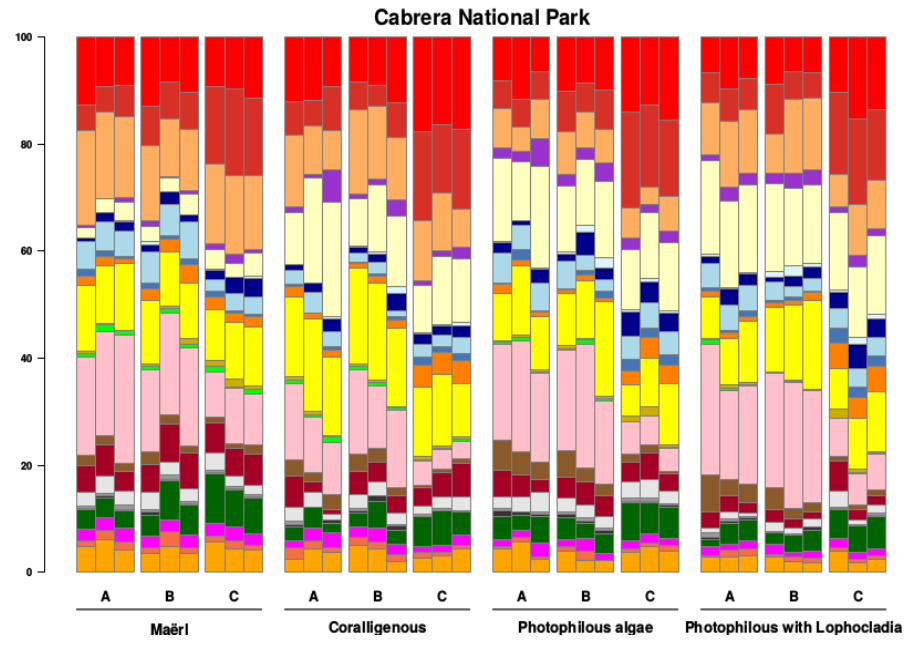
Negative binomial distribution

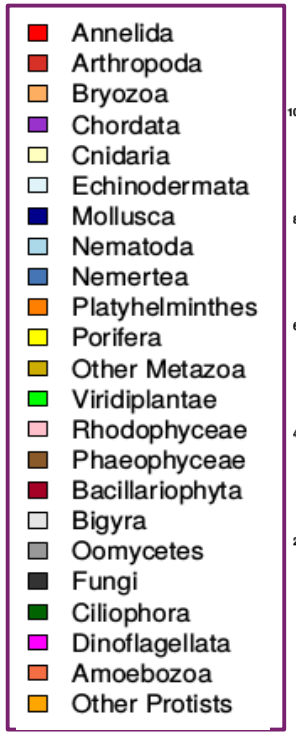
The metabarcoding dataset as a relational database



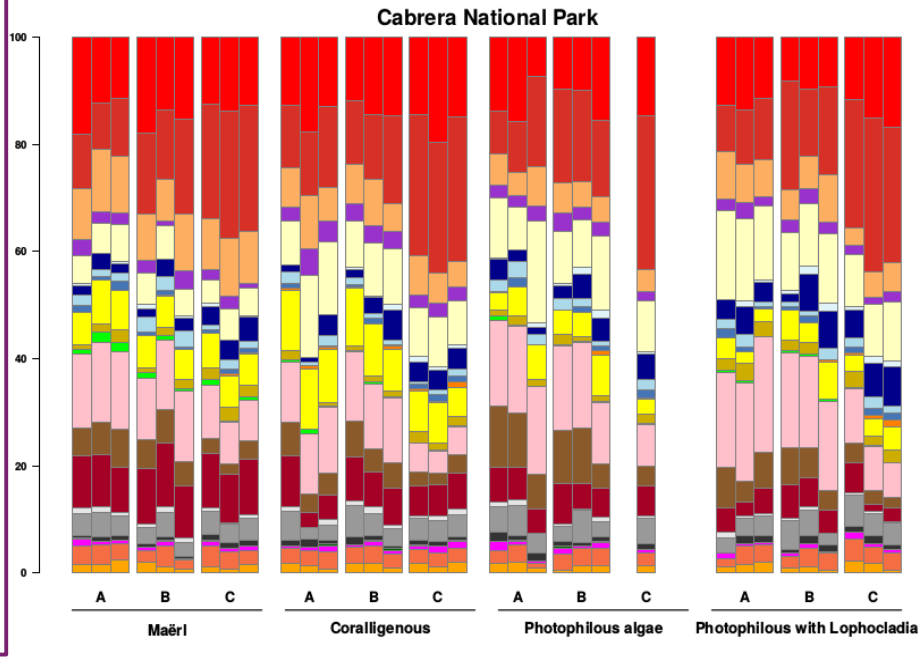
PRESENTING RESULTS: DIVERSITY (% MOTUS)

