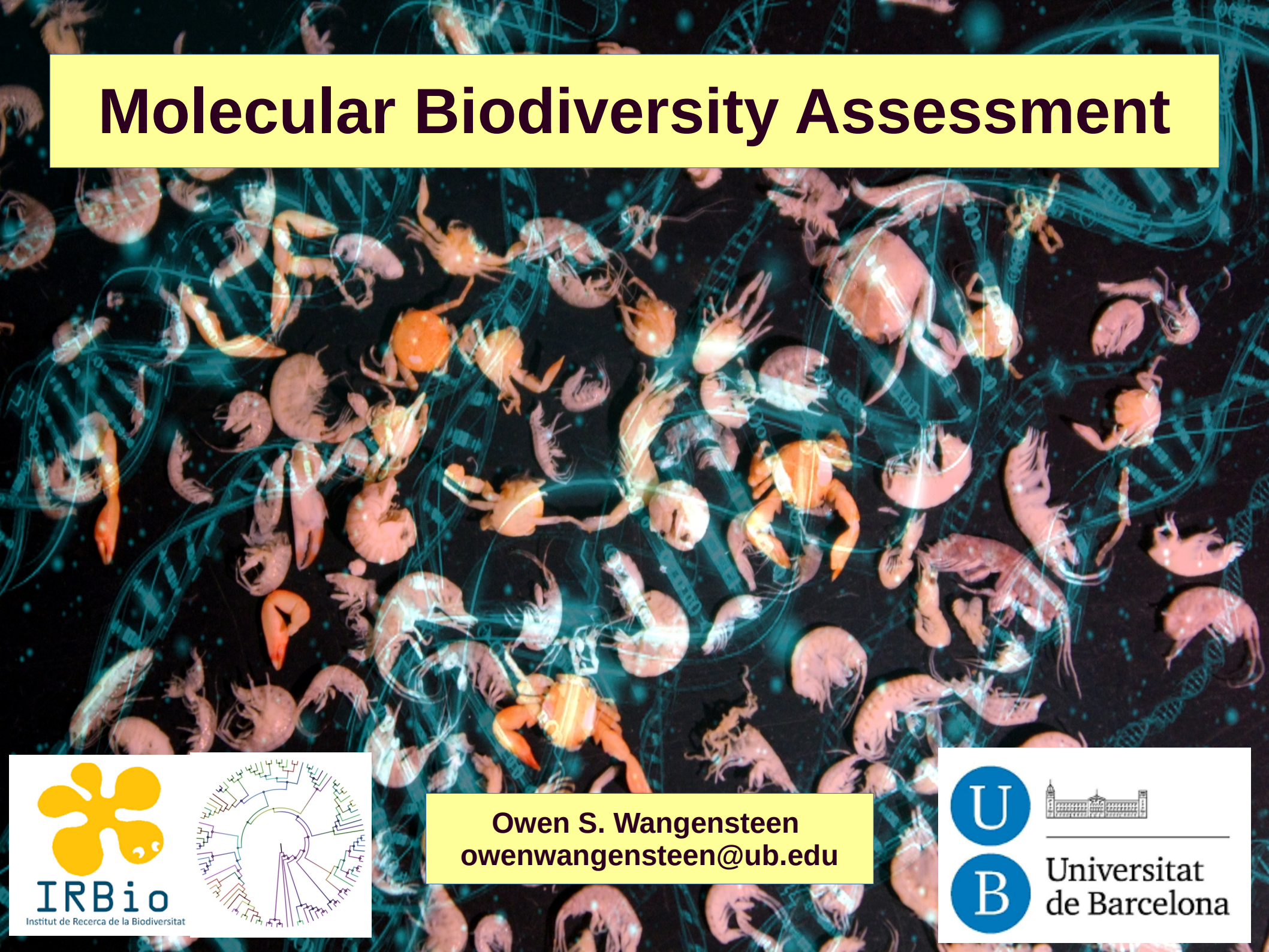
Molecular Biodiversity Assessment



Owen S. Wangensteen
owenwangensteen@ub.edu

IRBio
Institut de Recerca de la Biodiversitat

Universitat de Barcelona

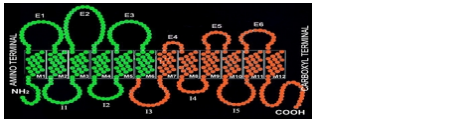
OUTLINE:

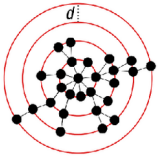
1. Importance of biodiversity assessment and biomonitoring

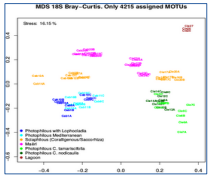
2. Molecular approaches to biodiversity:
   DNA-barcoding, DNA-metabarcoding.
   Other metagenomics techniques.

3. The metabarcoding approach:
   Community DNA vs extra-organismal DNA. Quantitative value.
   Metabarcoding markers, metabarcoding primer design.

4. Metabarcoding workflow and pipelines
   - 4.1. Sampling and pre-processing
   - 4.2. DNA extraction and PCR
   - 4.3. Bioinformatics: demultiplexing and QC
   - 4.4. Bioinformatics: MOTU definition / denoising /clustering
   - 4.5. Bioinformatics: taxonomic assignment
   - 4.6. Bioinformatics: final dataset refining

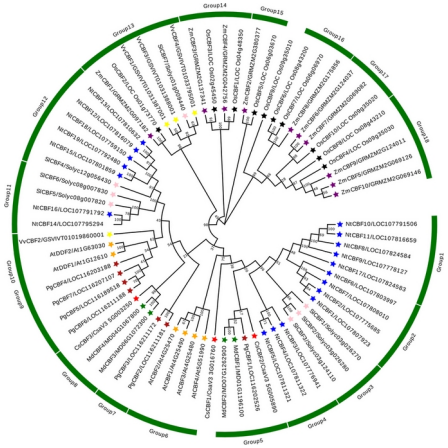






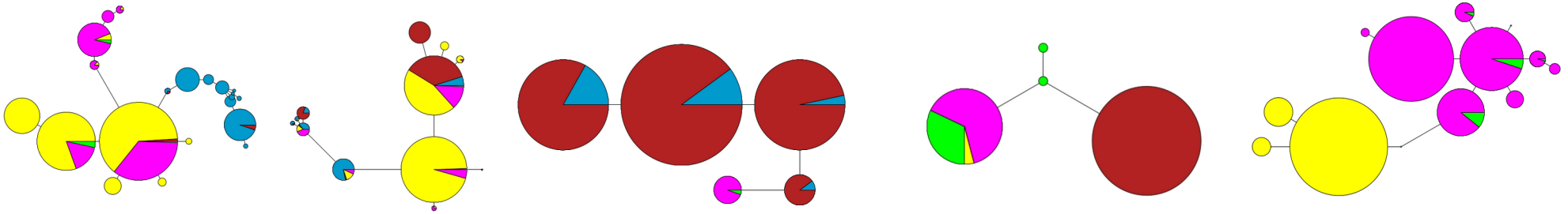
1. IMPORTANCE OF BIODIVERSITY ASSESSMENT AND BIOMONITORING

What is a class on **Biodiversity Assessment** like you doing
in a course on **Phylogenomics and Population Genomics** like this?



1- Classical (morphological) biodiversity assessment methods are on their way to be replaced by **molecular biodiversity assessment** methods. Bioinformatic analyses of HT-sequencing datasets and taxonomic assignment of detected sequences is crucial for these applications.

