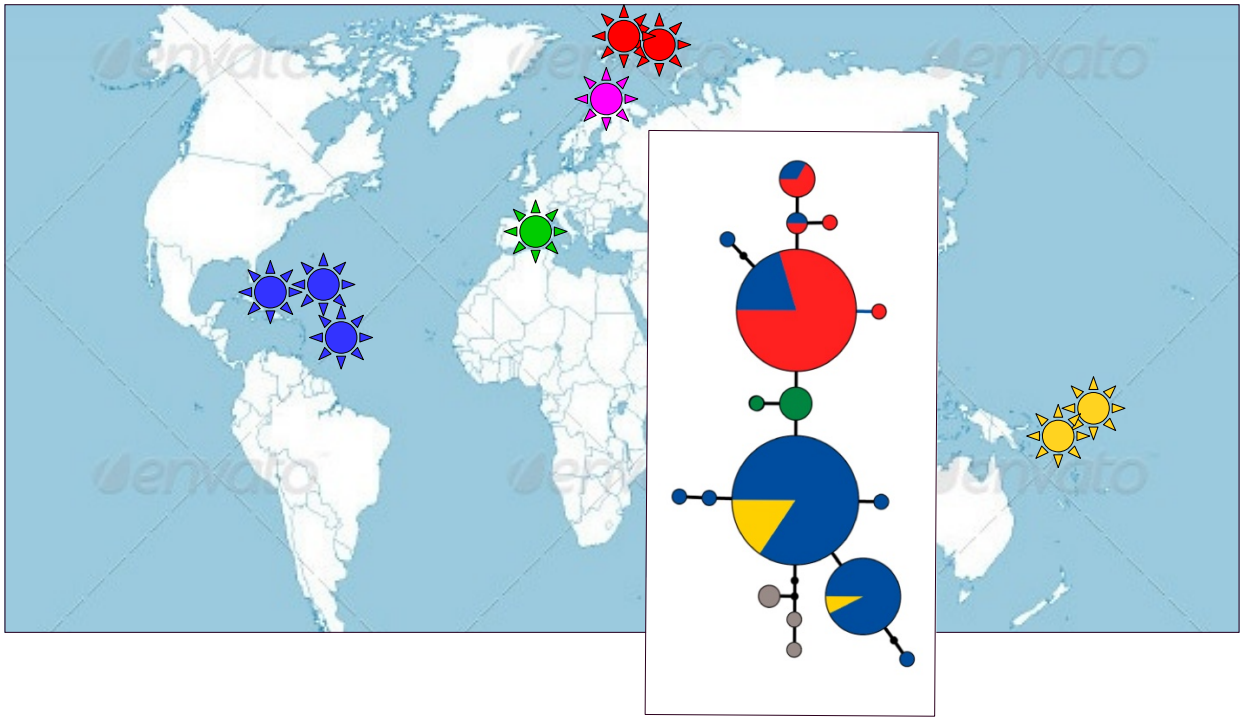
Methods to retrieve intra-species diversity information from (COI) metabarcoding data



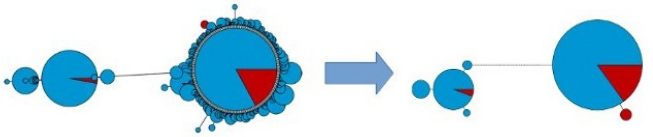
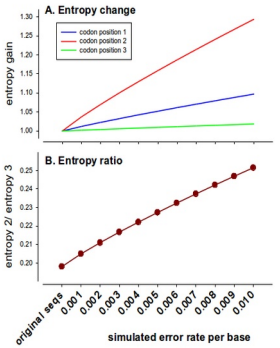
@owenwangensteen
@UiTGenetics

Owen S. Wangensteen, Adrià Antich, Creu Palacín, Kim Præbel, Xavier Turon

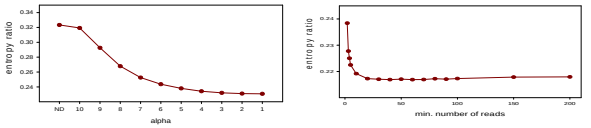
Outline of this talk

1. Natural variability, intra-species diversity and taxonomic resolution: Not all markers are created equal!

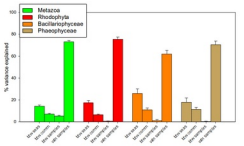

2. The origin of sequence diversity in metabarcoding datasets: Natural variability vs errors and artifacts Clustering vs denoising algorithms

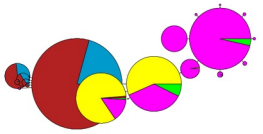



3. The codon entropy ratio as a method to optimise denoising algorithms applied to coding sequences

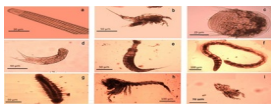
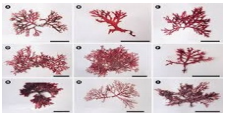


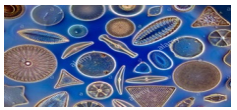
4. (Clustering OR denoising) vs (clustering AND denoising). Clustering first or denoising first?



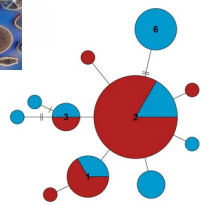


5. DnoisE: a new algorithm based on codon entropy information for denoising of coding sequences and the MJOLNIR pipeline



6. Some results: haplotype networks, metaphylogeographic patterns



1.

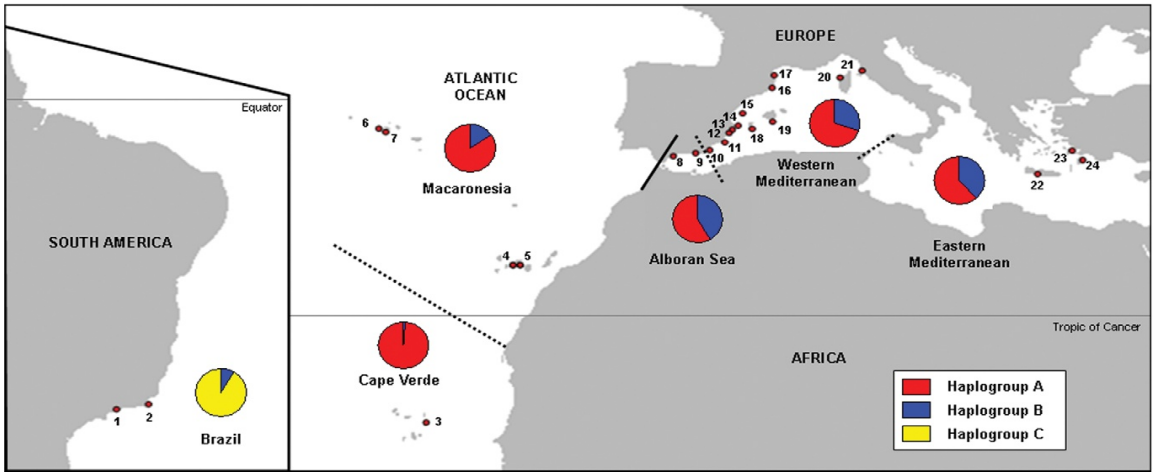
Beyond biodiversity assessment: population genetics & phylogeography

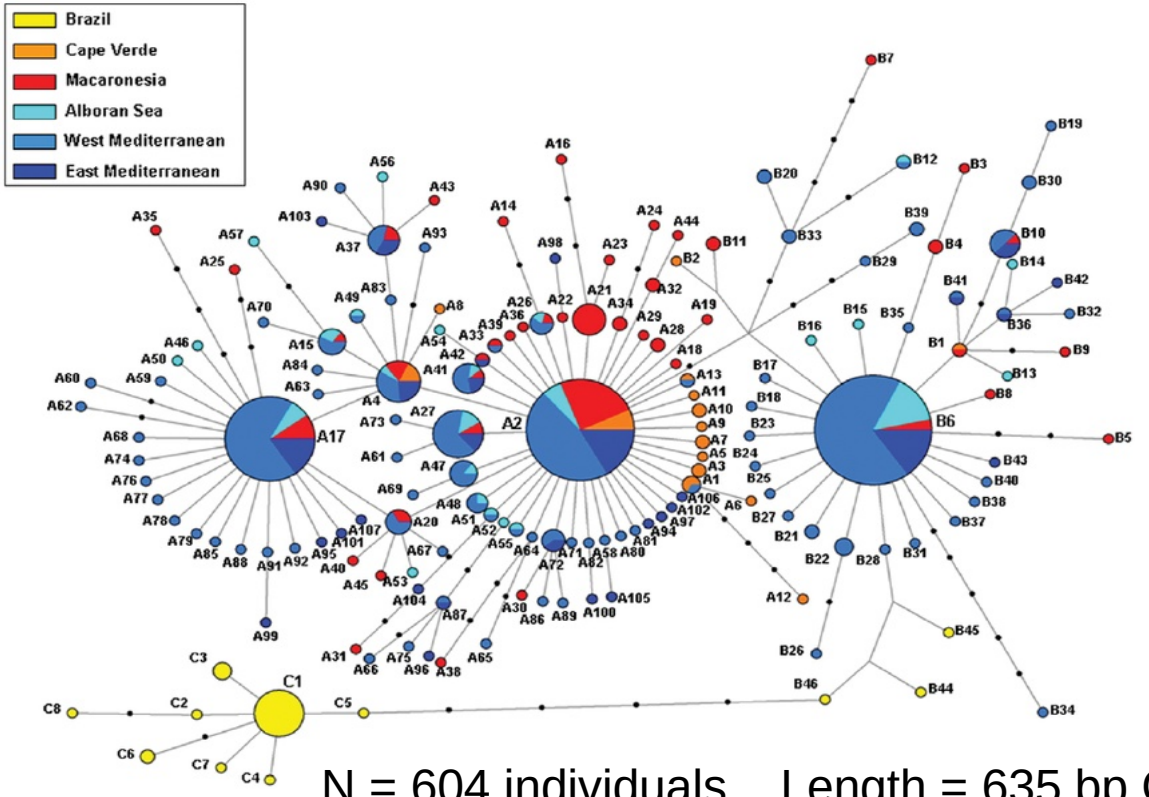
Intra-species diversity is the basis of population genetics & phylogeography

This information is crucial in conservation biology! Defining conservation units, assessing barriers to gene flow, fisheries management, design of protected areas, etc…

These studies are currently performed one species at a time, by genotyping many individuals of the same species

Small-sized organisms (meiofauna, plankton, epibionts, microeukaryota) are typically left out