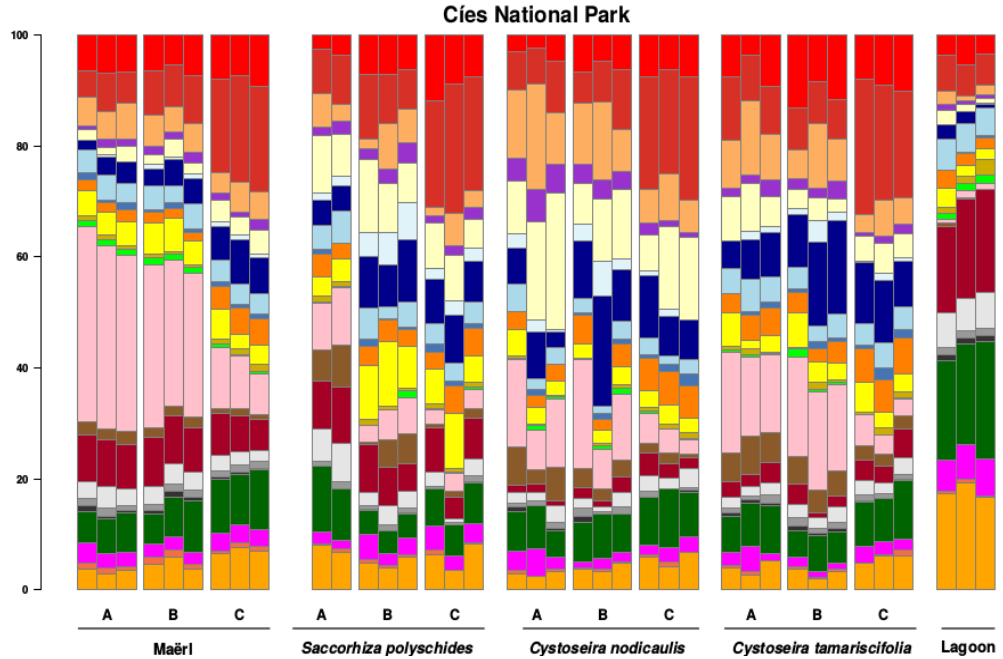
18S % Assigned MOTUs

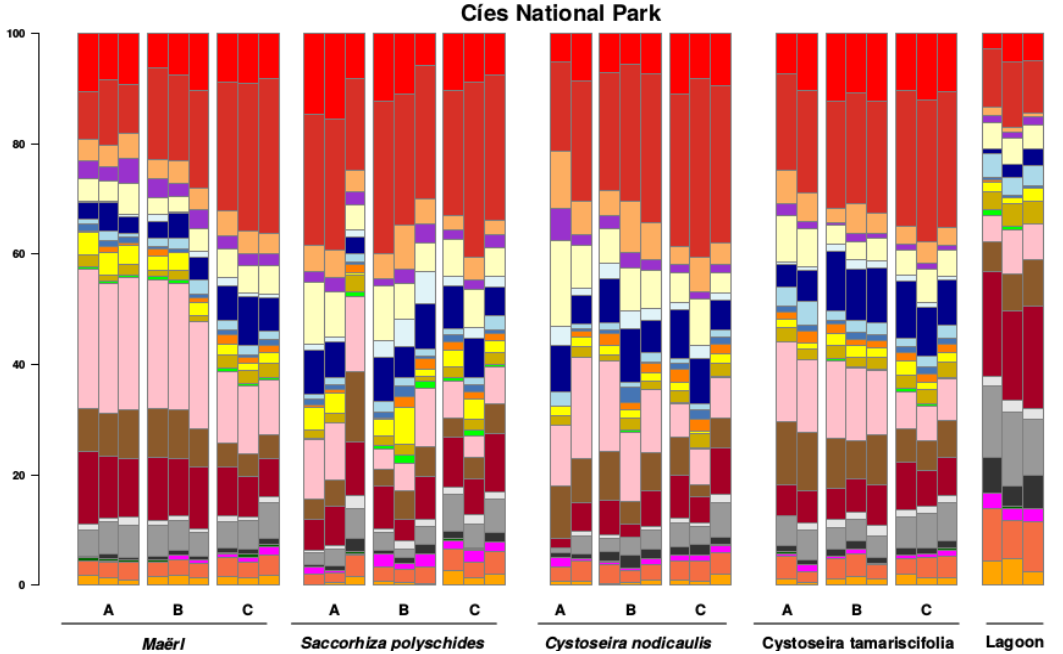




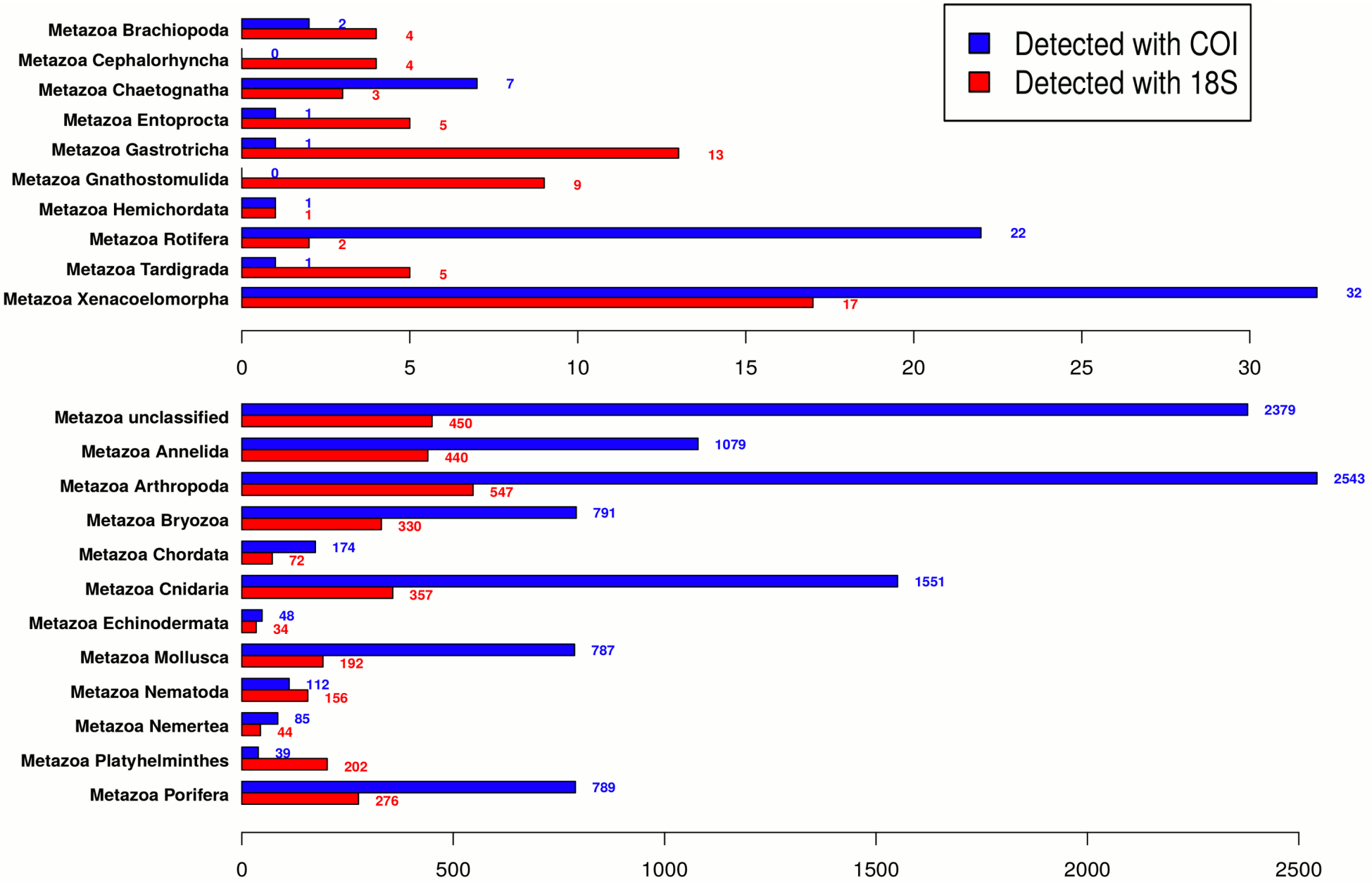
COI % Assigned MOTUs