2- Some molecular biodiversity assessment methods can be used to infer population genetic patterns for many species at one time, even for species without taxonomic assignment (**metaphylogeography**)



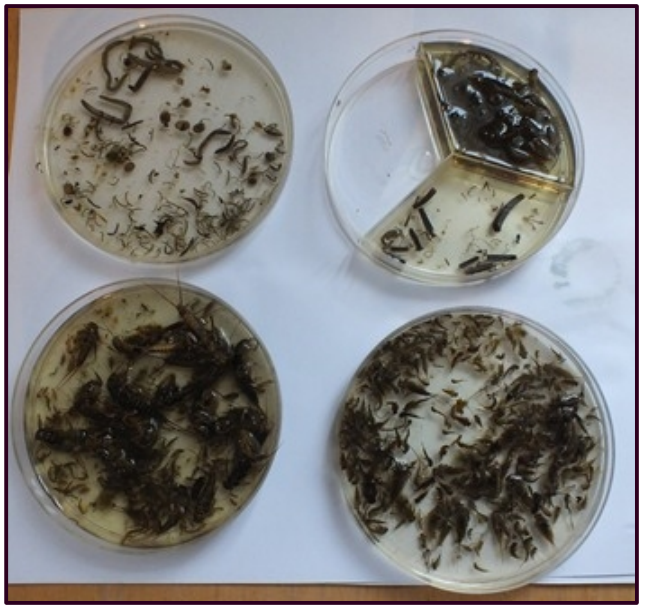
BIOMONITORING AND ECOLOGICAL QUALITY

Some examples:

- River benthic macroinvertebrates

- Marine benthic macroinvertebrates

- Freshwater microalgae (diatoms)



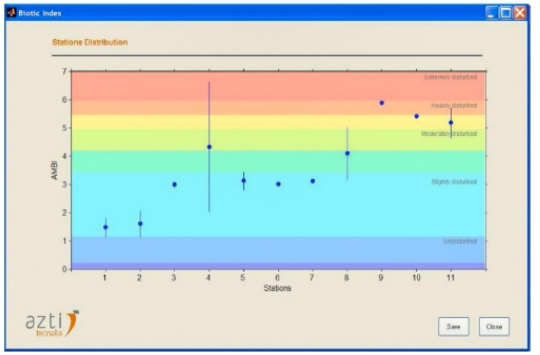
ECOLOGICAL QUALITY INDICES BASED ON BIODIVERSITY

Some examples:

- **RICT: River benthic macroinvertebrates**

- **AMBI: Marine benthic macroinvertebrates**

- **IBD: Freshwater diatoms**



Drawbacks:

- **Need of an experienced taxonomist**
- **Need of adaptive implementation to different biogeographical areas**
- **Slow! Thorough analysis of some complex samples can take weeks!**
- **Subjectivity, lack of repeatability and lack of traceability**

ERRORS IN MORPHOLOGY-BASED ECOLOGICAL QUALITY INDICES

J. N. Am. Benthol. Soc., 2010, 29(4):1279–1291
© 2010 by The North American Benthological Society
DOI: 10.1899/09-183.1
Published online: 7 September 2010

First audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program: human error greatly lowers precision of assessment results

Peter Haase[1,3], Steffen U. Pauls[1,2,4], Karin Schindehütte[1,5], AND Andrea Sundermann[1,6]

Errors and biases:

- 29 % of specimens and 21% of taxa were overlooked by the analysts
- 30 % of assigned taxa differed between analysts and auditors
- 34 % of audited samples were assigned to the wrong ecological quality class

In comparison, molecular biodiversity assessment methods:

- No need of an experienced taxonomist. A lab technician can do all the work
- Fast and scalable! Hundreds of samples can be multiplexed in the same run
- Objectivity, repeatability, perfect traceability, sequence datasets can be stored forever and they are accessible online for everyone
- Accuracy of the assessment depends on the existence of good reference sequence databases

DNA-BARCODING

THE ROYAL SOCIETY

Received 29 July 2002
Accepted 30 September 2002
Published online 8 January 2003

Biological identifications through DNA barcodes

Paul D. N. Hebert*, Alina Cywinska, Shelley L. Ball and Jeremy R. deWaard

Department of Zoology, University of Guelph, Guelph, Ontario N1G 2W1, Canada

Although much biological research depends upon species diagnoses, taxonomic expertise is collapsing. We are convinced that the sole prospect for a sustainable identification capability lies in the construction of systems that employ DNA sequences as taxon 'barcodes'. We establish that the mitochondrial gene cytochrome *c* oxidase I (COI) can serve as the core of a global bioidentification system for animals. First, we demonstrate that COI profiles, derived from the low-density sampling of higher taxonomic categories, ordinarily assign newly analysed taxa to the appropriate phylum or order. Second, we demonstrate that species-level assignments can be obtained by creating comprehensive COI profiles. A model COI profile, based upon the analysis of a single individual from each of 200 closely allied species of lepidopterans, was 100% successful in correctly identifying subsequent specimens. When fully developed, a COI identification system will provide a reliable, cost-effective and accessible solution to the current problem of species identification. Its assembly will also generate important new insights into the diversification of life and the rules of molecular evolution.

Keywords: molecular taxonomy; mitochondrial DNA; animals; insects; sequence diversity; evolution

Paul Hebert (2003)

A short DNA fragment (~600 pb) of a highly variable genetic locus is enough to unequivocally identify a species with 100% success

THE BARCODE OF LIFE DATABASE

www.boldsystems.org

Barcodes: 17,552,000
BINs: 1,109,000

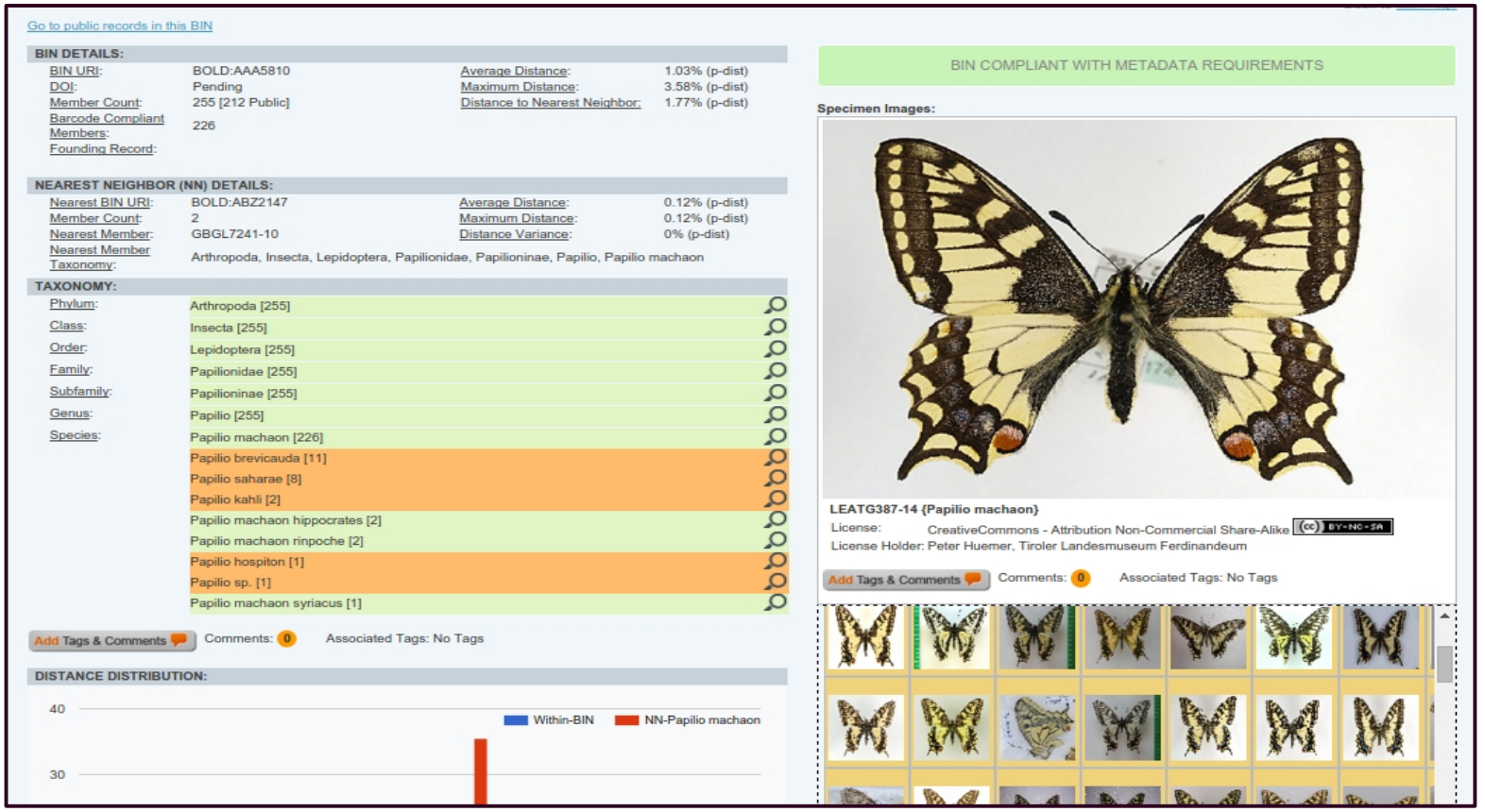
Species with barcodes: 354,000 (formally described)