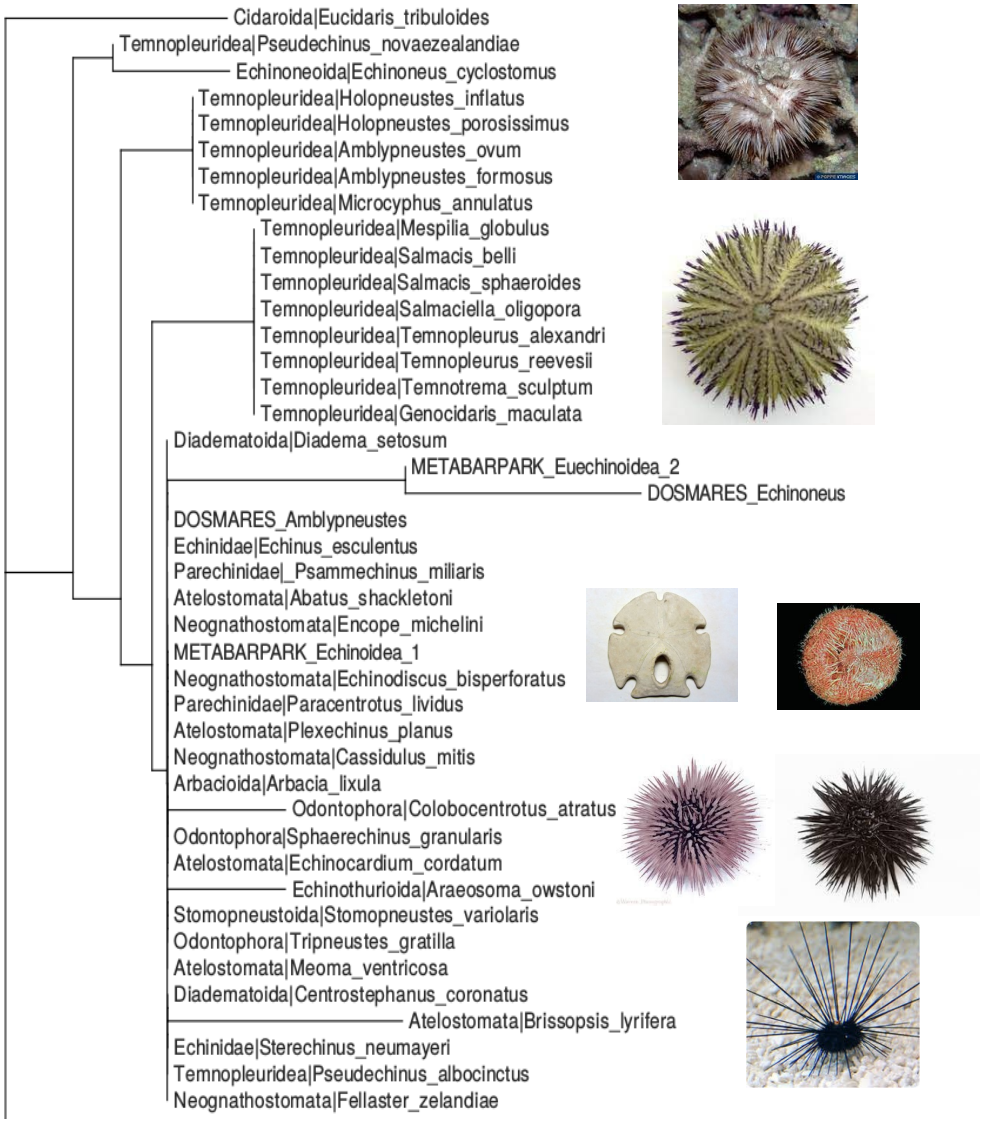


N = 604 individuals Length = 635 bp COI

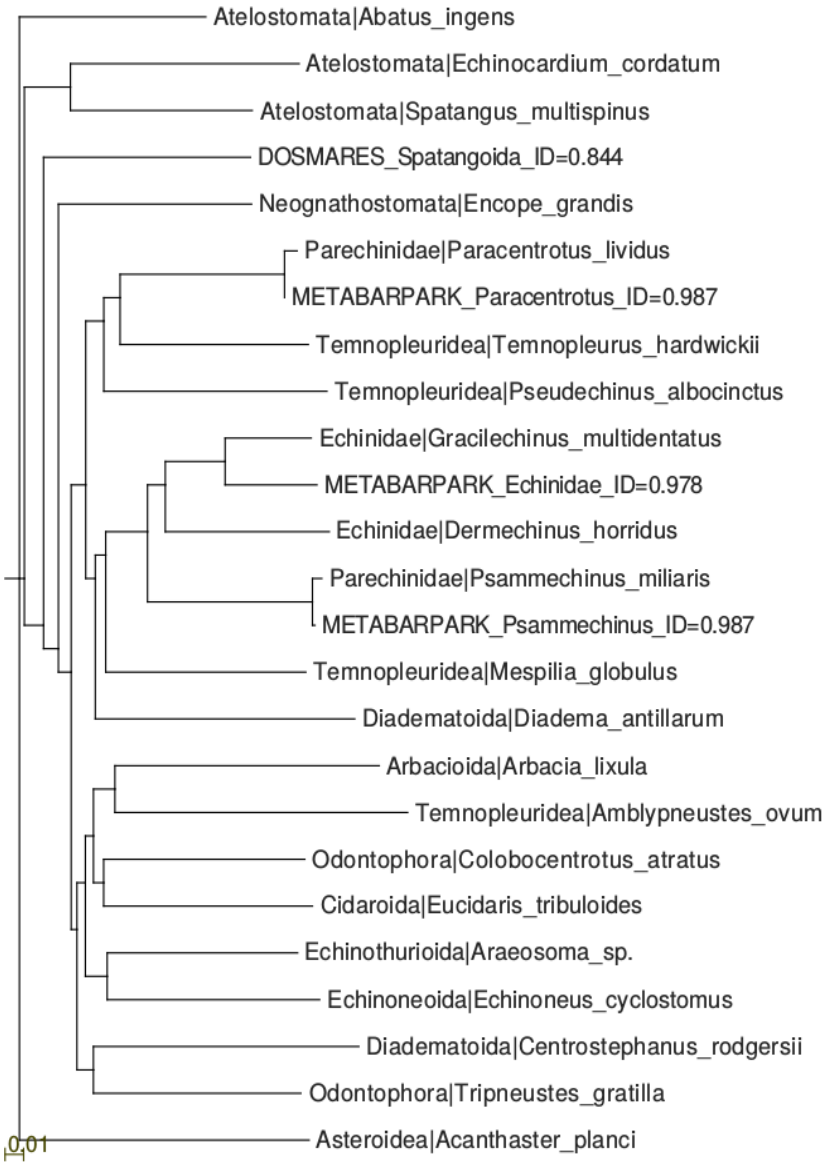
1.

Natural variability of a marker and taxonomic resolution

Echinoidea with 18S V7
(primers Guardiola et al., 2015)



Echinoidea with COI
(primers Wangensteen et al., 2018)



1.

Inter-species vs intra-species variability

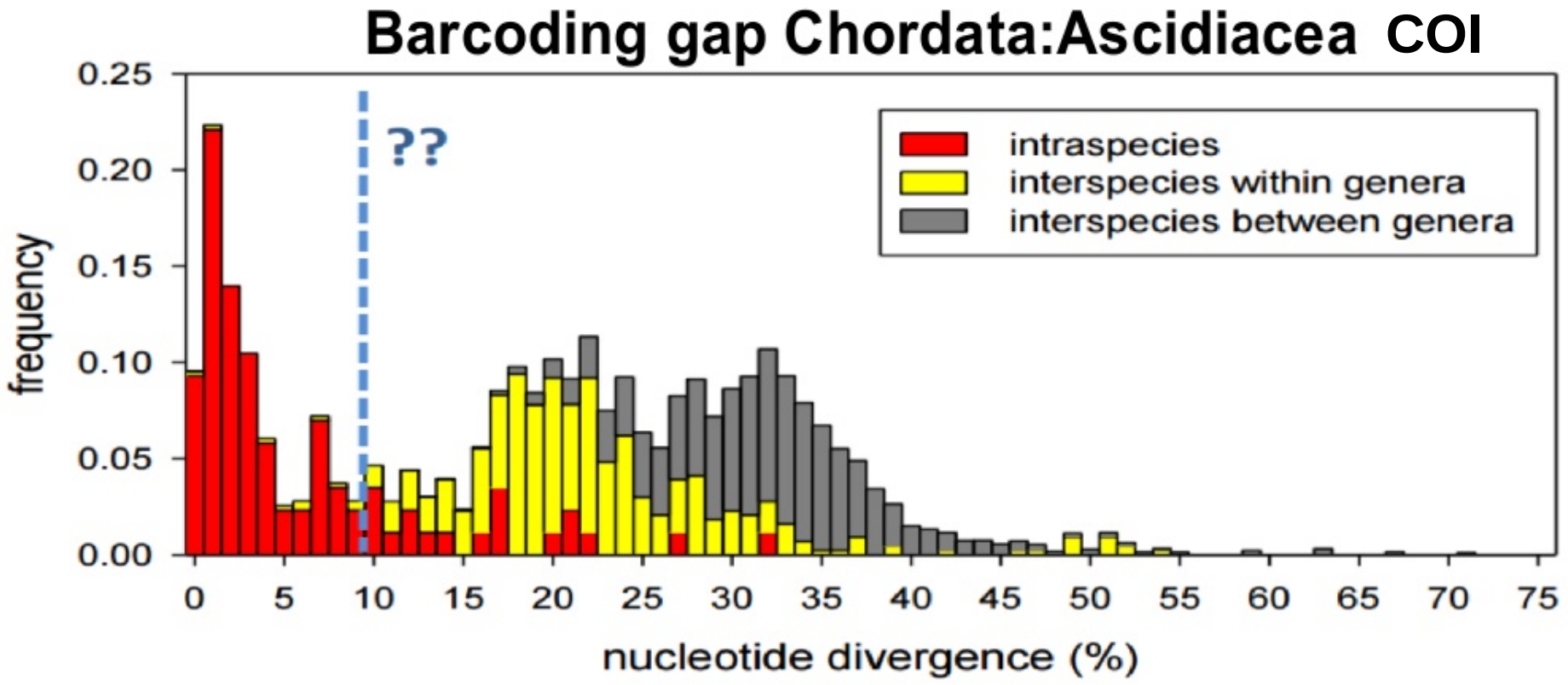
18S is a low-resolution marker.
Not very useful for distinguishing species
In the case of Echinoidea, not even families or orders!

Many species ➡ one 18S sequence

COI is very good for discriminating among species

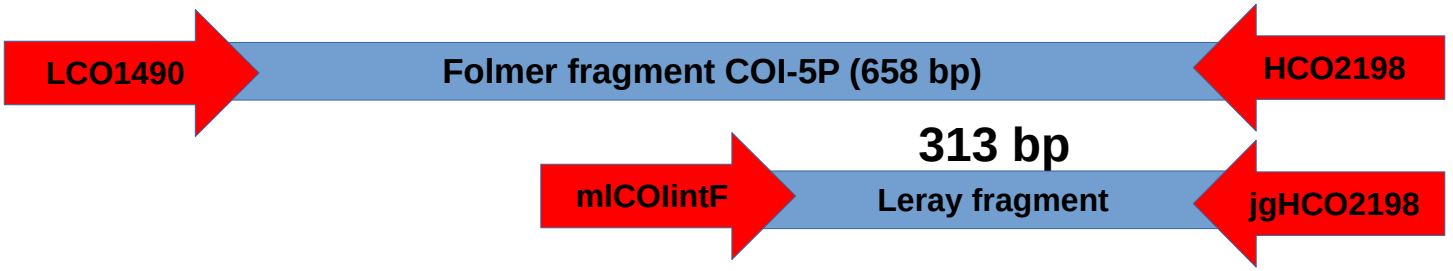
Can we also use COI data to retrieve intra-species diversity?
Haplotype information?

1 species ➡ many COI sequences



1.

We can do COI community-DNA metabarcoding using universal primers!



Fragment introduced by M. Leray et al. 2013 (most metazoa)
Modified Leray-XT primer set works well in most marine Eukarya
(Wangensteen et al. *PeerJ* 2018)

miCOIintF-XT: 5'-GGWACWRGWTGRACWITITAYCCYCC-3'

jgHCO2198: 5'-TAIACYTCIGGRTGICCRAARAAYCA-3'

Hypothesis: COI hypervariability will allow us to study intra-specific diversity (haplotypes)

2.

CASE STUDY: PROJECT METABARPARK

Cíes Islands

C. nodicaulis

C. tamariscifolia

Saccorhiza

Atlantic Maërl

Lophocladia

Photophilous

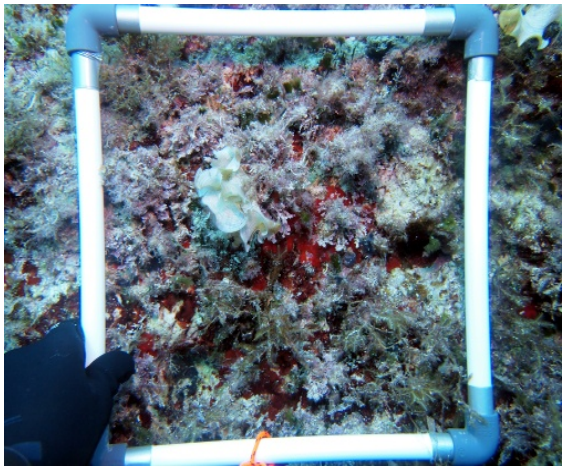
Precoralligenous

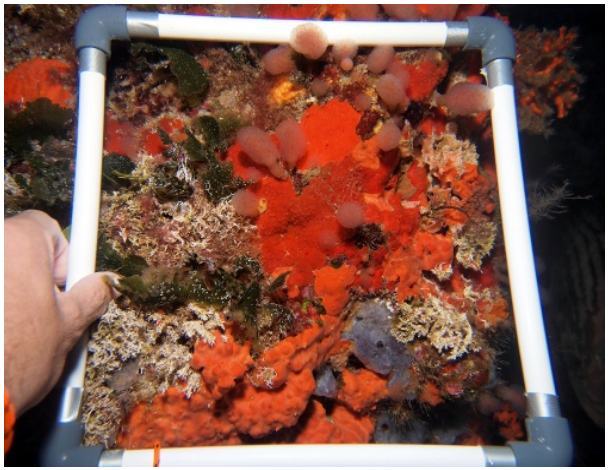
Mediterr. Maërl

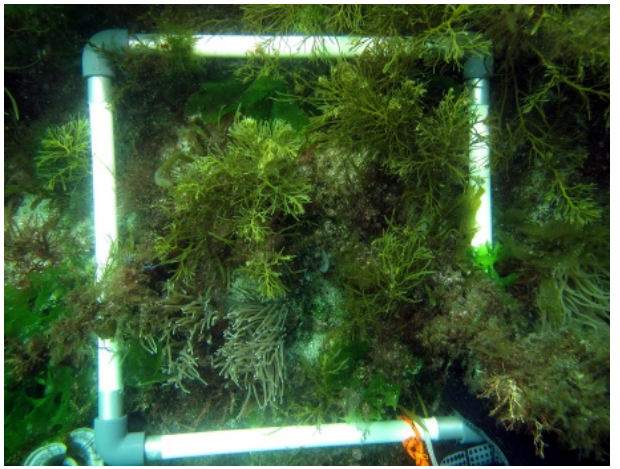
Cabrera Archipelago

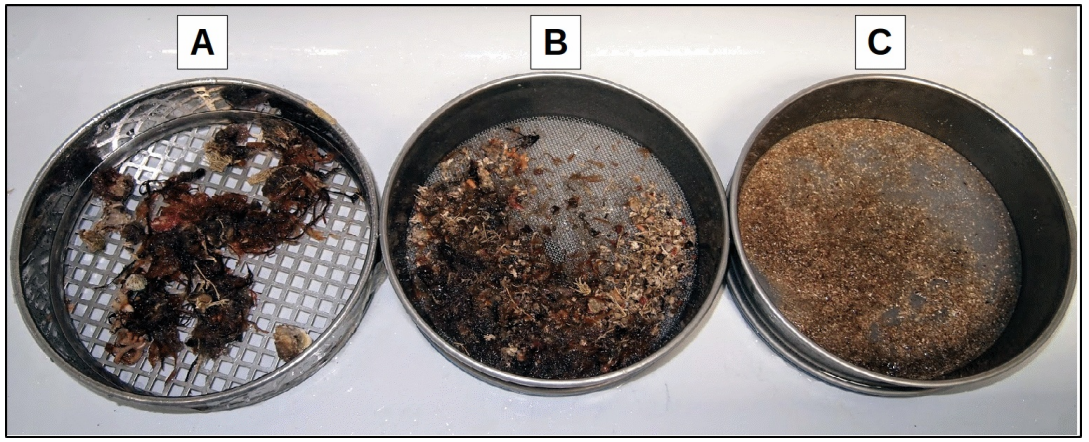
2.