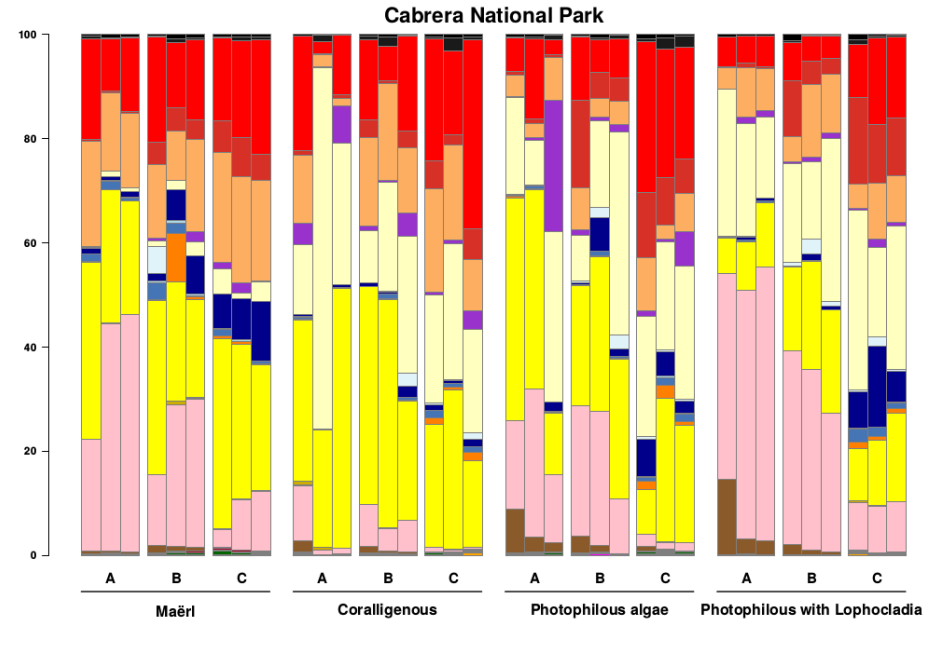


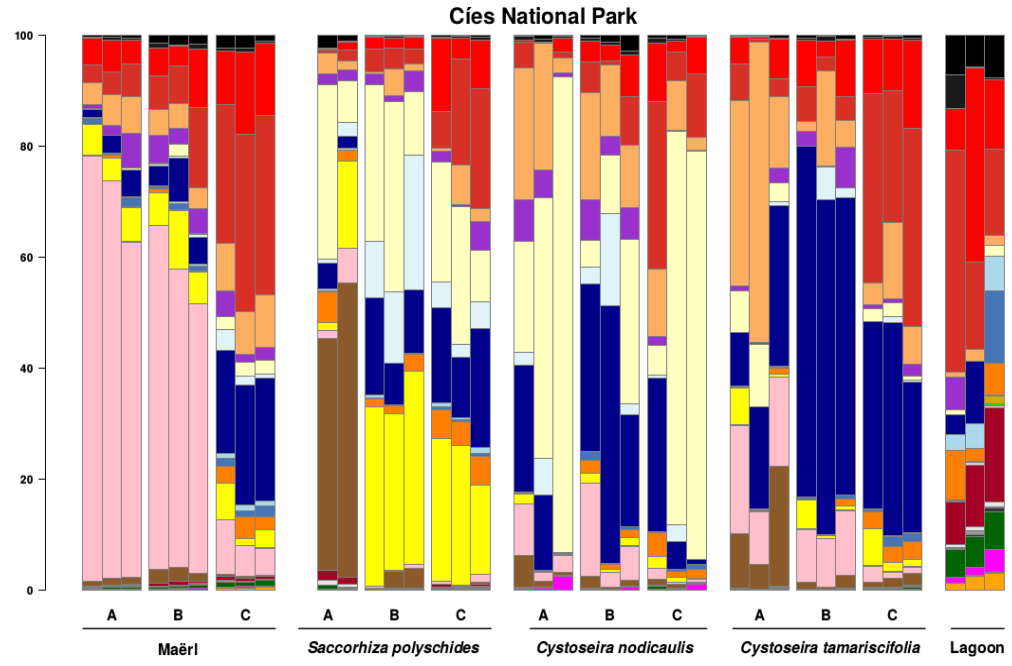
DIVERSITY: COMPARING TWO MARKERS

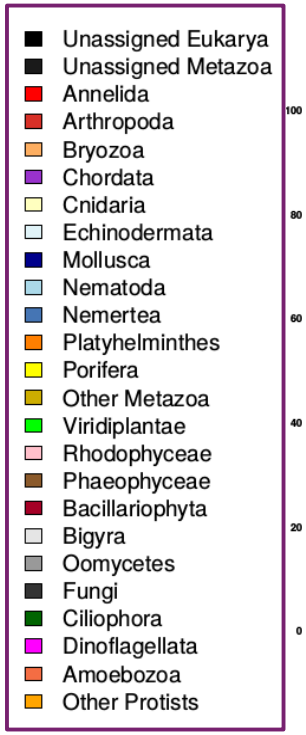


PRESENTING RESULTS: BIOMASS (% READS)