DNA-BARCODING IDENTIFICATION

Sanger sequencing

```
AACATTATATTTTATTTTTGGTATTTGAGCAAGTATATTAGGAACTTCTCTTAGTTTATTAATTCGAACTGAATT
AGGAACACCGGGCTCATTAATTGGAGATGATCAAATTTATAATACTATTGTTACAGCTCACGCTTTTATTATAAT
TTTTTTTATAGTTATACCTATTATAATTGGAGGATTTGGAAATTGATTAATTCCTTTAATATTAGGAGCCCCTGA
CATAGCTTTCCCACGAATAAATAATATAAGATTTTGATTATTACCTCCTTCATTAACTCTTTTAATTTCAAGAAT
AATTGTAGAAAATGGAGCTGGAACTGGTTGAACTGTTTACCCCCCTCTCTCTTCTAACATTGCACATGGAAGAAG
ATCAGTAGATTTAGTTATTTTTTCTTTACATTTAGCAGGTATCTCATCAATTTTAGGGGCAATTAATTTCATCAC
AACAATTATTAATATACGTATTAATAATATATCATTTGATCAAATACCTTTATTTGTTTGAGCTGTTGGAATTAC
AGCTTTATTATTACTTCTTTCCTTACCAGTTTTAGCTGGAGCTATTACTATATTATTAACAGATCGAAATTTAAA
TACTTCTTTTTTTGATCCTGCTGGAGGGGGAGACCCTATTTTATACCAACATTTATTT
```

REFERENCE DATABASE
www.boldsystems.org



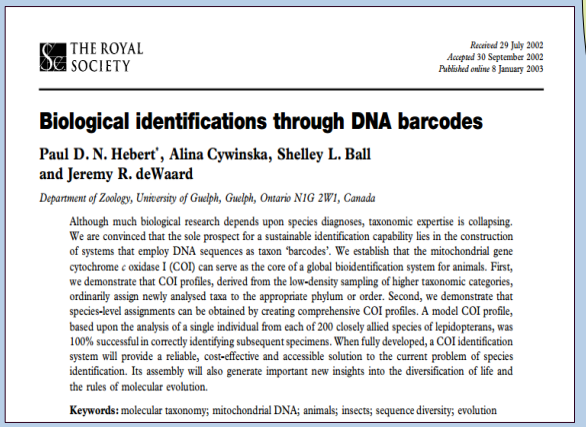
BIN: BOLD:AAA5810

Papilio machaon

METABARCODING

DNA-BARCODING

(2003)



THE ROYAL SOCIETY

Received 29 July 2002
Accepted 30 September 2002
Published online 8 January 2003

Biological identifications through DNA barcodes

Paul D. N. Hebert', Alina Cywinska, Shelley L. Ball and Jeremy R. deWaard

Department of Zoology, University of Guelph, Guelph, Ontario N1G 2W1, Canada

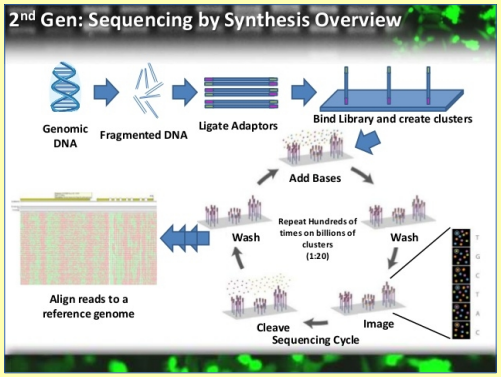
Although much biological research depends upon species diagnoses, taxonomic expertise is collapsing. We are convinced that the sole prospect for a sustainable identification capability lies in the construction of systems that employ DNA sequences as taxon 'barcodes'. We establish that the mitochondrial gene cytochrome *c* oxidase I (COI) can serve as the core of a global bioidentification system for animals. First, we demonstrate that COI profiles, derived from the low-density sampling of higher taxonomic categories, ordinarily assign newly analysed taxa to the appropriate phylum or order. Second, we demonstrate that species-level assignments can be obtained by creating comprehensive COI profiles. A model COI profile, based upon the analysis of a single individual from each of 200 closely allied species of lepidopterans, was 100% successful in correctly identifying subsequent specimens. When fully developed, a COI identification system will provide a reliable, cost-effective and accessible solution to the current problem of species identification. Its assembly will also generate important new insights into the diversification of life and the rules of molecular evolution.

Keywords: molecular taxonomy; mitochondrial DNA; animals; insects; sequence diversity; evolution

METABARCODING

HIGH-THROUGHPUT SEQUENCING

(2002-2010)



DNA BARCODING vs METABARCODING

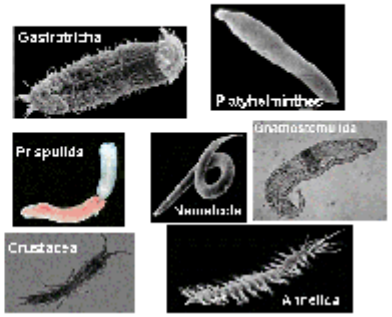
DNA BARCODING

DNA Extraction → Amplification of a gene fragment using universal primers (18S, COI, etc.) → Sanger sequencing → *Sphaerechinus granularis*

1 DNA sequence

DNA METABARCODING

DNA Extraction → Amplification of a gene fragment using universal primers (18S, COI, etc.) → Next-Generation Sequencing → Comparison to reference databases



10^5 – 10^7 DNA sequences

Detects not only living individuals, but also symbionts and parasites, gut contents, body fragments, dead remnants and extracellular DNA

HIGH THROUGHPUT SEQUENCING
SHORT FRAGMENTS

Roche GS-FLX+ 454

- Up to 1 M seqs
- 1000 bp max length
- 700 Mb per run

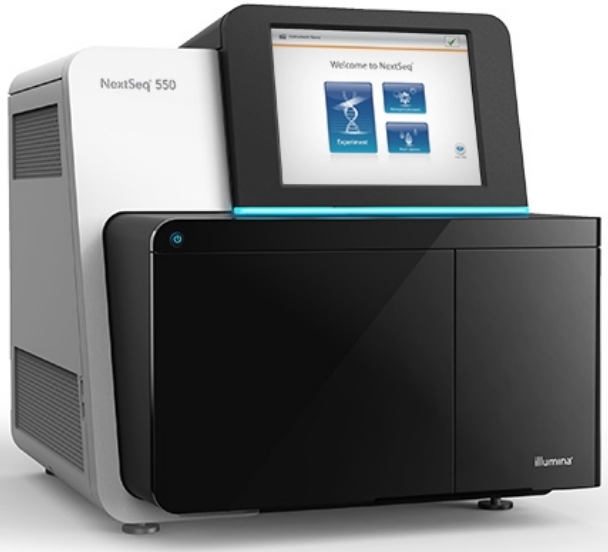
Illumina MiSeq

- Up to 25 M seqs
- 2x 300 bp max length
- 15 Gb per run

Illumina HiSeq 2500

- Up to 600 M seqs
- 2x 250 bp max length
- 300 Gb per run

HIGH THROUGHPUT SEQUENCING
SHORT FRAGMENTS



Illumina NextSeq 550

- Up to 400 M seqs
- 2 x 150 bp max length
- 120 Gb per run

Illumina NovaSeq 6000

- Up to 20,000 M seqs
- 2 x 150 bp max length
  (2 x 250 bp SP kit)
- 3000 Gb per run