SAMPLING AND METABARCODING









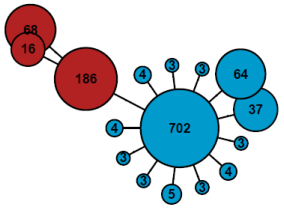
2 Parks x 4 Communities x 3 Replicates x 3 Fractions x 2 years = **144 samples**

Amplified with COI Leray-XT primer set, 313 bp (Wangensteen et al. 2018)

25 millions raw reads → **4 millions of unique sequences!!!**

2.

Where this extreme sequence diversity comes from?



Natural (real) diversity:

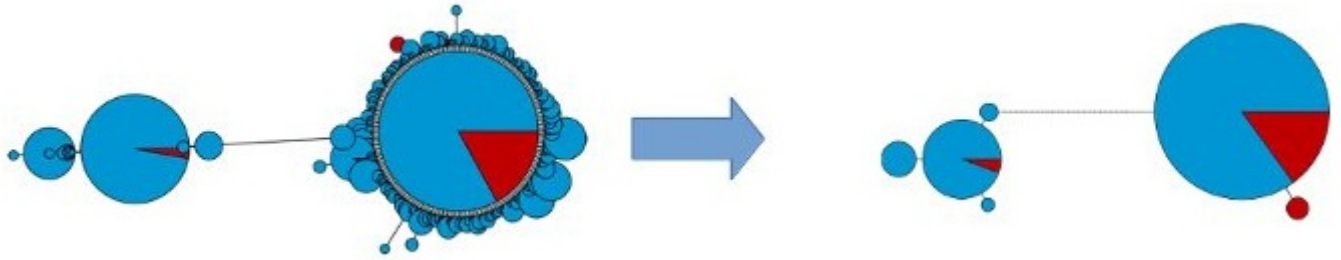
- Inter-species
- Intra-species
- Intra-individual

Artifacts and errors:

- PCR amplification errors
- Chimaeric sequences
- Sequencing errors (typical ER:1%)

To retrieve the true intra-species diversity information we must remove the errors

Raw sequence variants **Denoised network**



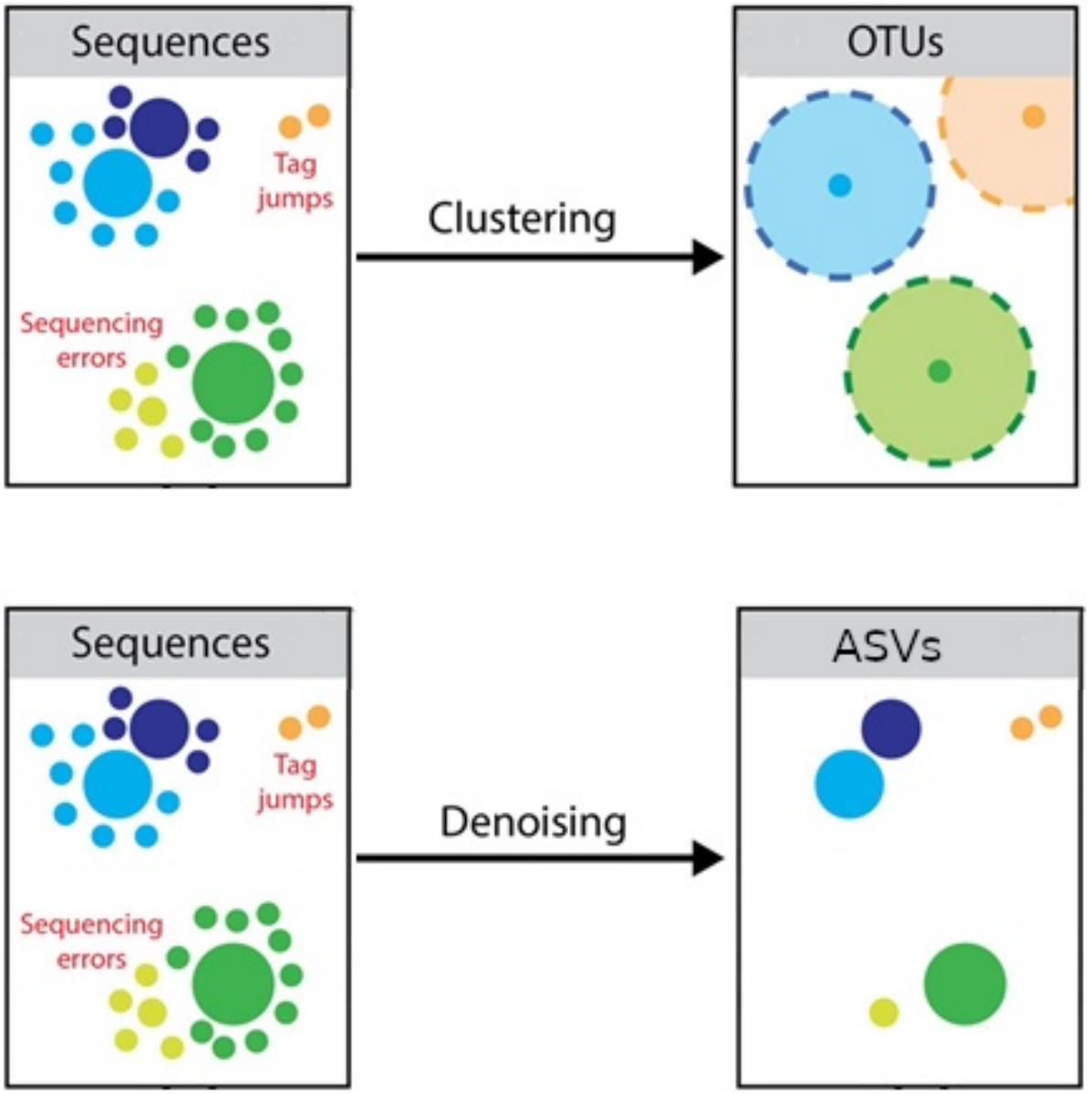
Variability due to errors is usually higher than natural variability. Many variants, but in lower abundances

How to distinguish true variability from random errors?

2.

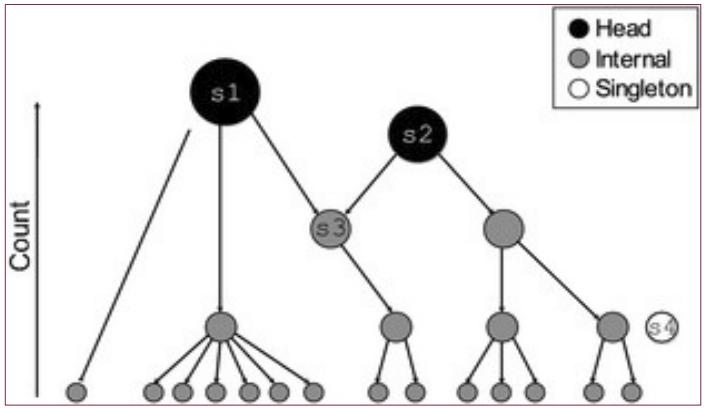
HOW TO DEAL WITH SEQUENCING ERRORS?

CLUSTERING vs DENOISING

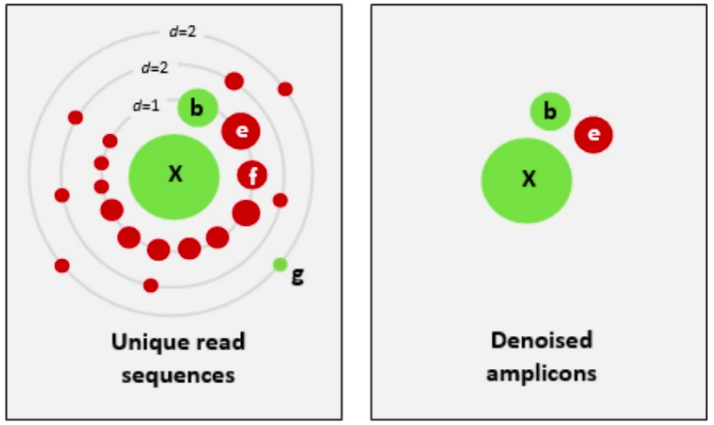


2.

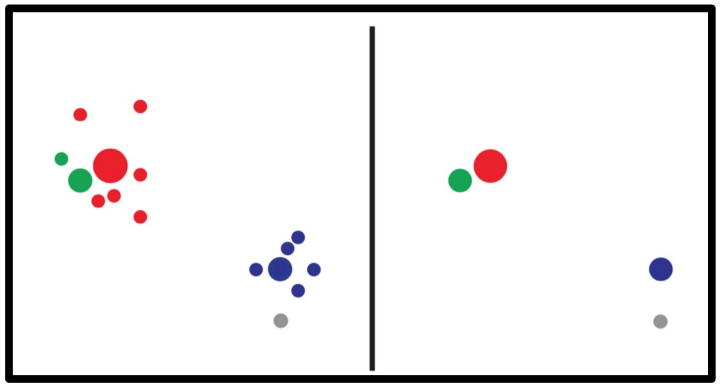
DENOISING ALGORITHMS



obiclean (OBITools) → Heads



UNOISE (USEARCH) → ZOTUs



DADA2 Amplicon Sequencing. Exactly. Version 1.14

→ ASVs

2.

DENOISING ALGORITHMS

PROBLEM: the performance of every denoising algorithm depends on a series of adjustable stringency parameters

Trade-off: removing most errors / keeping true haplotypes

These parameters have been optimized for bacterial 16S markers using prokaryotic mock communities.

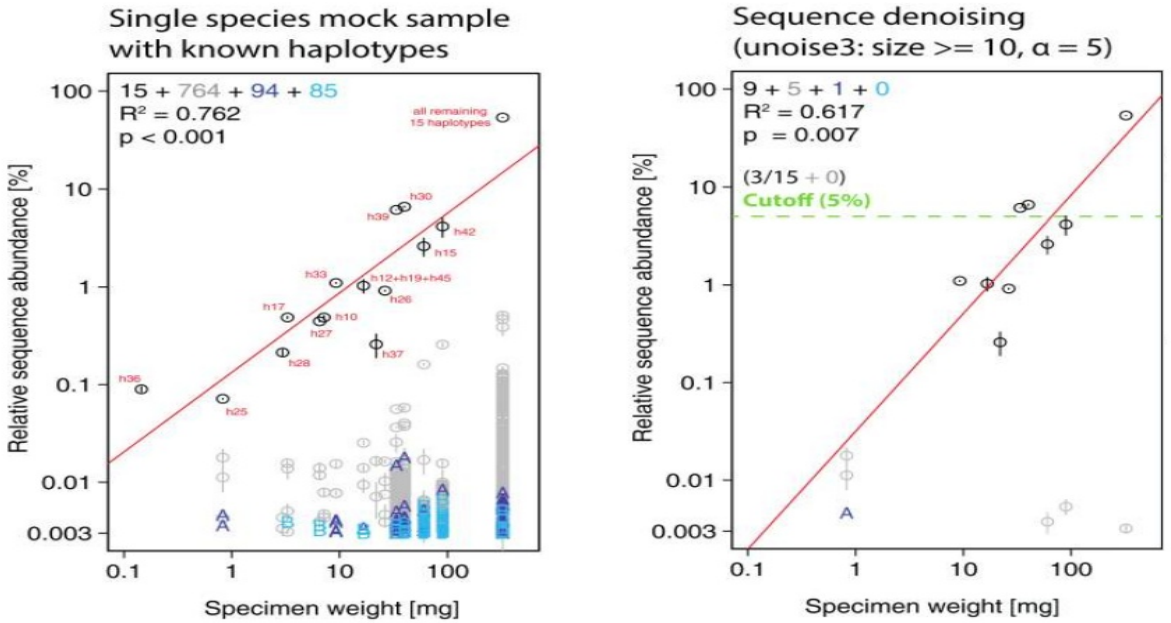
They are usually applied to other eukaryotic markers, without any further critical considerations.