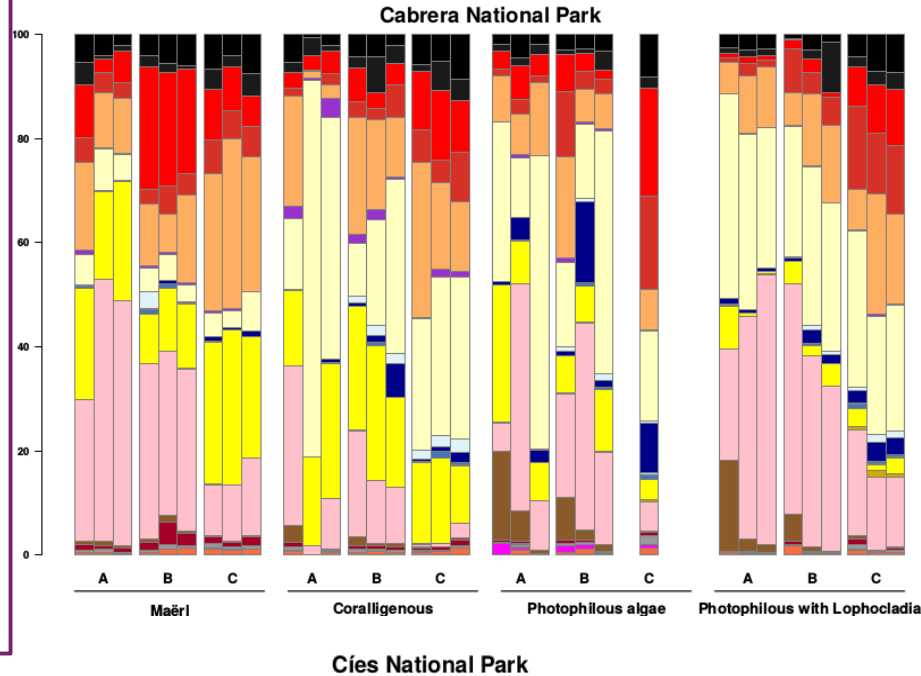
18S % Total reads

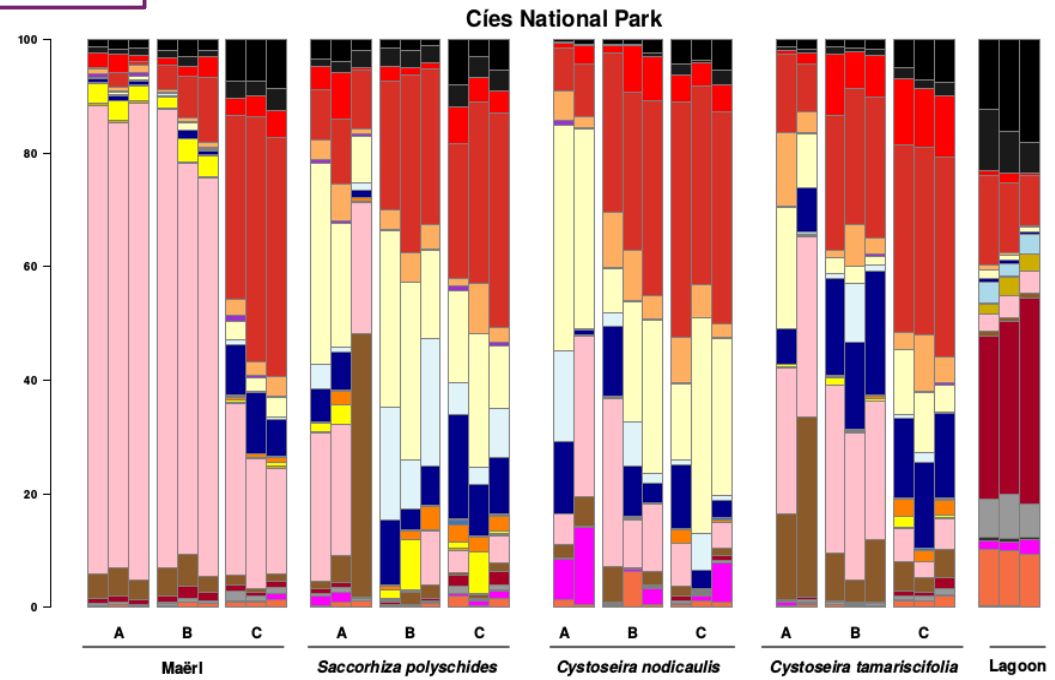