HIGH THROUGHPUT SEQUENCING

LONGER FRAGMENTS

Oxford NT MinION

- Up to 1 M seqs
- Variable. Up to 900 kb length!
- 5 Gb per run

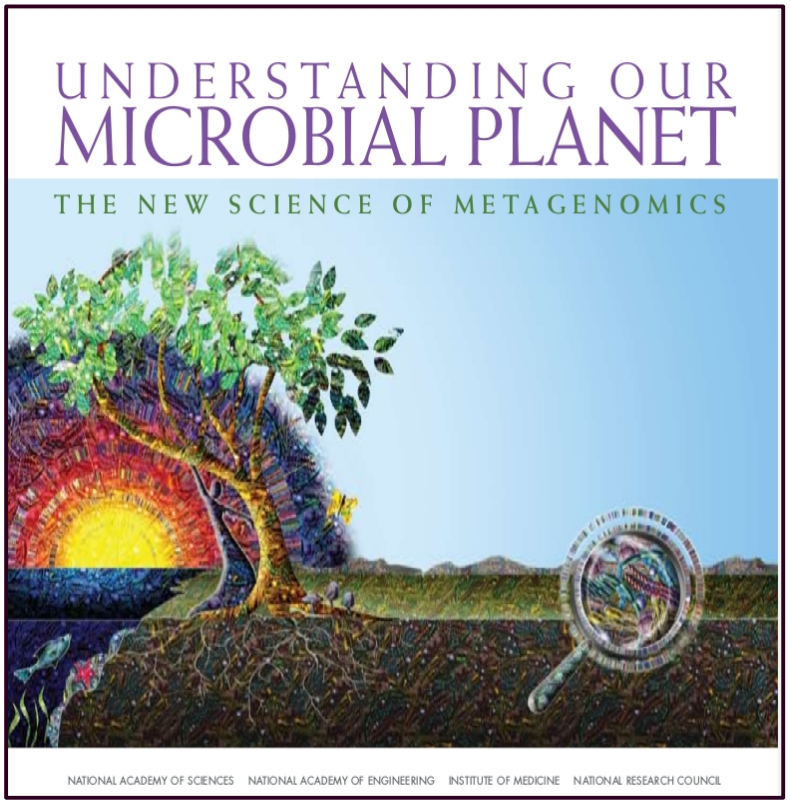
5-15% Error rates!
(for single reads)

PacBio Sequel

- Up to 500 k seqs
- Up to 35 kb max length
- 10 Gb per run

10-15% Error rates!
(for single reads)

PROKARYOTE 16S METABARCODING

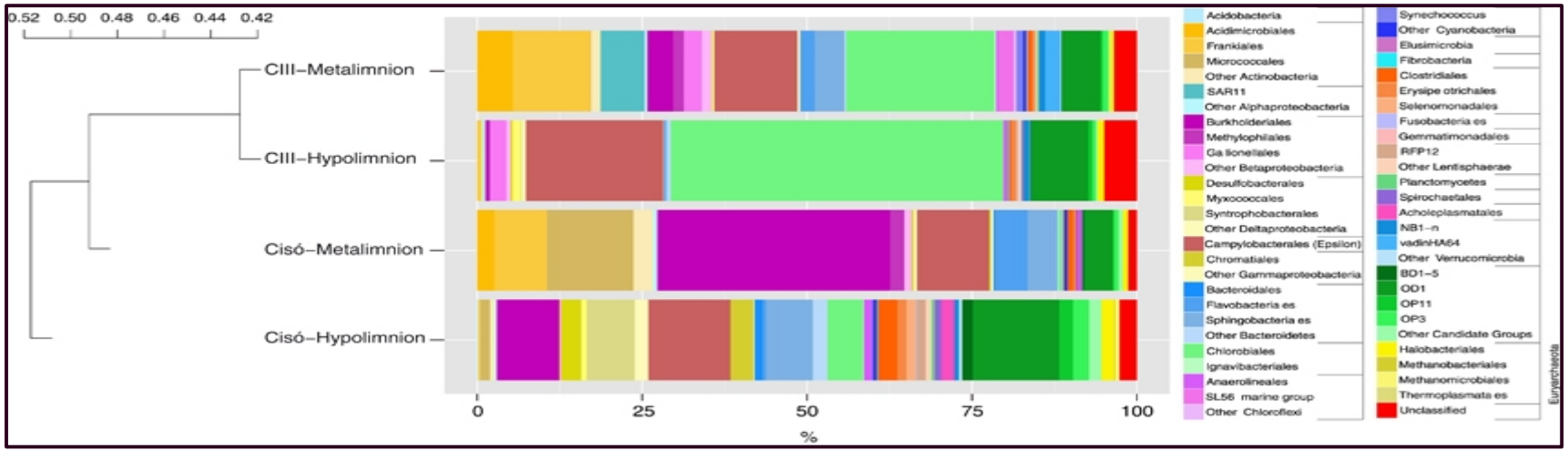


National Academy of Sciences 2007

https://nap.nationalacademies.org/resource/11902/metagenomics_final.pdf

- **Human microbiomes**
- **Marine microbes**
- **Soil prokaryote communities**
- **Bioremediation**
- **Water treatment ...**

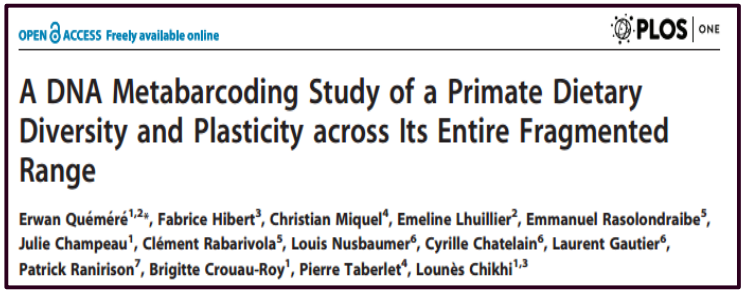
Well-established pipelines and markers (16S)
Commercial applications (several companies)



EUKARYOTIC METABARCODING

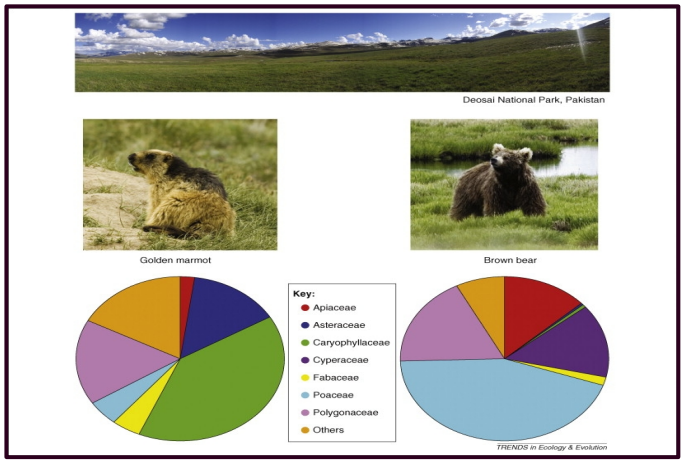
DEVELOPMENT WORK IN PROGRESS:

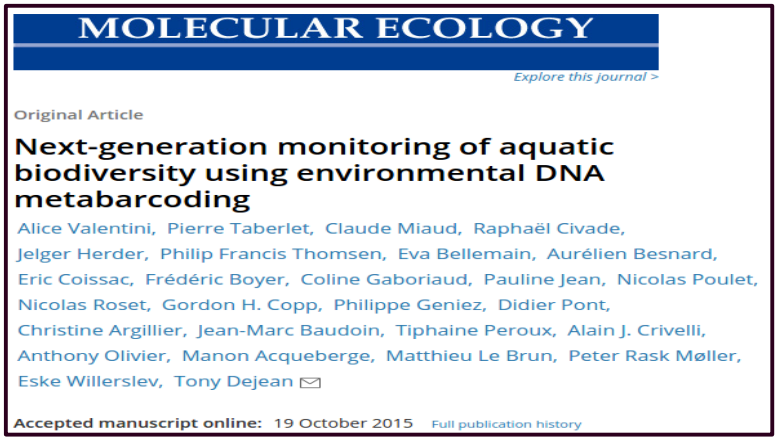
- **Different sampling techniques, preservation and DNA extraction procedures**
- **Some confusion between environmental DNA / community DNA / extraorganismal DNA**
- **Need to decide which markers yield the best results**
- **Need to find good metabarcoding primers: universal and specific**
- **Need to standardize analysis pipelines**
- **Need to improve statistical data treatment**


OPEN ACCESS Freely available online PLOS ONE

A DNA Metabarcoding Study of a Primate Dietary Diversity and Plasticity across Its Entire Fragmented Range

Erwan Quéméré[1,2*], Fabrice Hibert[3], Christian Miquel[4], Emeline Lhuillier[2], Emmanuel Rasolondraibe[5], Julie Champeau[1], Clément Rabarivola[5], Louis Nusbaumer[6], Cyrille Chatelain[6], Laurent Gautier[6], Patrick Ranirison[7], Brigitte Crouau-Roy[1], Pierre Taberlet[4], Lounès Chikhi[1,3]