- over-denoising: removing true natural variability

- under-denoising: keeping spurious ASVs coming from sequencing errors

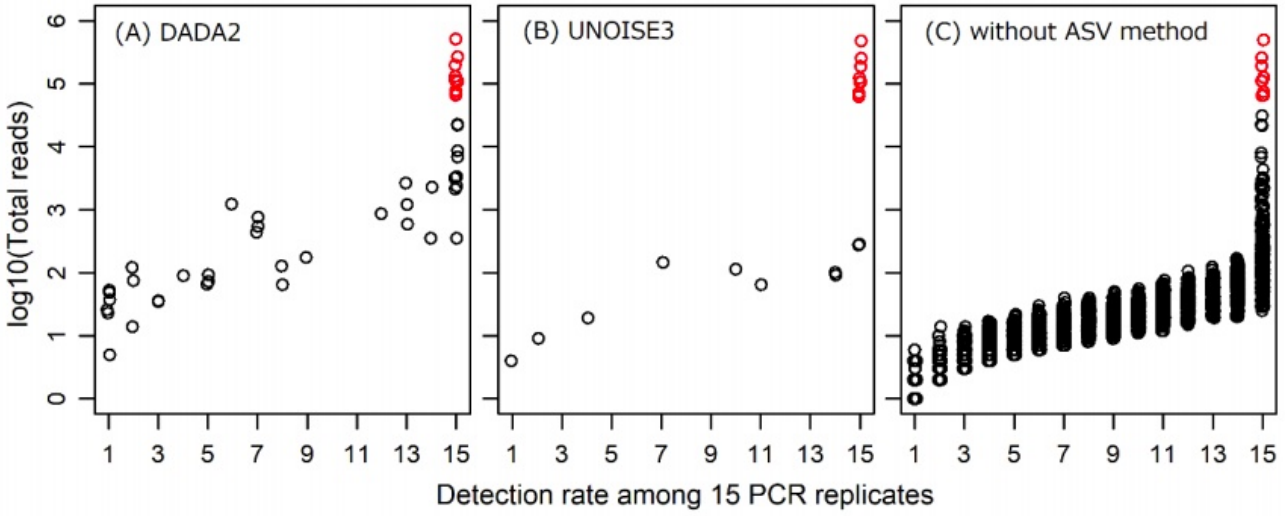
2.

EXAMPLES USING MOCK SAMPLES



over-denoising with COI (bulk-DNA)

Elbrecht et al. (2018) http://doi.org/10.7717/peerj.4644



under-denoising with mt-CR (e-DNA)

Tsuji et al. (2020) http:doi.org/10.1111/1755-0998.13200

3.

ECOLOGICAL APPLICATIONS
ECOLOGICAL SOCIETY OF AMERICA
Volume 30, Issue 2
March 2020
https://doi.org/10.1002/eap.2036

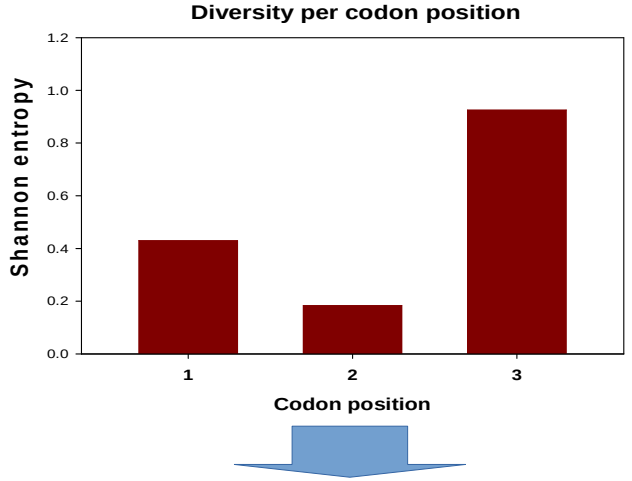
From metabarcoding to metaphylogeography: separating the wheat from the chaff

Xavier Turon, Adrià Antich, Creu Palacín, Kim Præbel, Owen Simon Wangensteen

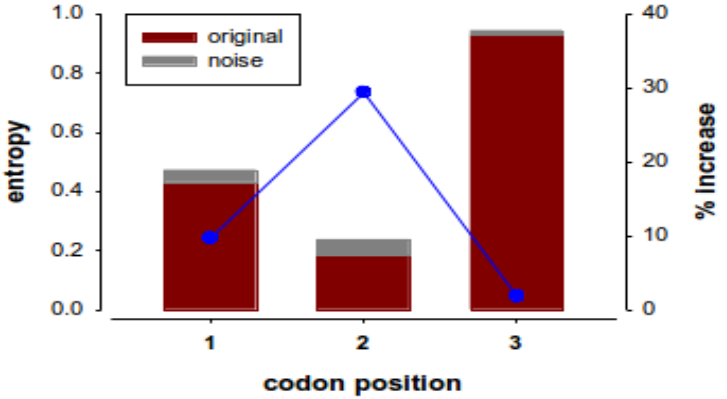
Trade-off: removing most errors / keeping true haplotypes

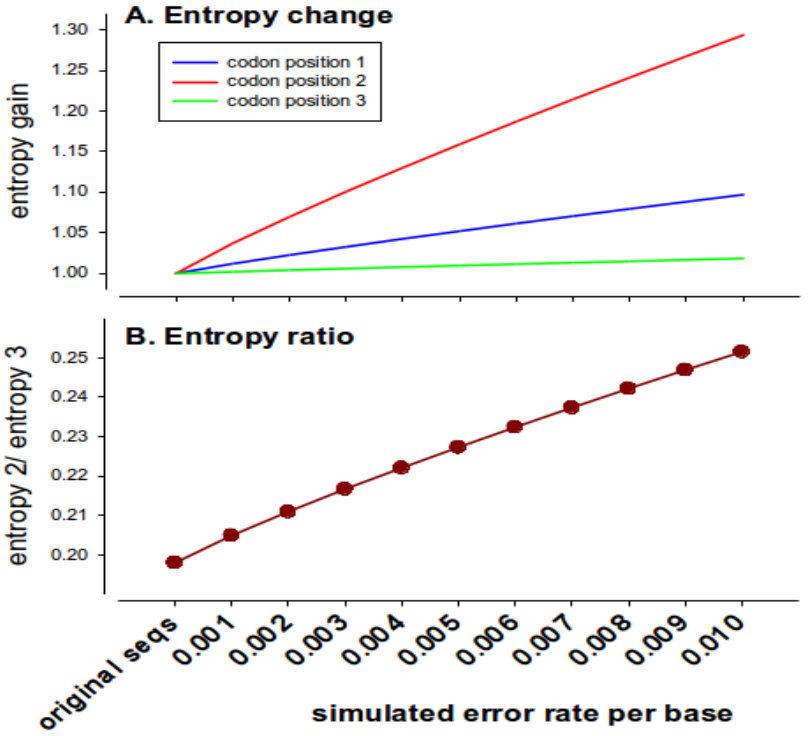
In coding sequences, entropy is differentially affected by random errors, depending on the codon position

Natural variability in COI



Natural variability + random errors





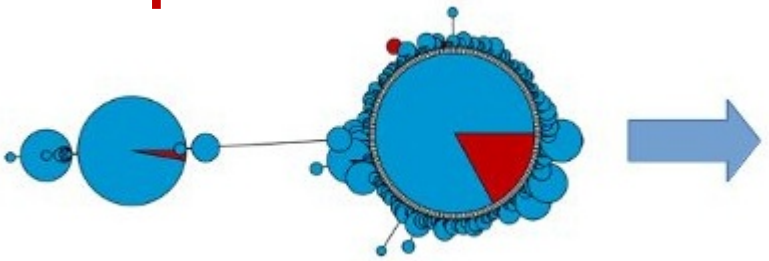
3.

ECOLOGICAL APPLICATIONS
ECOLOGICAL SOCIETY OF AMERICA
Volume 30, Issue 2
March 2020
https://doi.org/10.1002/eap.2036 |

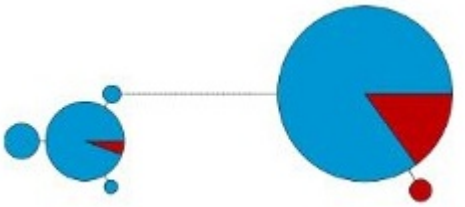
From metabarcoding to metaphylogeography: separating the wheat from the chaff

Xavier Turon ✉, Adrià Antich, Creu Palacín, Kim Præbel, Owen Simon Wangensteen

Raw sequence variants

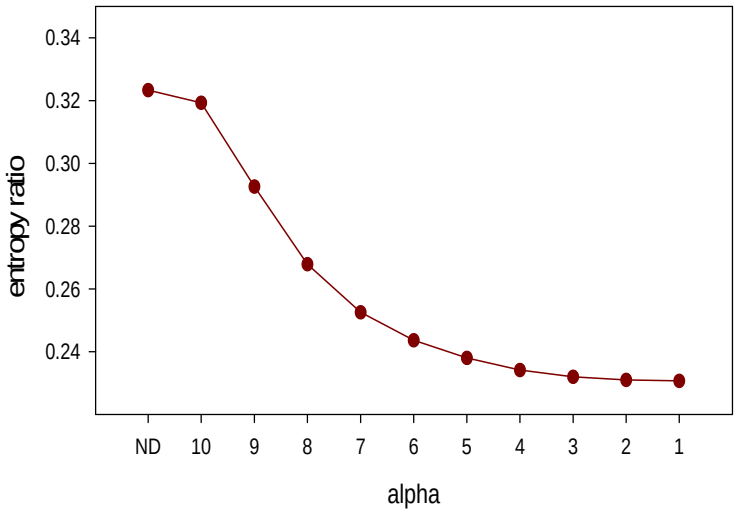


Denoised network

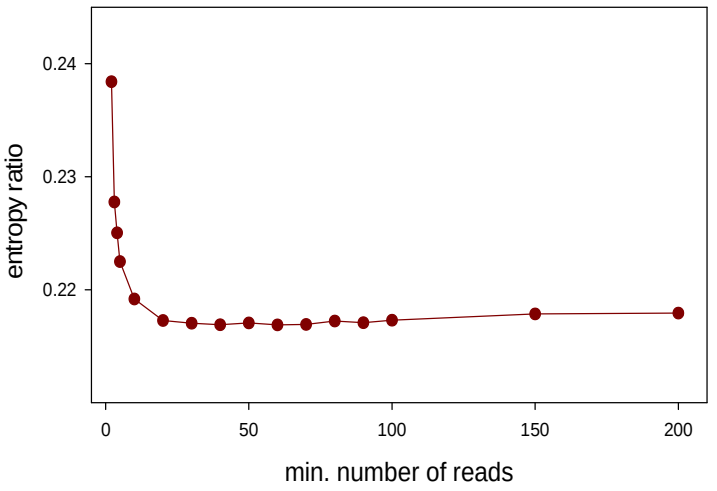


Codon Entropy Ratio (codon-3rd/codon-2nd) allows to find the right threshold for denoising & filtering