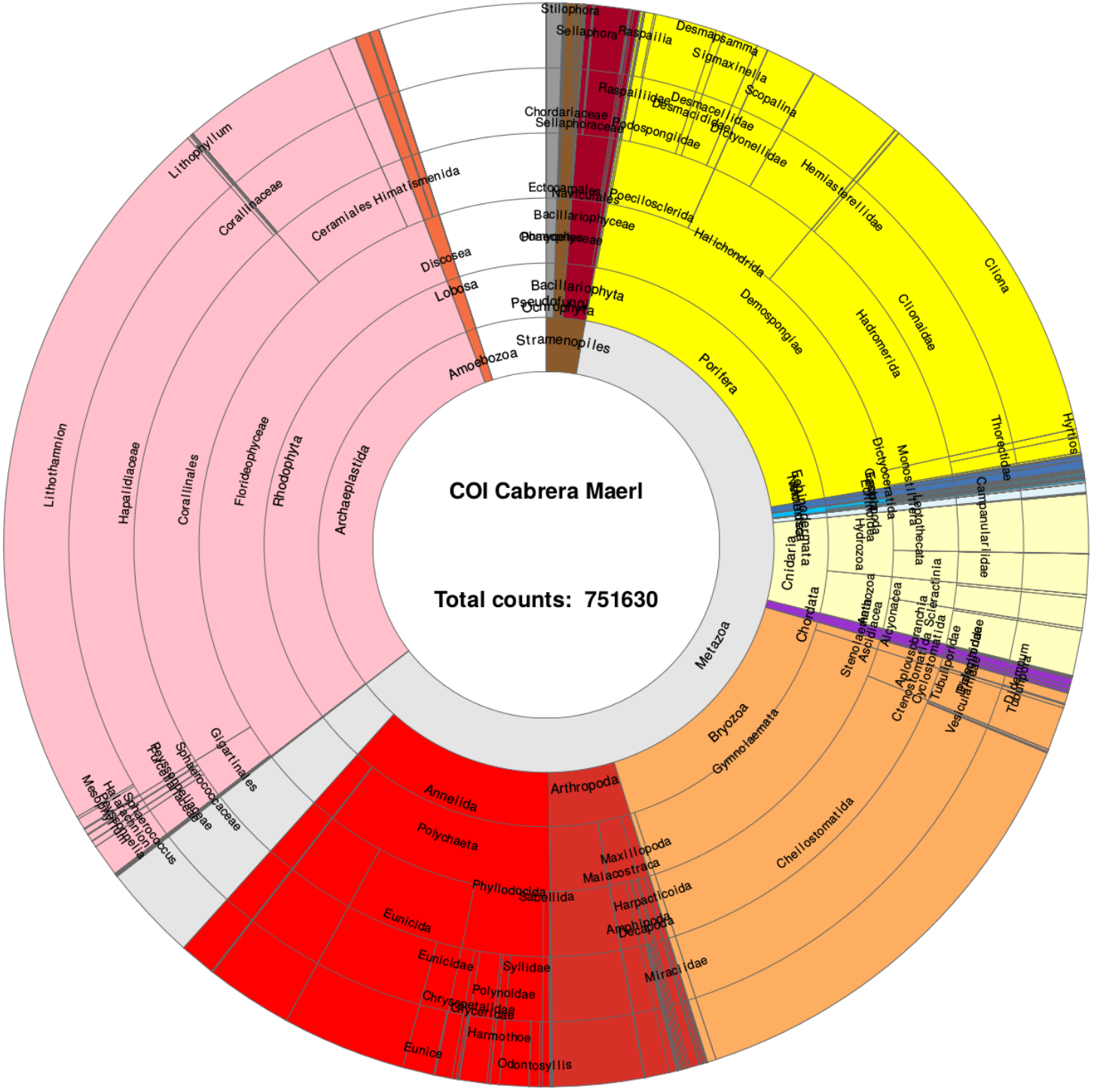
COI % Total reads



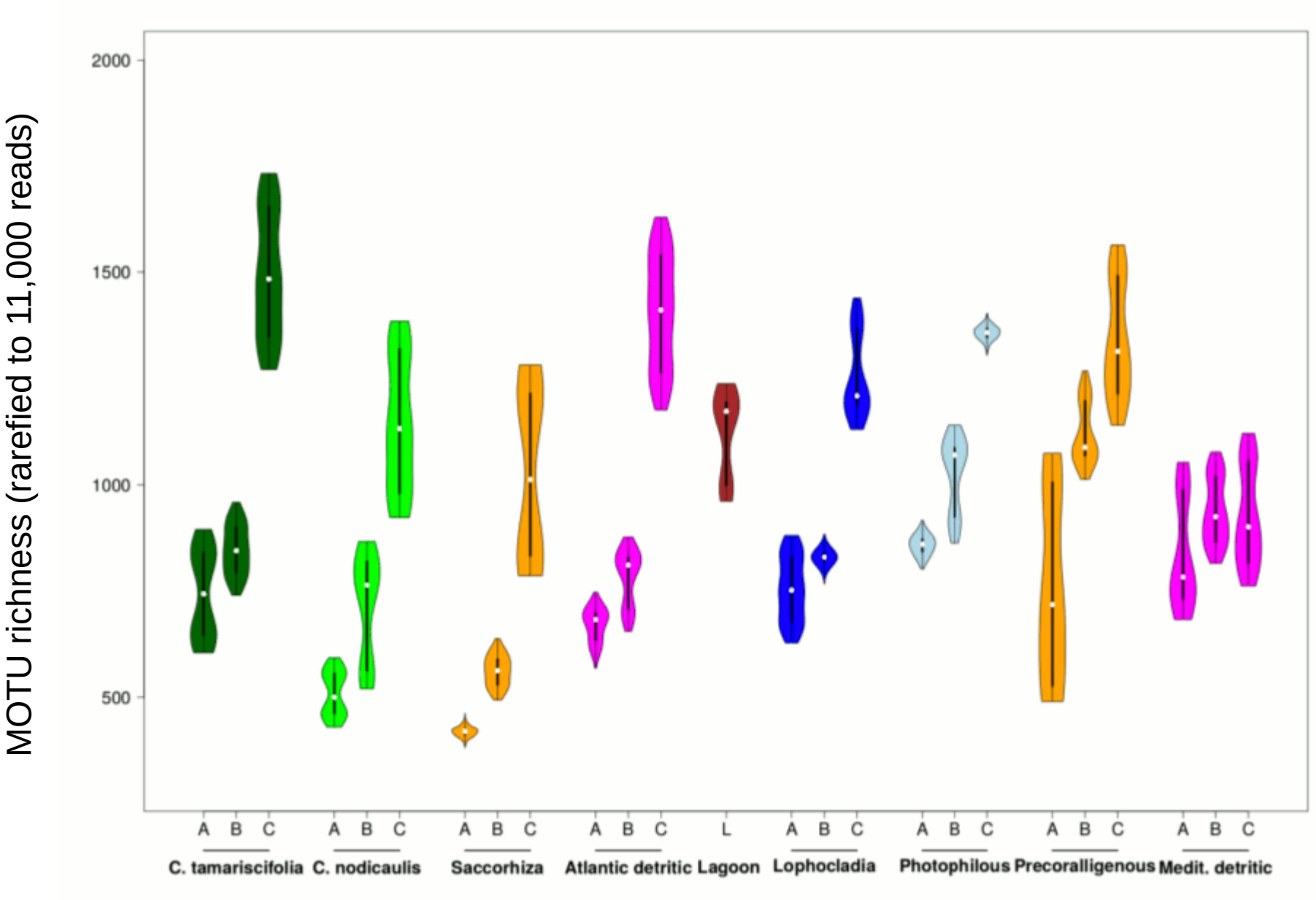