MOLECULAR ECOLOGY

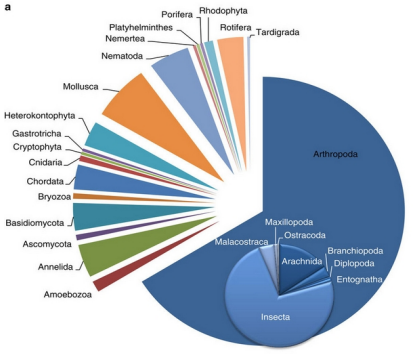
Explore this journal >

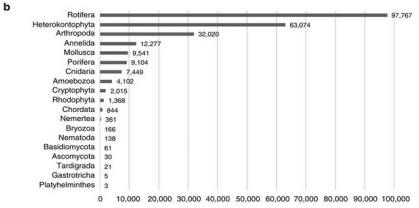
Original Article

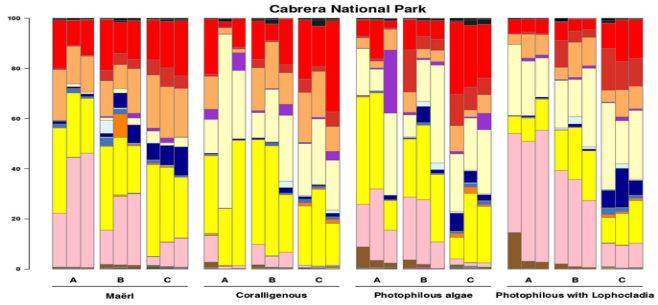
Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding

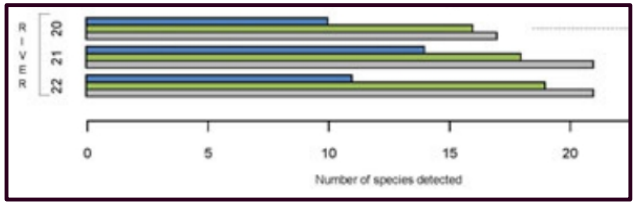
Alice Valentini, Pierre Taberlet, Claude Miaud, Raphaël Civade, Jelger Herder, Philip Francis Thomsen, Eva Bellemain, Aurélien Besnard, Eric Coissac, Frédéric Boyer, Coline Gaboriaud, Pauline Jean, Nicolas Poulet, Nicolas Roset, Gordon H. Copp, Philippe Geniez, Didier Pont, Christine Argillier, Jean-Marc Baudoin, Tiphaine Peroux, Alain J. Crivelli, Anthony Olivier, Manon Acqueberge, Matthieu Le Brun, Peter Rask Møller, Eske Willerslev, Tony Dejean

Accepted manuscript online: 19 October 2015 Full publication history

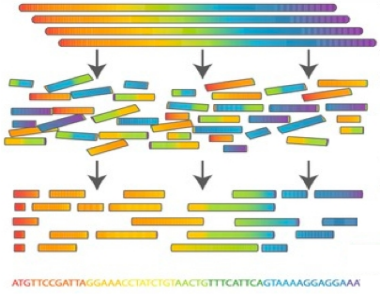








VARIETY OF METAGENOMICS TECHNIQUES

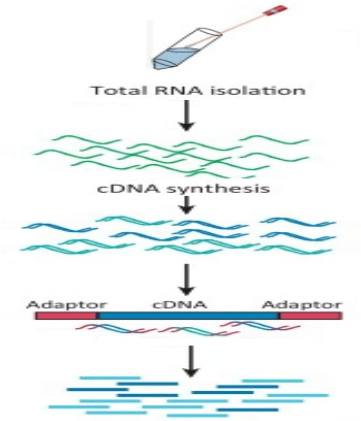


TOTAL DNA SEQUENCING (SHOTGUN SEQUENCING):

- METAGENOMICS (mostly Prokaryotic, currently)

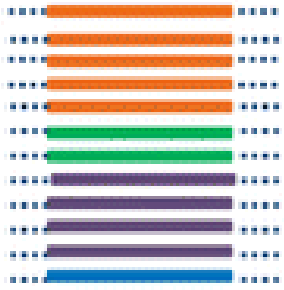
- MITOGENOMICS (MITOCHONDRIAL DNA METAGENOMICS)

- TARGET-ENRICHMENT TECHNIQUES (still experimental)



TOTAL RNA SEQUENCING (cDNA SHOTGUN):

- METATRANSCRIPTOMICS (mostly Prokaryotic, currently)



AMPLICON SEQUENCING (METABARCODING):

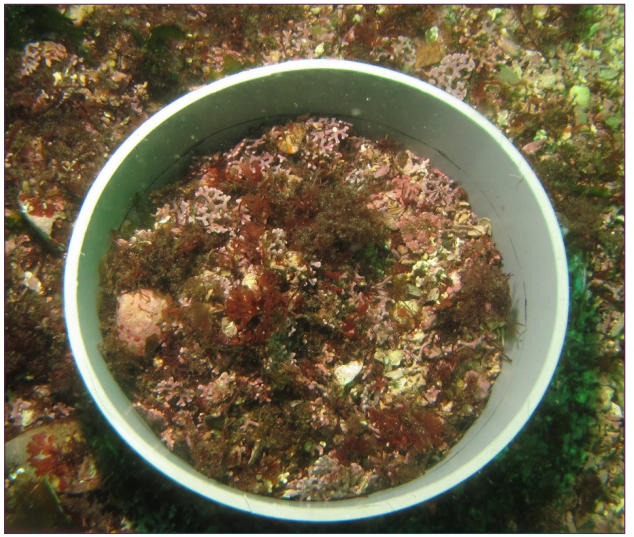
- DNA METABARCODING

- RNA METABARCODING (RIBOSOMAL MARKERS)