Denoising



Abundance filtering

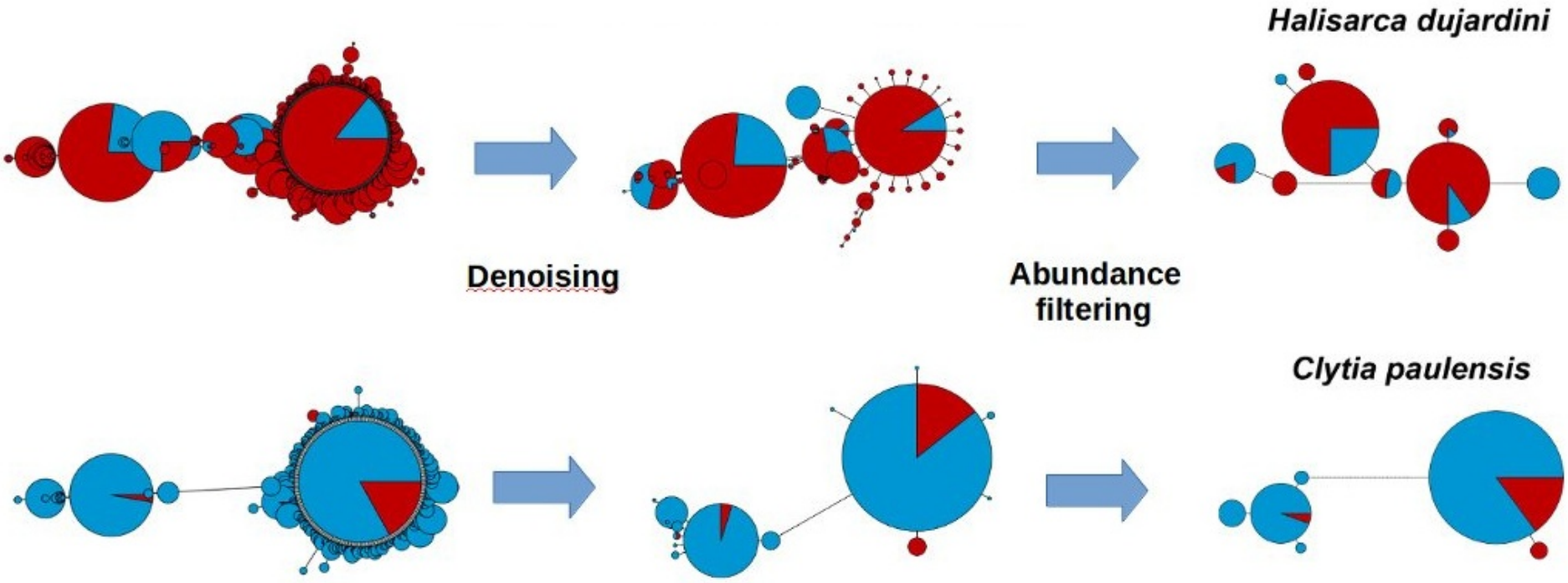


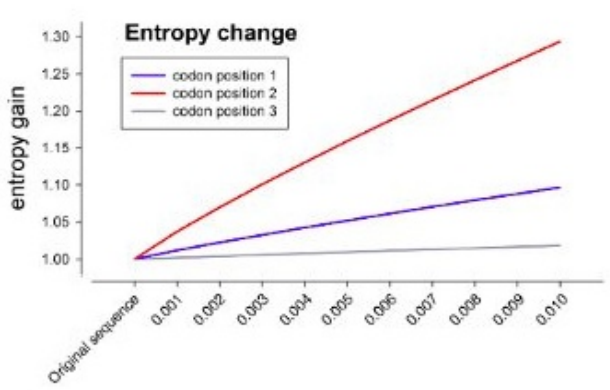
3.

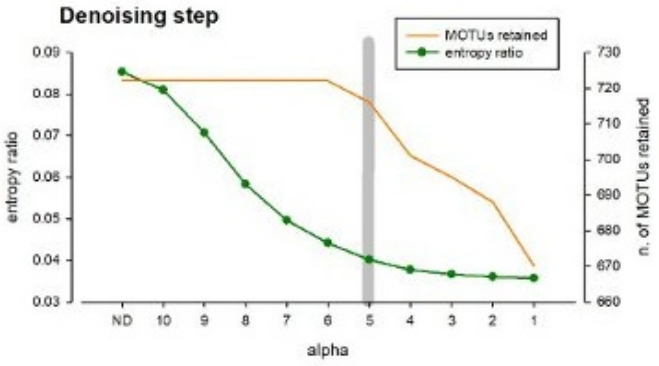
ECOLOGICAL APPLICATIONS
ECOLOGICAL SOCIETY OF AMERICA
Volume 30, Issue 2
March 2020
https://doi.org/10.1002/eap.2036 |

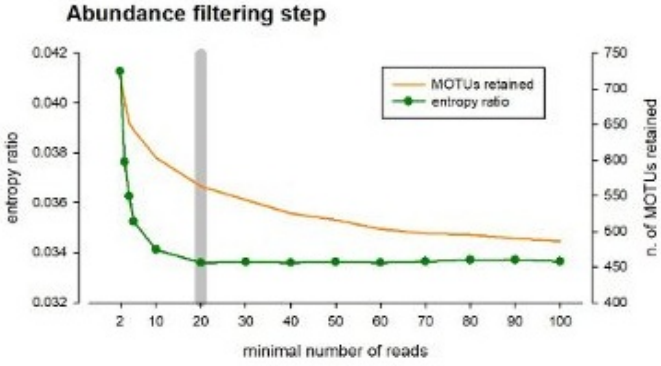
From metabarcoding to metaphylogeography: separating the wheat from the chaff

Xavier Turon ✉, Adrià Antich, Creu Palacín, Kim Præbel, Owen Simon Wangensteen







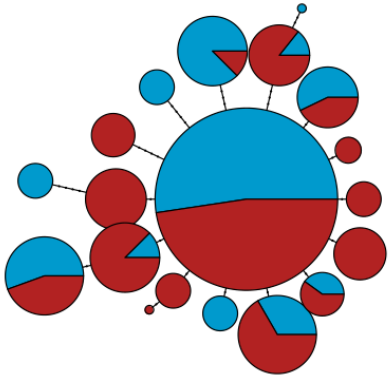


optimal alpha = 5

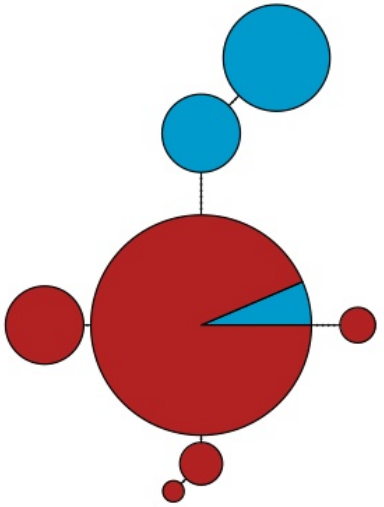
optimal minimal abundance = 20 reads

3.

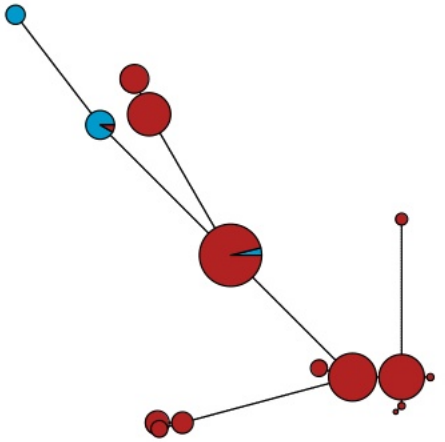
Some of the haplotype networks generated (for 444 MOTUs)



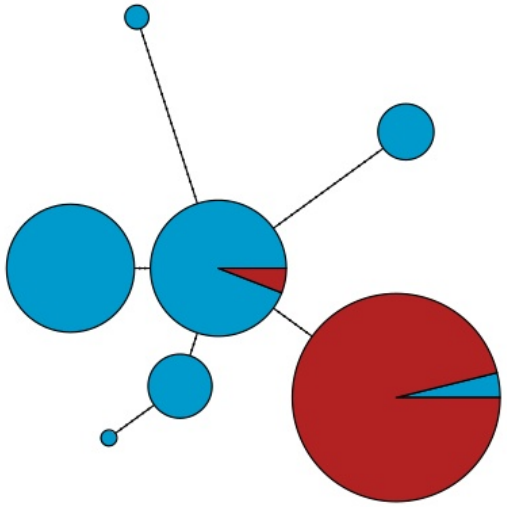
MBPA_000065793
Pusillina inconspicua 0.99
(Mollusca: Gastropoda)



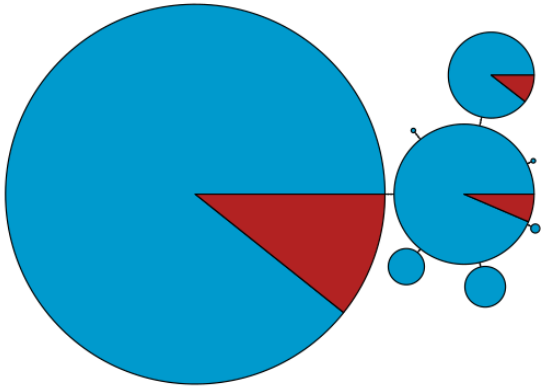
MBPA_000011152
Ord. Corallinales 0.92
(Rhodophyta)



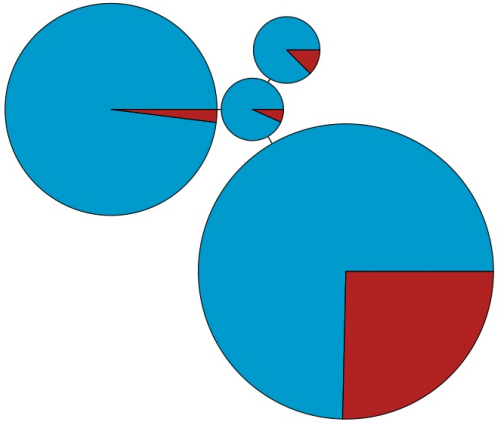
MBPA_000047601
Stenothoe monoculoides 0.99
(Arthropoda: Amphipoda)



MBPA_000043002
Phorbas plumosus 1.00
(Porifera: Poecilosclerida)



MBPA_000035627
Fam. Cyclopidae 0.82
(Arthropoda: Maxillopoda)



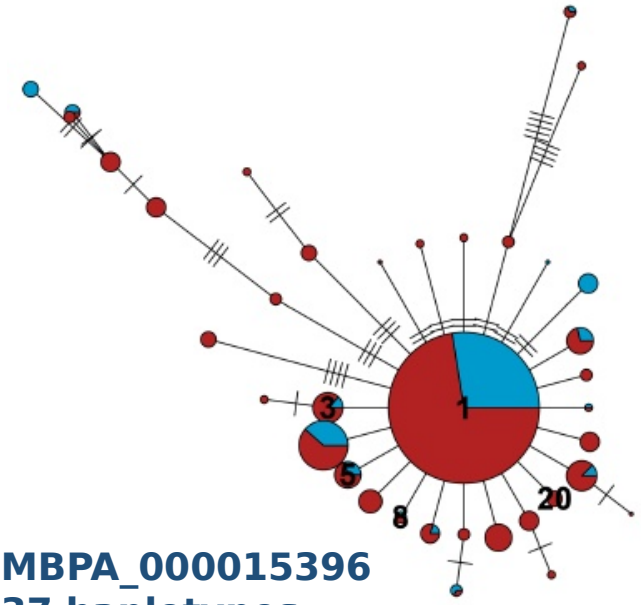
MBPA_000032954
Fam. Bacillariaceae 0.83
Stramenopiles: Bacillariophyceae

MEDITERRANEAN

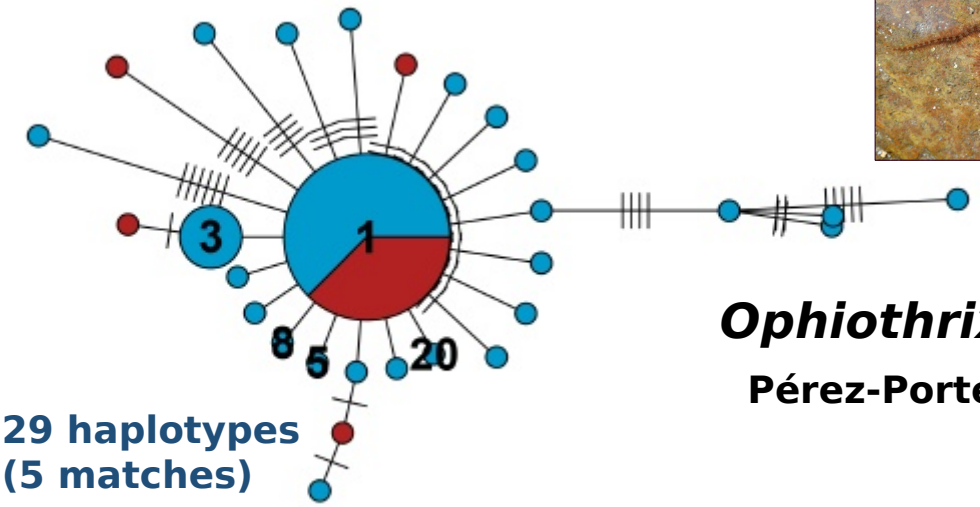
ATLANTIC

3.