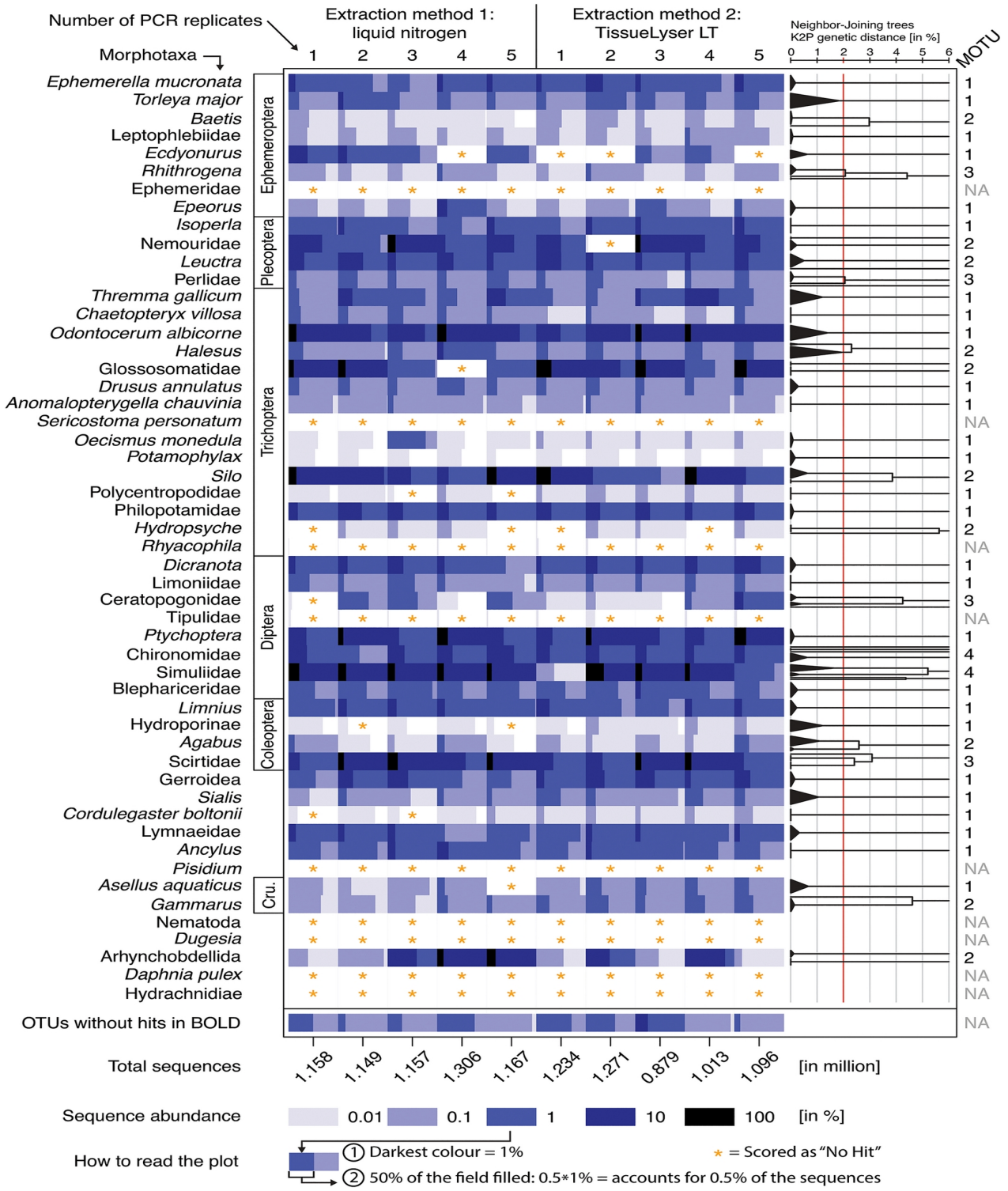
TAXONOMIC SUMMARY: KRONA PLOTS

Cabrera Maërl COI



RESULTS: α-DIVERSITY (VIOLIN PLOTS)