TYPES OF GENETIC SAMPLES

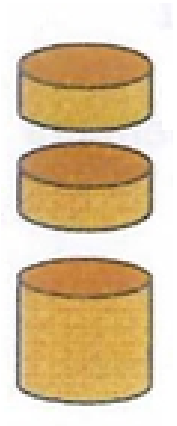
ENVIRONMENTAL DNA

COMMUNITY DNA

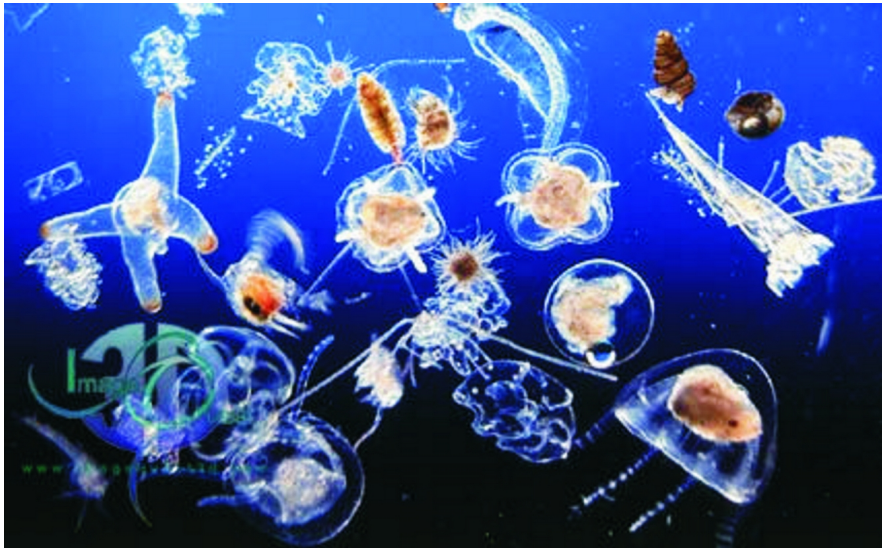


THIS DISTINCTION MAY BE BLURRY

MARINE SEDIMENTS



PLANKTON COMMUNITIES

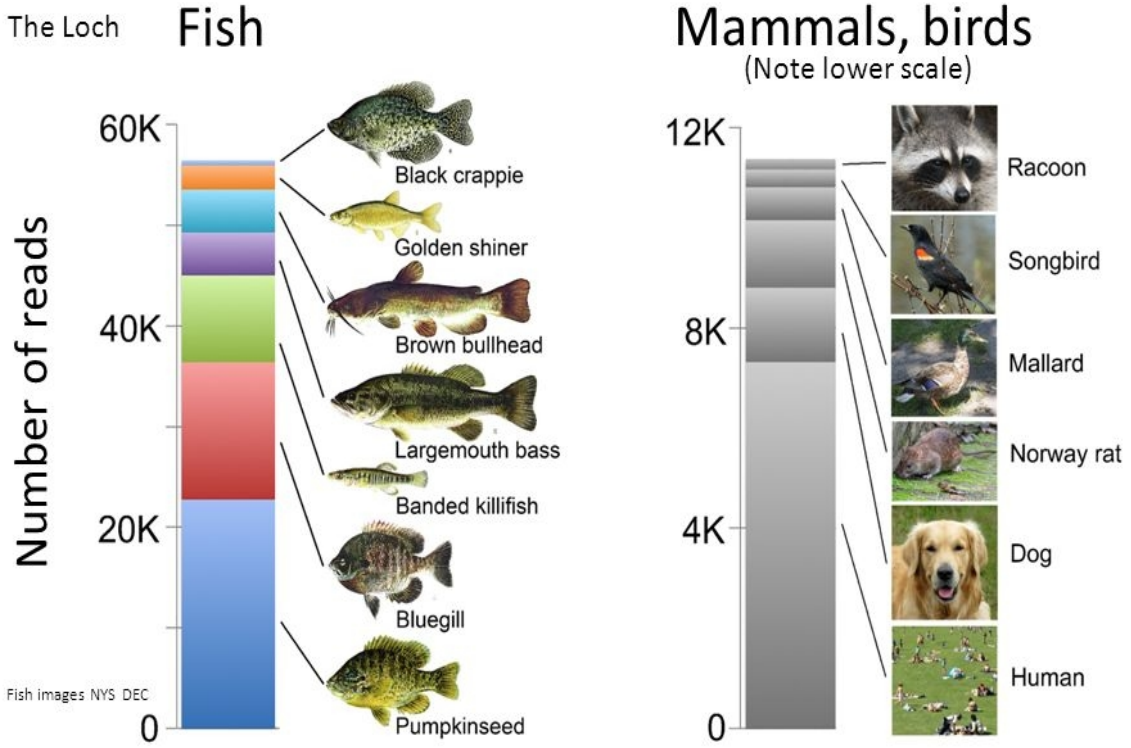


AND DEPENDS ON THE TARGET TAXA (SPECIFICITY OF PRIMERS USED)

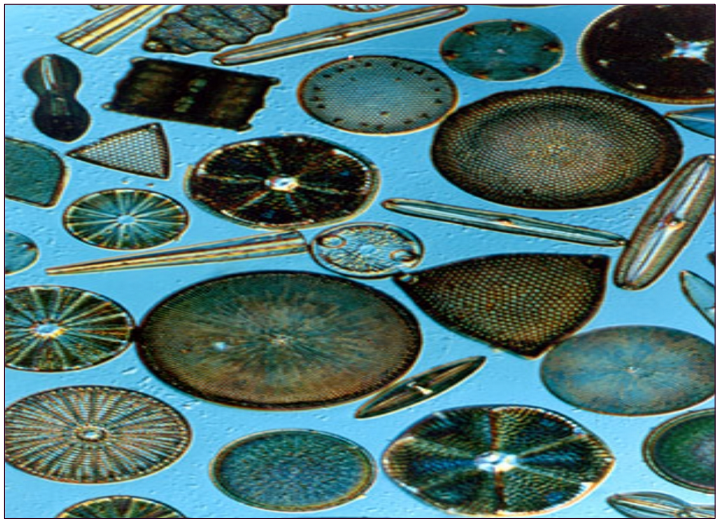
EXTRA-ORGANISMAL DNA

A dozen species in ¼ cup of water

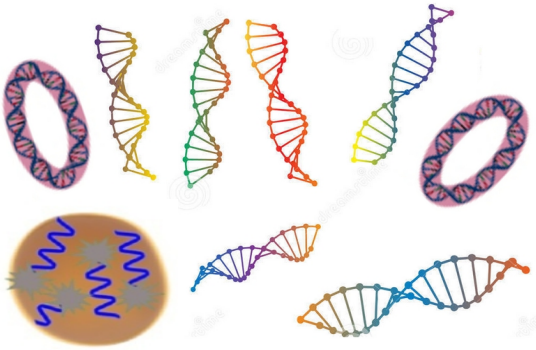
- Analyze 6.33 ng DNA (roughly 60 mL H2O); 7 x 10^4 reads
- Detected 7 species of fish, also mammals, birds



COMMUNITY DNA



EVERY eDNA SAMPLE IS ACTUALLY A COMBINATION OF TWO FRACTIONS:



eDNA **=** **extra-organismal DNA** **+** **community DNA**

- total DNA
- envDNA

- “free” DNA
- “dissolved” DNA
- “true” eDNA
- extracellular DNA
- “dead” DNA
- "trace" DNA

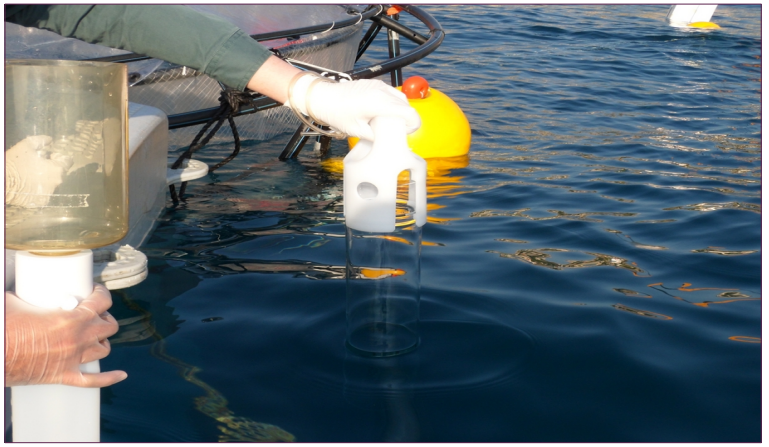
- “living” DNA
- organismal eDNA

Trade-offs between reducing complex terminology and producing accurate interpretations from environmental DNA: Comment on “Environmental DNA: What's behind the term?” by Pawlowski et al., (2020)

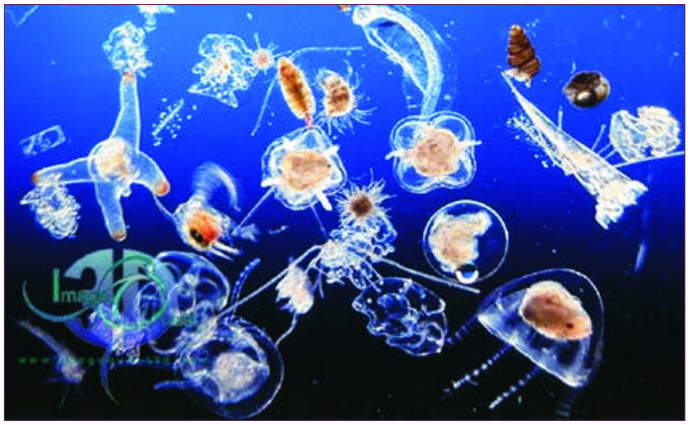
https://onlinelibrary.wiley.com/doi/epdf/10.1111/mec.15942