Ground-truthing using classical phylogeographic studies



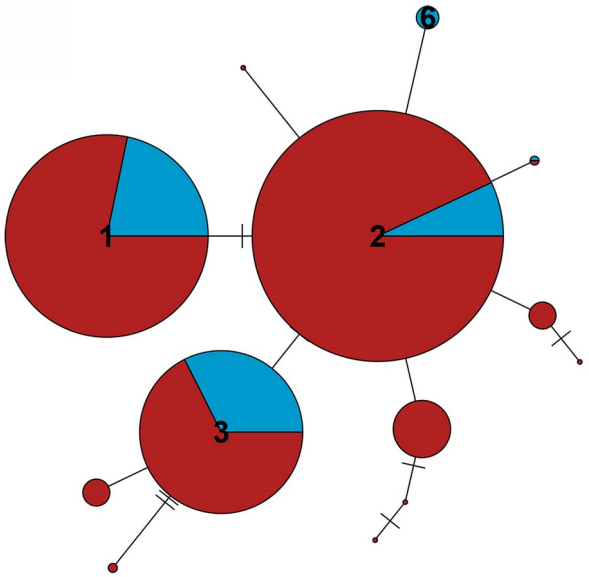
MBPA_000015396
37 haplotypes



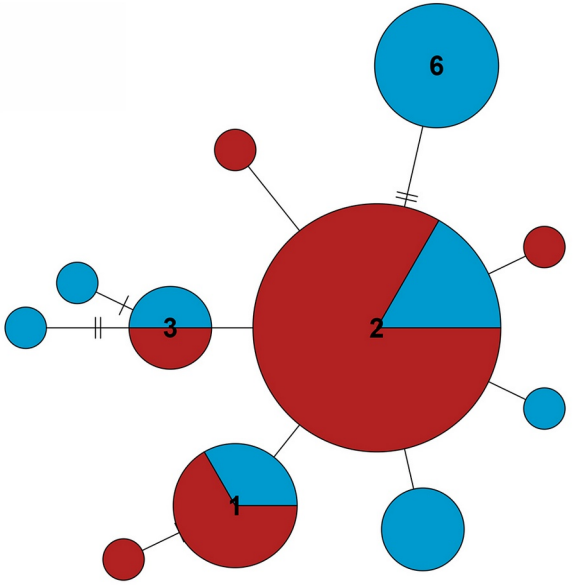
29 haplotypes
(5 matches)

Ophiothrix aff. fragilis

Pérez-Portela et al. (2013)



MBPA_000000697 **13 haplotypes**



11 haplotypes (4 matches)

Paracentrotus lividus

Duran et al. (2004)

4.

CLUSTERING OR DENOISING?

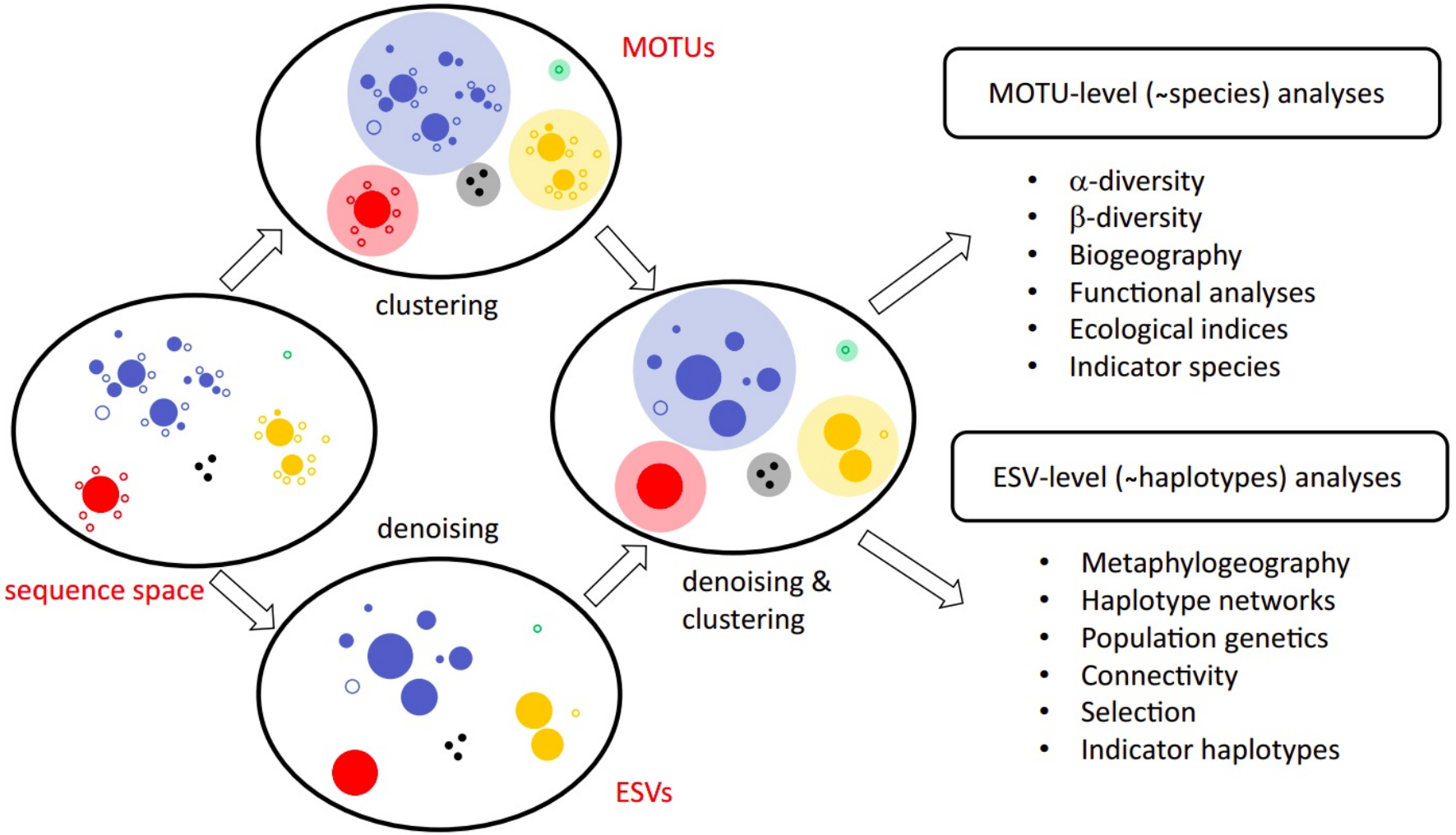
BMC Bioinformatics

Antich *et al. BMC Bioinformatics* (2021) 22:177
https://doi.org/10.1186/s12859-021-04115-6

To denoise or to cluster? That is not the question. Optimizing pipelines for COI metabarcoding and metaphylogeography

A. Antich, C. Palacin, O.S. Wangensteen, X. Turon

Why choosing between Denoising OR clustering?
When you can have both Denoising AND clustering!



4.

CLUSTERING AND DENOISING!

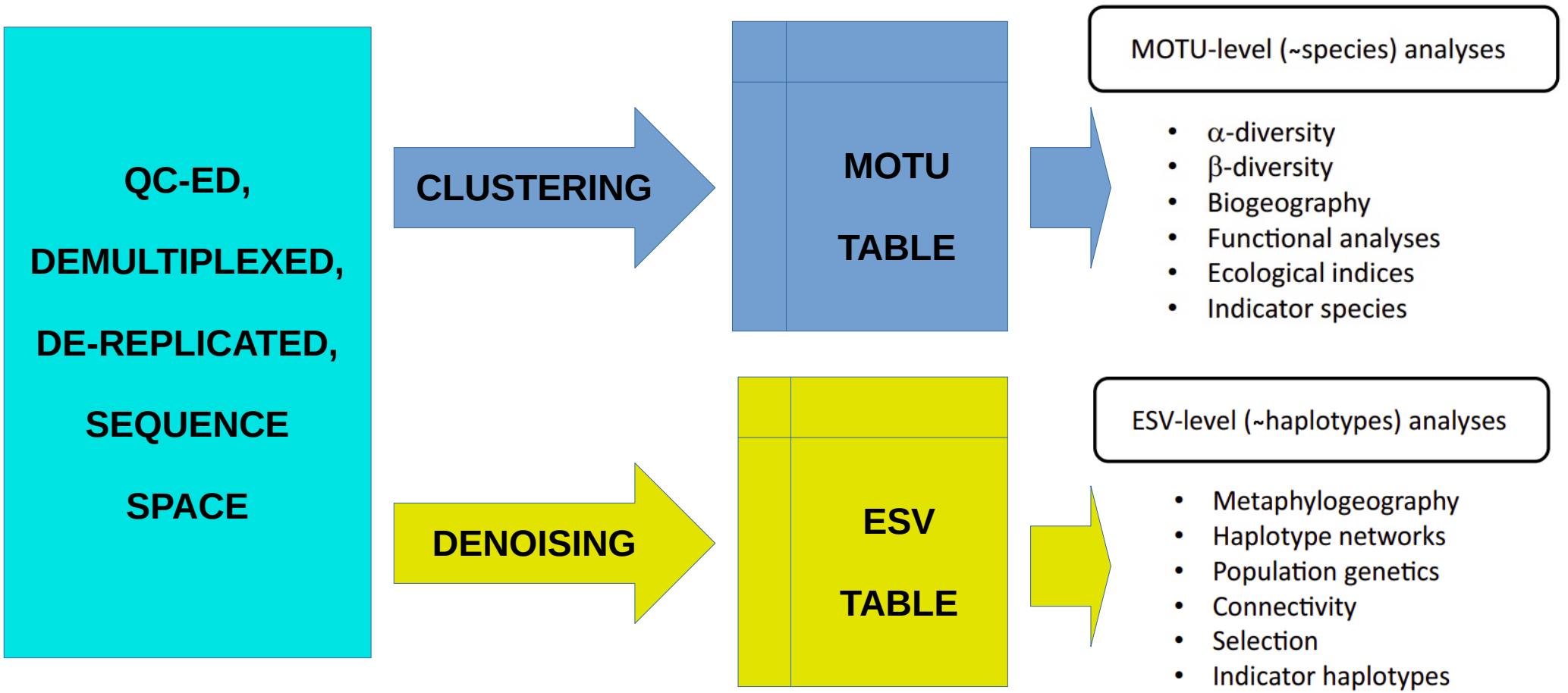
BMC Bioinformatics

Antich *et al. BMC Bioinformatics* (2021) 22:177
https://doi.org/10.1186/s12859-021-04115-6

To denoise or to cluster? That is not the question. Optimizing pipelines for COI metabarcoding and metaphylogeography

A. Antich, C. Palacin, O.S. Wangensteen, X. Turon

COMPREHENSIVE METABARCODING PIPELINE FOR A HYPERVARIABLE MARKER



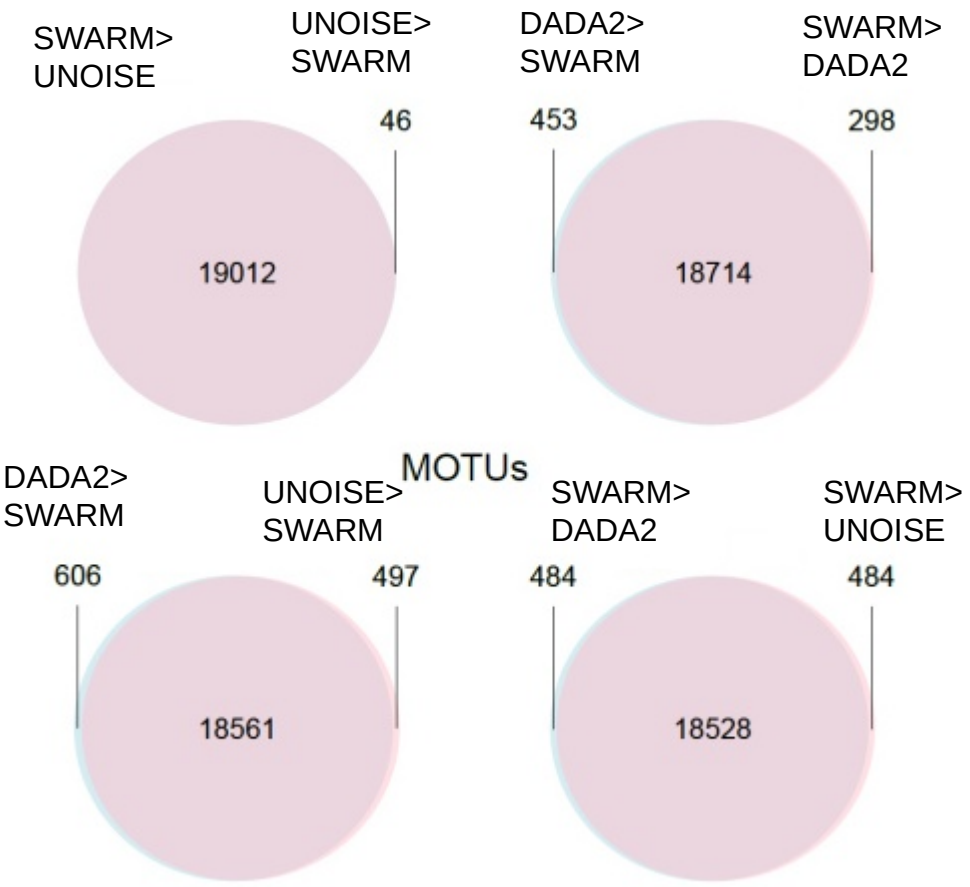
4. CLUSTERING FIRST OR DENOISING FIRST?

BMC Bioinformatics

Antich *et al. BMC Bioinformatics* (2021) 22:177
https://doi.org/10.1186/s12859-021-04115-6

To denoise or to cluster? That is not the question. Optimizing pipelines for COI metabarcoding and metaphylogeography

A. Antich, C. Palacin, O.S. Wangensteen, X. Turon



SWARM first, followed by denoising
=
Denoising first, followed by SWARM

The order does not really matter!

5.