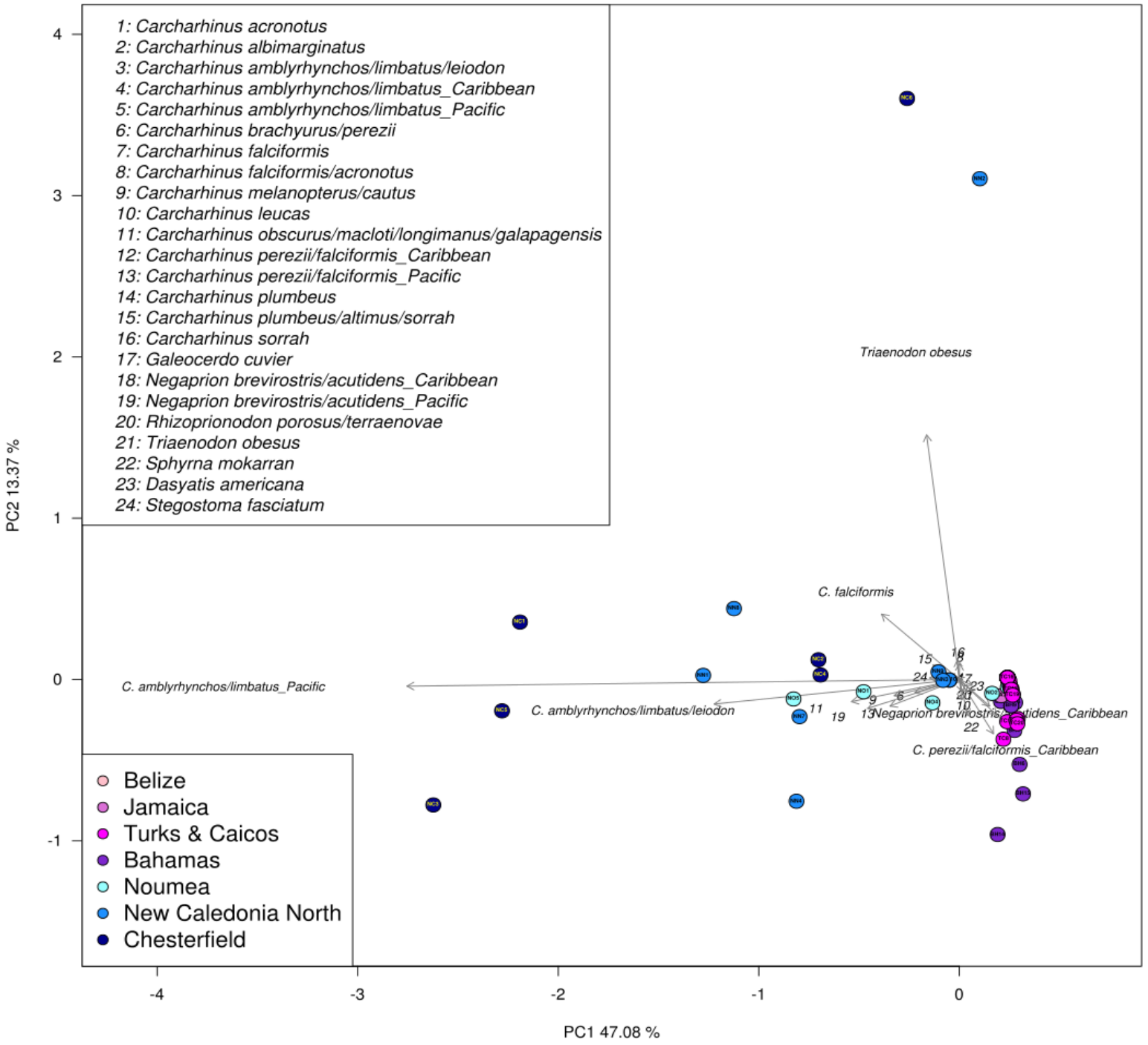




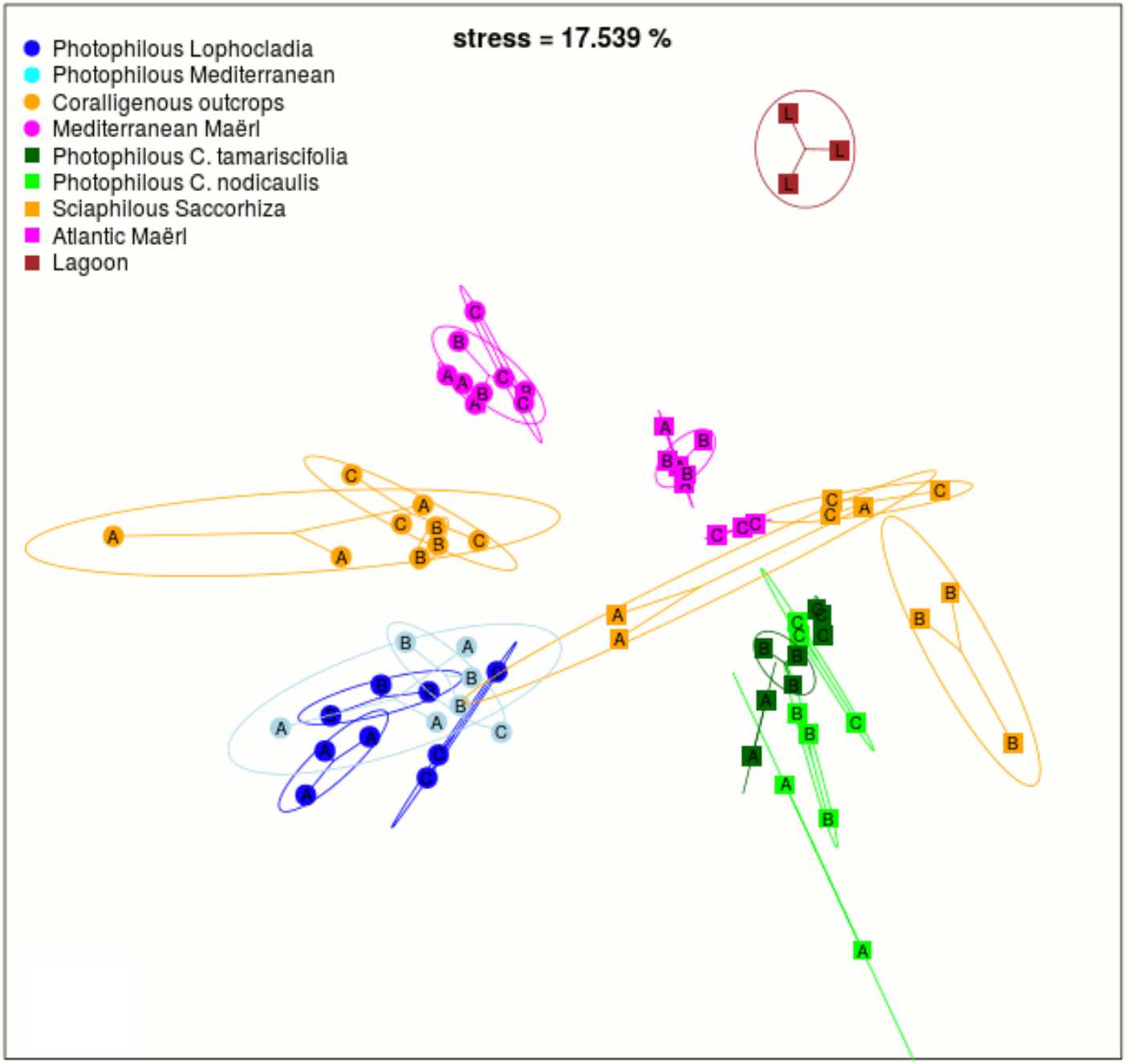
RESULTS: β-DIVERSITY (HEAT MAPS)