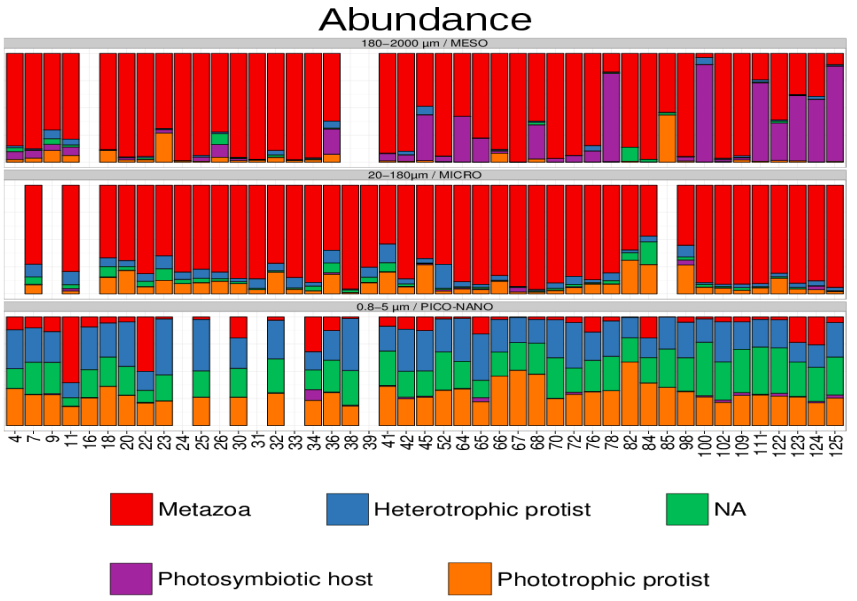
WITH IMPLICATIONS FOR THE QUANTITATIVE VALUE

EXTRA-ORGANISMAL DNA



COMMUNITY DNA





PRESENCE / ABSENCE DATA

QUANTITATIVE DATA

SAMPLING STRATEGY:
A QUESTION OF TARGET CONCENTRATION

THE NEEDLE? **or** **THE HAY?**

SUMMARY:

COMMUNITY DNA METABARCODING

- High target DNA concentration

- Deterministic PCR
- Very high repeatability
- No need to use PCR replicates
- A few ecological replicates

- Quantitative value of N. of reads
- Relative abundance of reads reflects relative biomass of MOTUs

- No contamination issues
- Negative controls will be OK

EXTRA-ORGANISMAL DNA METABARCODING

- Very low target DNA concentration

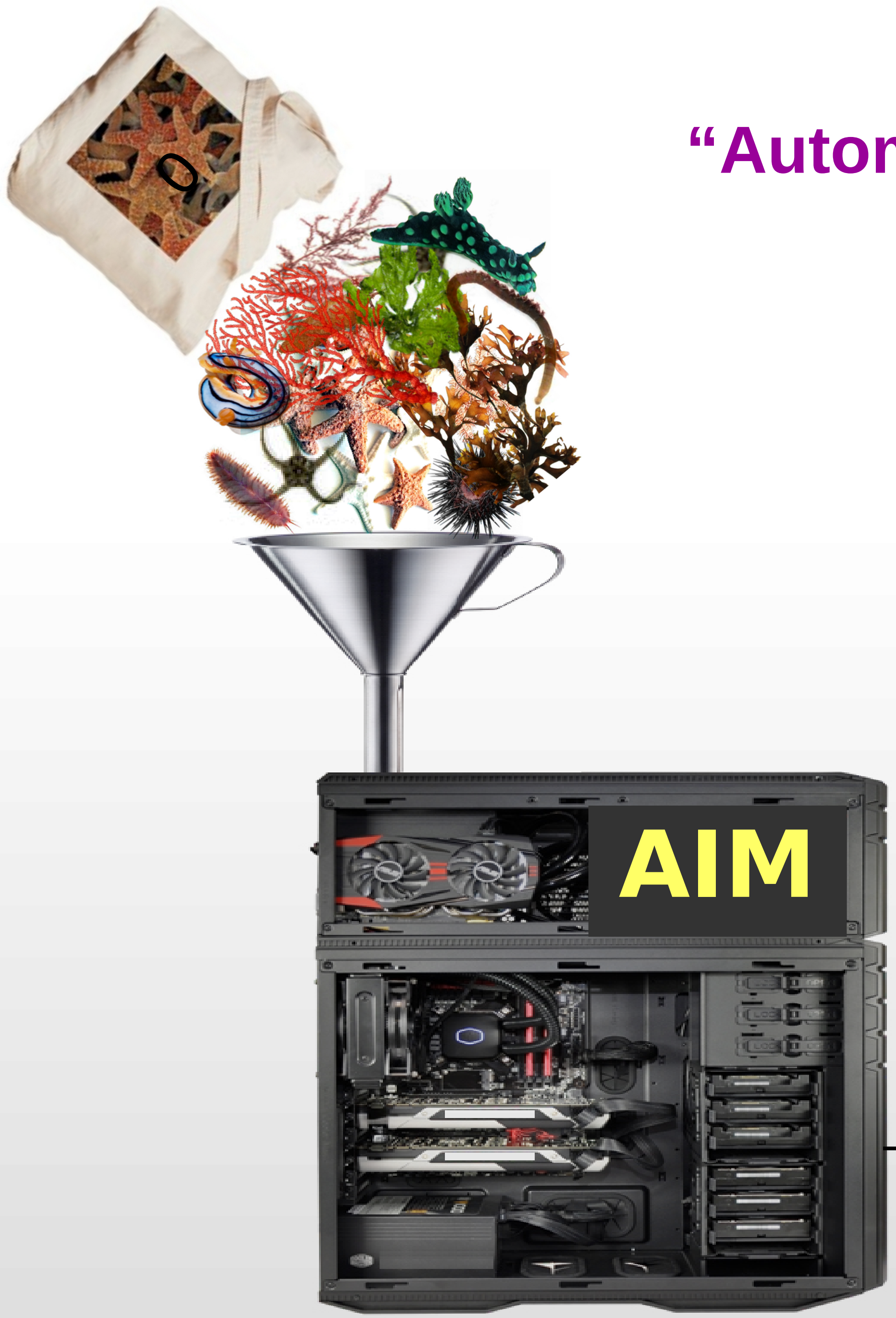
- Stochastic PCR
- Very low repeatability
- Need for many PCR replicates
- As many as possible ecol. replicates

- Stochasticity in N. of reads
- Use percentage of positive PCRs as a measure of MOTU abundance

- Very sensitive to contaminations
- Problematic MOTUs in negatives

COMMUNITY DNA METABARCODING

"Automated Inventory Machine"