DnoisE: A NEW DENOISING ALGORITHM USING CODON ENTROPY INFORMATION

DnoisE is based in the UNOISE3 algorithm (Edgar 2016)

$$\beta(d) = 0.5^{\alpha * d+1}$$

Edgar's abundance skew criterion to decide when a possible daughter sequence comes from a possible mother sequence

***d* is the Levenshtein distance (minimum number of changes to transform one sequence into the other) Changes in any position count equally for *d*!**

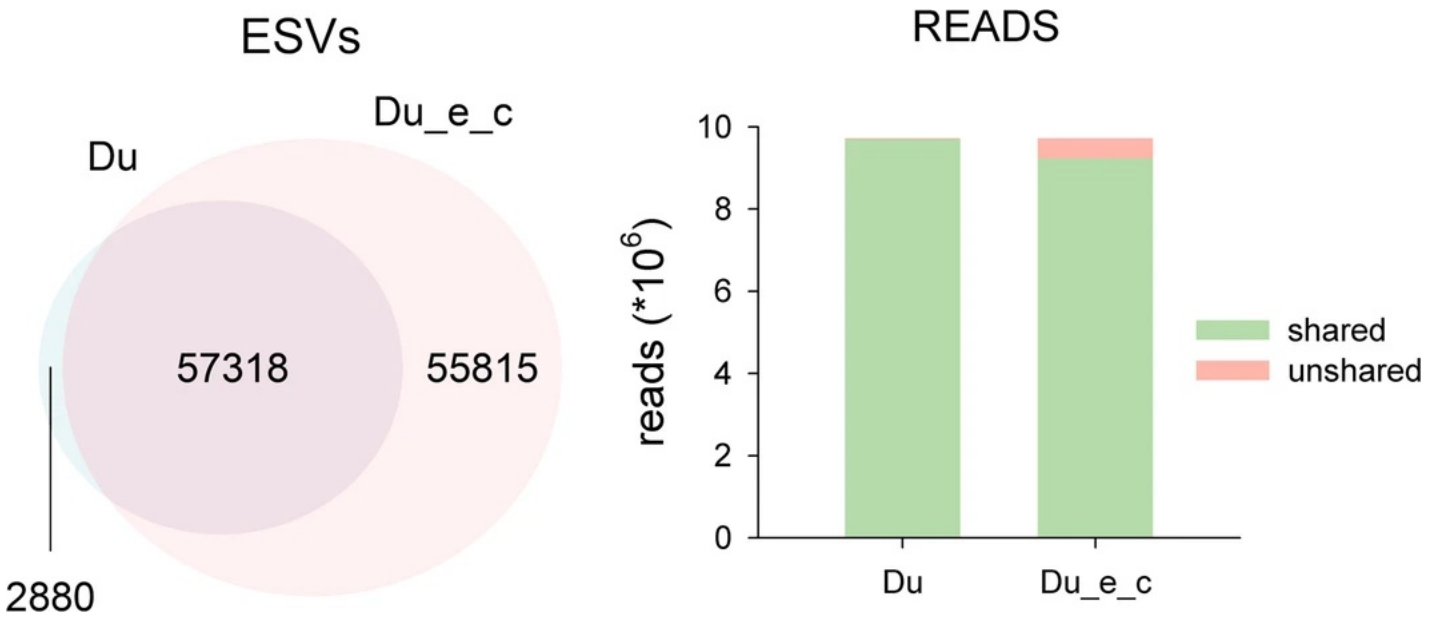
DnoisE uses a corrected d_{corr} distance, assigning different weights if the change is in the first, second or third codon position

$$d_{corr} = \sum_{i=1}^{3} d(i) * entropy(i)/(entropy(1) + entropy(2) + entropy(3))$$

These weights can either be calculated from the analysis of the particular dataset, or average values calculated for eukaryotic communities can be used. There is still much experimental work to do to get the right values!

5.

DnoisE RESULTS: MANY MORE ESVs ARE RETAINED AS VALID HAPLOTYPES



Our dataset from DnoisE has 113,133 valid ESVs, compared to 60,198 ESVs from UNOISE3 (88% more haplotypes are retained)

As expected, most of the new retained haplotypes have changes in the third position of the codon

The average of ESVs / MOTU is 113,133 / 19,012 = 6 haplotypes / MOTU

This can be compared to the typical results obtained from classical phylogeographical studies

5.

THE MJOLNIR PIPELINE

Metabarcoding **J**oining **O**bitools & **L**inkage **N**etworks **I**n **R**

MJOLNIR will integrate clustering by SWARM and denoising by DnoisE within a user-friendly pipeline

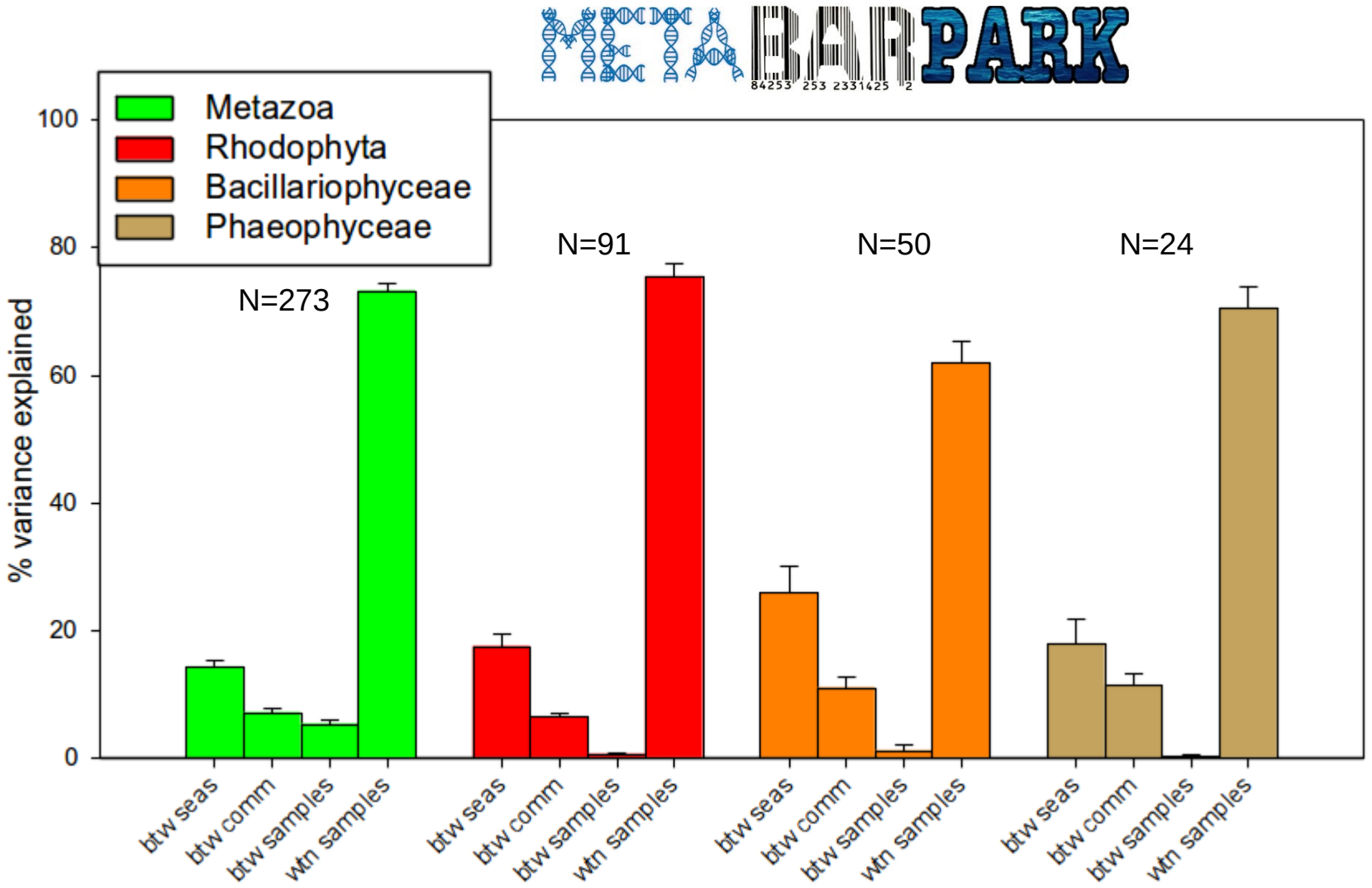
https://github.com/uit-metabarcoding/MJOLNIR

6.

EXAMPLE I: METAPHYLOGEOGRAPHICAL PATTERNS BY TAXONOMIC GROUPS

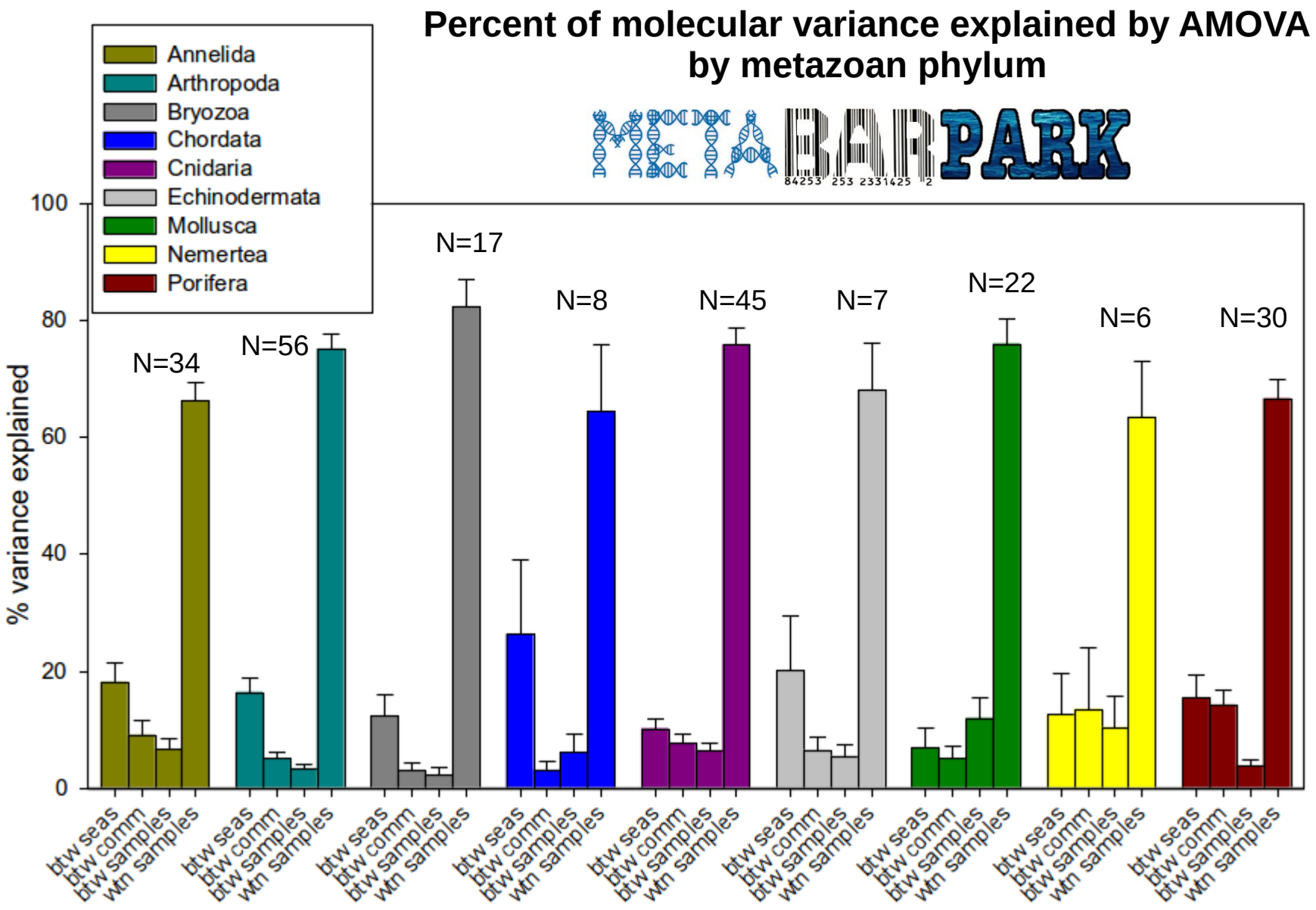
Patterns of molecular variance (AMOVA)