Elbrecht & Leese 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? *PloS ONE* 10, e130324

RESULTS: ORDINATION METHODS: PCA



RESULTS: ORDINATION METHODS
NON-METRIC MULTIDIMENSIONAL SCALING



ECOLOGICAL INDICES CALCULATIONS

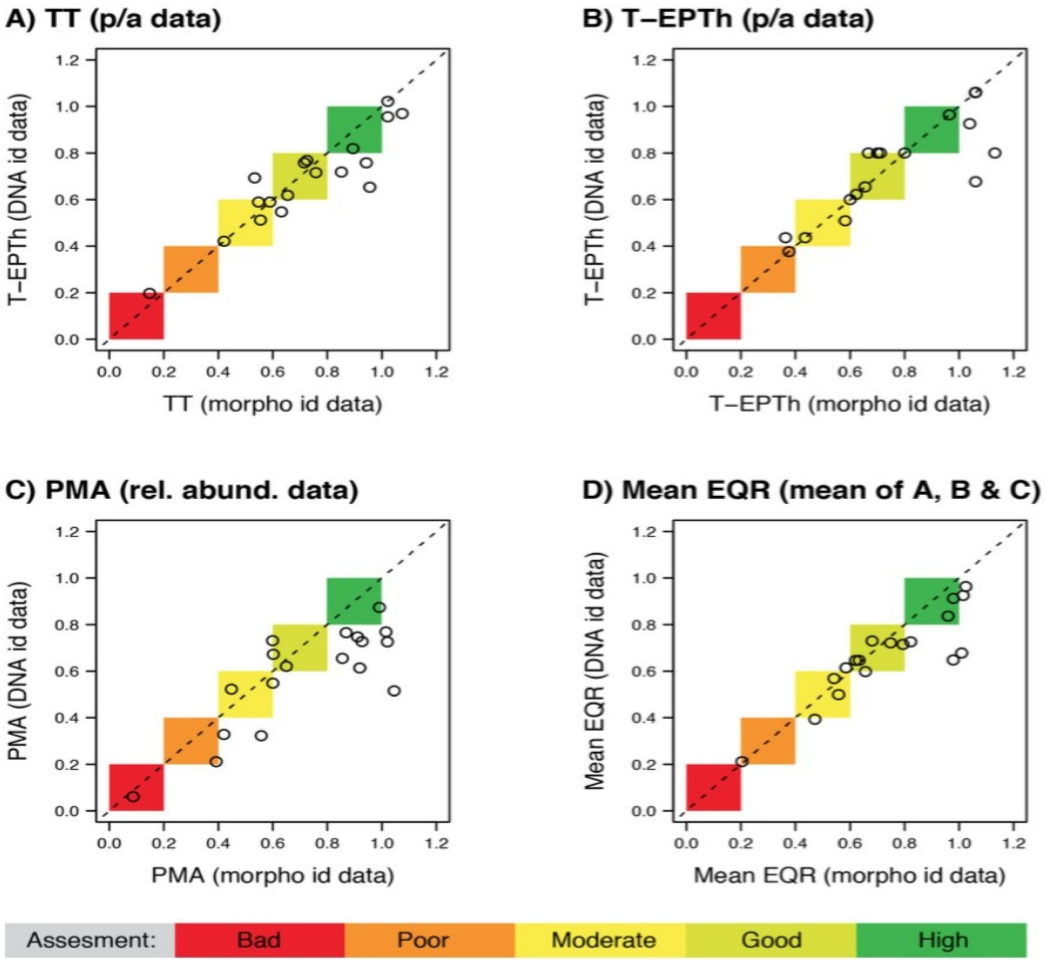


Figure 3: Comparison of Finnish macroinvertebrate WFD assessment indices calculated with taxa lists based on morphological- and DNA-based (BF2+BR2 primer) identification. The three indices are shown as normalized Ecological Quality Ratios (EQR) ranging from 0 (Bad status) to 1 (High status with no anthropogenic 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. Assessing strengths and weaknesses of DNA metabarcoding based macroinvertebrate identification for routine stream monitoring. *PeerJ.* In review.


THANK YOU!