Percent of molecular variance explained by AMOVA by taxonomic group



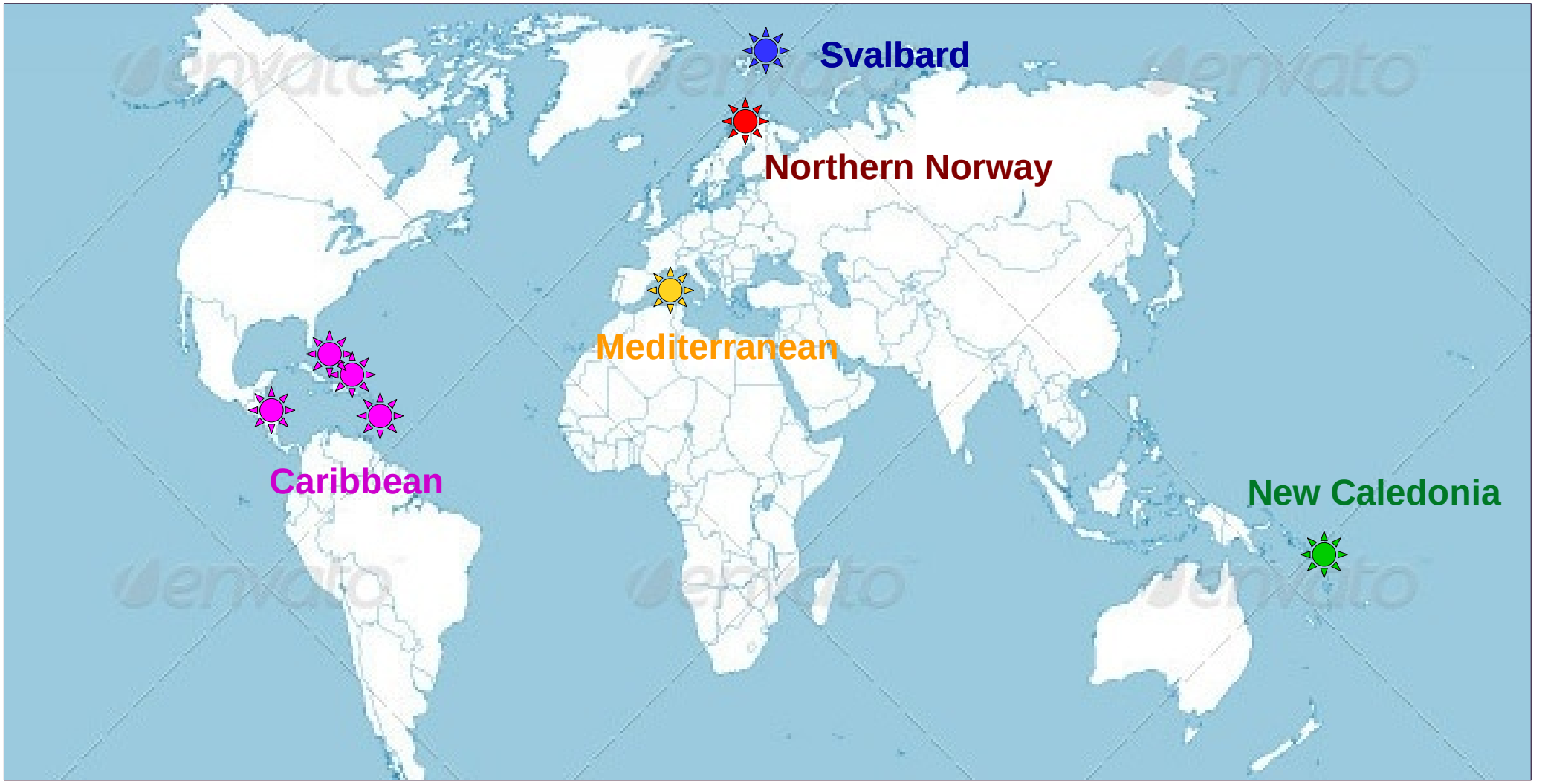
6.

EXAMPLE I: METAPHYLOGEOGRAPHICAL PATTERNS FOR METAZOAN PHYLA



6.

EXAMPLE II: 340 SAMPLES OF FILTERED SEAWATER FROM 5 GEOGRAPHICAL AREAS



- DNA extracted using different methods
- Amplified in 3 different laboratories
- Sequenced in 6 different runs, in 2 different MiSeq sequencers

6.

SUMMARY OF THIS PIPELINE

59.4 Million initial reads

22.5 Million reads after quality filtering

Belonging to 6,613,246 unique sequences

43,667 non-singleton eukaryotic MOTUs after clustering by SWARM d=13

452 eukaryotic MOTUs present in at least 3 geographical areas,
with abundances of > 10 reads in each area

Shared MOTUs were then de-clustered into unique sequences.

- Then they were denoised using Unoise2 with alpha=5

- Then low-abundance haplotypes were filtered out using the threshold selected using the entropy-ratio criterium

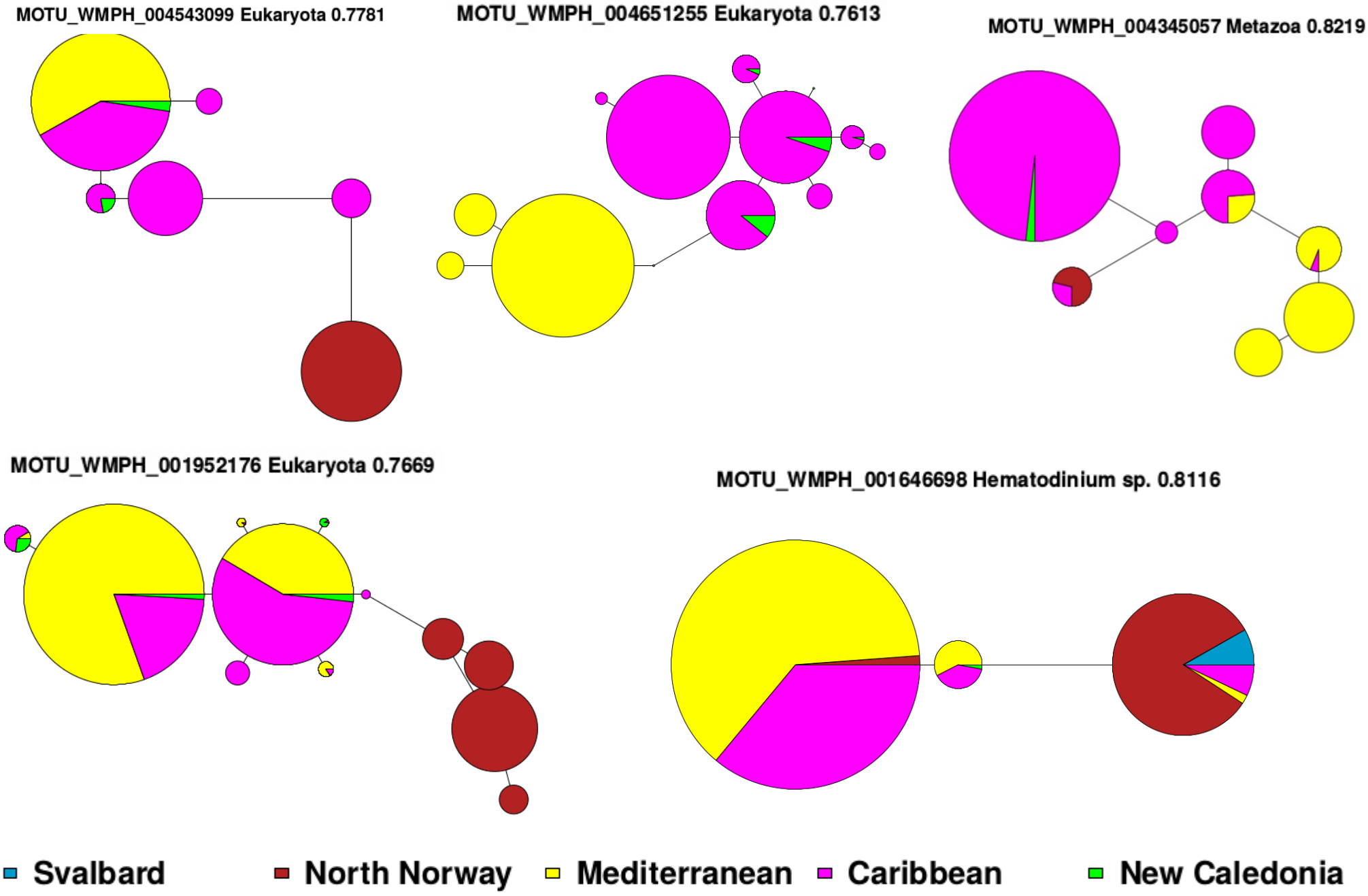
6.

TAXONOMIC DISTRIBUTION OF SHARED MOTUS (PRESENT IN ≥ 3 AREAS)

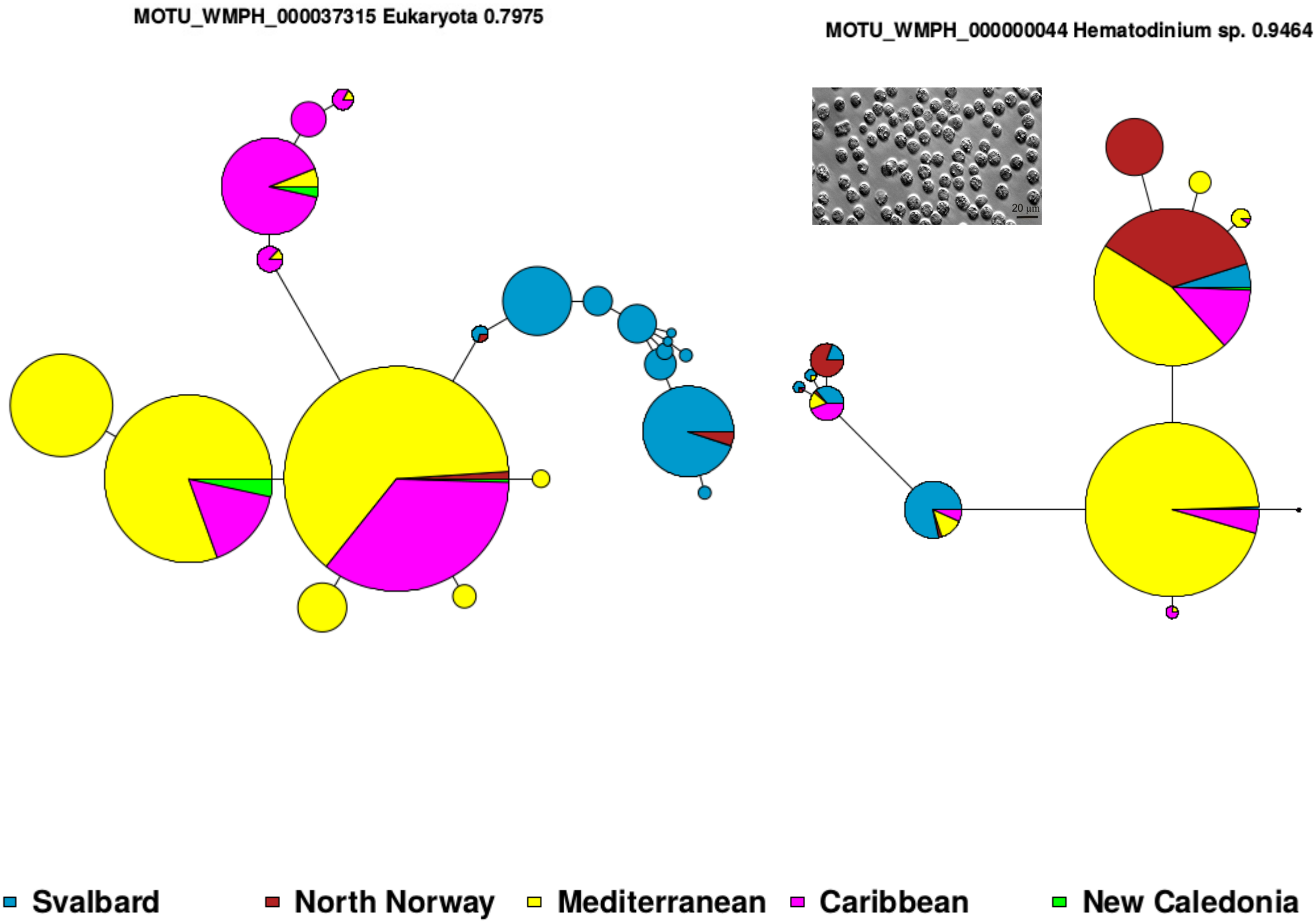
Rhodophyta	5
Chlorophyta	13
Dinoflagellata	46
Cryptophyta	5
Haptophyta	66
Cercozoa	1
Bacillariophyta	12
Ochrophyta	18
Oomycota	13
Ascomycota	3
Basidiomycota	1
Mucoromycota	1
Arthropoda	6
Chordata	4
Cnidaria	2
Porifera	1
Unassigned Eukaryota	255

6.

SOME EXAMPLES OF HAPLOTYPE NETWORKS



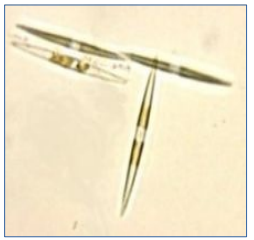
6.


MOTU_WMPH_000037315 Eukaryota 0.7975
MOTU_WMPH_000000044 Hematodinium sp. 0.9464
20 µm
Svalbard
North Norway
Mediterranean
Caribbean
New Caledonia

6.

MOTU_WMPH_001440436 Clytia hemisphaerica 0.9808

MOTU_WMPH_000896558 Pseudo-nitzschia cuspidata 0.9776



MOTU_WMPH_000300732 Chrysochromulina 0.8076

MOTU_WMPH_000050469 Emiliania huxleyi 0.9904

Svalbard North Norway Mediterranean Caribbean New Caledonia


TUSEN TAKK!!!