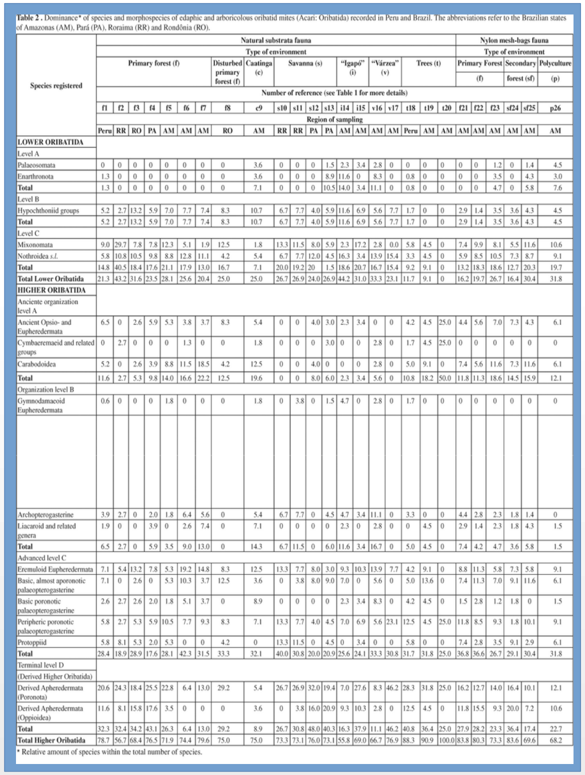
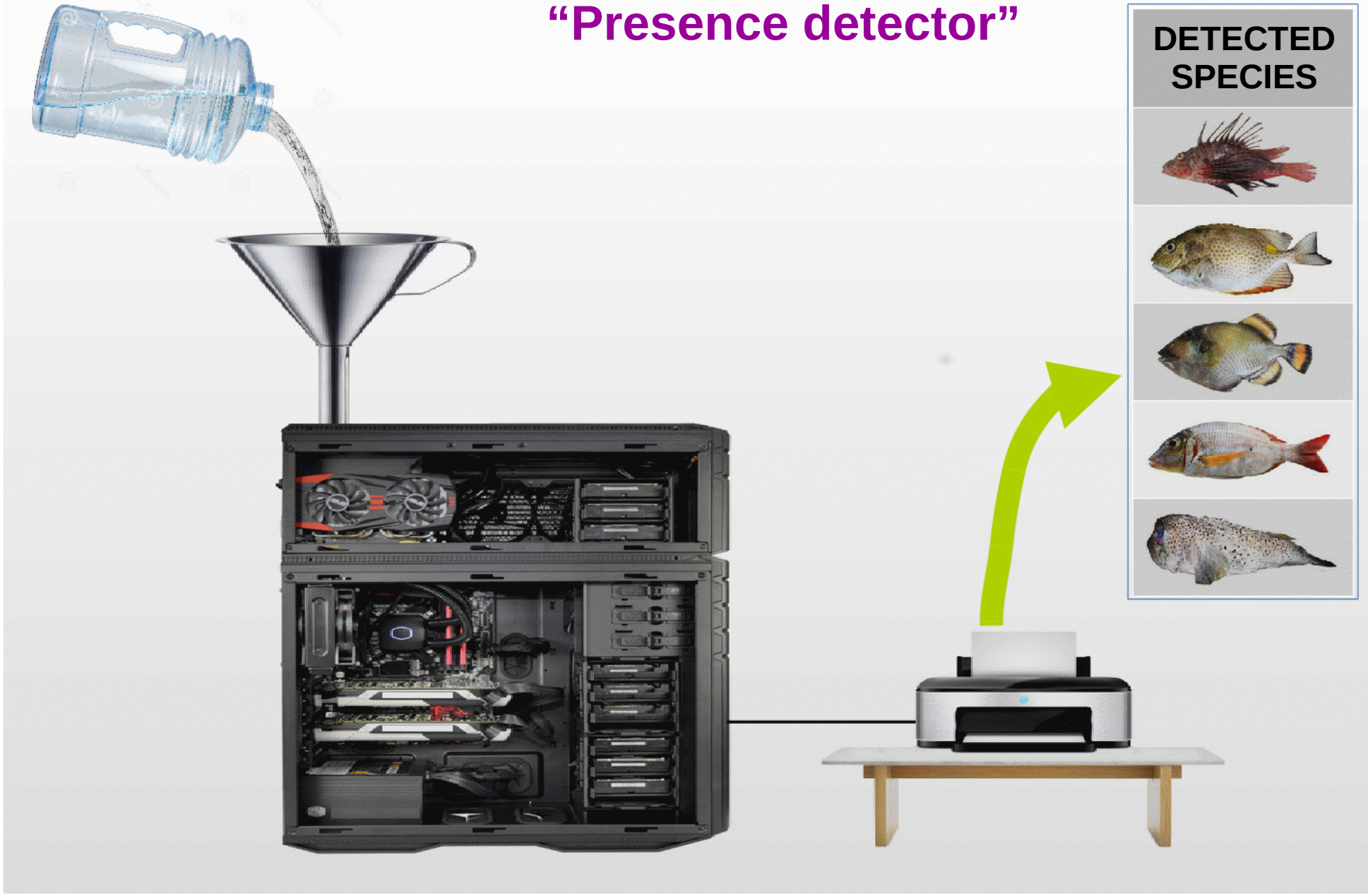


Table 2 . Dominance* of species and morphospecies of edaphic and arboricolous oribatid mites (Acari: Oribatida) recorded in Peru and Brazil. The abbreviations refer to the Brazilian states of Amazonas (AM), Pará (PA), Roraima (RR) and Rondônia (RO).

Species registered	Natural substrata fauna																				Nylon mesh-bags fauna					
Type of environment	Primary forest (f)							Disturbed primary forest (f)	Caatinga (c)	Savanna (s)				"Igapó" (i)		"Várzea" (v)		Trees (t)			Primary Forest (f)			Secondary forest (sf)		Polyculture (p)
Number of reference (see Table 1 for more details)	f1	f2	f3	f4	f5	f6	f7	f8	c9	s10	s11	s12	s13	i14	i15	v16	v17	t18	t19	t20	f21	f22	f23	sf24	sf25	p26
Region of sampling	Peru	RR	RO	PA	AM	AM	AM	RO	AM	RR	RR	PA	PA	AM	AM	AM	AM	Peru	AM	AM	AM	AM	AM	AM	AM	AM
LOWER ORIBATIDA																										
Level A																										
Palaeosomata	0	0	0	0	0	0	0	0	3.6	0	0	0	1.5	2.3	3.4	2.8	0	0	0	0	0	0	1.2	0	1.4	4.5
Enarthronota	1.3	0	0	0	0	0	0	0	3.6	0	0	0	8.9	11.6	0	8.3	0	0.8	0	0	0	0	3.5	0	4.3	3.0
Total	1.3	0	0	0	0	0	0	0	7.1	0	0	0	10.5	14.0	3.4	11.1	0	0.8	0	0	0	0	4.7	0	5.8	7.6
Level B																										
Hypochthoniid groups	5.2	2.7	13.2	5.9	7.0	7.7	7.4	8.3	10.7	6.7	7.7	4.0	5.9	11.6	6.9	5.6	7.7	1.7	0	0	2.9	1.4	3.5	3.6	4.3	4.5
Total	5.2	2.7	13.2	5.9	7.0	7.7	7.4	8.3	10.7	6.7	7.7	4.0	5.9	11.6	6.9	5.6	7.7	1.7	0	0	2.9	1.4	3.5	3.6	4.3	4.5
Level C																										
Mixonomata	9.0	29.7	7.8	7.8	12.3	5.1	1.9	12.5	1.8	13.3	11.5	8.0	5.9	2.3	17.2	2.8	0.0	5.8	4.5	0	7.4	9.9	8.1	5.5	11.6	10.6
Nothroidea *s.l.*	5.8	10.8	10.5	9.8	8.8	12.8	11.1	4.2	5.4	6.7	7.7	12.0	4.5	16.3	3.4	13.9	15.4	3.3	4.5	0	5.9	8.5	10.5	7.3	8.7	9.1
Total	14.8	40.5	18.4	17.6	21.1	17.9	13.0	16.7	7.1	20.0	19.2	20	1.5	18.6	20.7	16.7	15.4	9.2	9.1	0	13.2	18.3	18.6	12.7	20.3	19.7
Total Lower Oribatida	21.3	43.2	31.6	23.5	28.1	25.6	20.4	25.0	25.0	26.7	26.9	24.0	26.9	44.2	31.0	33.3	23.1	11.7	9.1	0	16.2	19.7	26.7	16.4	30.4	31.8
HIGHER ORIBATIDA																										
Anciente organization level A																										
Ancient Opsio- and Eupheredermata	6.5	0	2.6	5.9	5.3	3.8	3.7	8.3	5.4	0	0	4.0	3.0	2.3	3.4	0	0	4.2	4.5	25.0	4.4	5.6	7.0	7.3	4.3	6.1
Cymbaeremaeid and related groups	0	2.7	0	0	0	1.3	0	0	1.8	0	0	0	3.0	0	0	2.8	0	1.7	4.5	25.0	0	0	0	0	0	0
Carabodoidea	5.2	0	2.6	3.9	8.8	11.5	18.5	4.2	12.5	0	0	4.0	0	0	0	2.8	0	5.0	9.1	0	7.4	5.6	11.6	7.3	11.6	6.1
Total	11.6	2.7	5.3	9.8	14.0	16.6	22.2	12.5	19.6	0	0	8.0	6.0	2.3	3.4	5.6	0	10.8	18.2	50.0	11.8	11.3	18.6	14.5	15.9	12.1
Organization level B																										
Gymnodamaeoid Eupheredermata	0.6	0	0	0	1.8	0	0	0	1.8	0	3.8	0	1.5	4.7	0	2.8	0	1.7	0	0	0	0	0	0	0	0
Archopterogasterine	3.9	2.7	0	2.0	1.8	6.4	5.6	0	5.4	6.7	7.7	0	4.5	4.7	3.4	11.1	0	3.3	0	0	4.4	2.8	2.3	1.8	1.4	0
Liacaroid and related genera	1.9	0	0	3.9	0	2.6	7.4	0	7.1	0	0	0	0	2.3	0	2.8	0	0	4.5	0	2.9	1.4	2.3	1.8	4.3	1.5
Total	6.5	2.7	0	5.9	3.5	9.0	13.0	0	14.3	6.7	11.5	0	6.0	11.6	3.4	16.7	0	5.0	4.5	0	7.4	4.2	4.7	3.6	5.8	1.5
Advanced level C																										
Eremuloid Eupheredermata	7.1	5.4	13.2	7.8	5.3	19.2	14.8	8.3	12.5	13.3	7.7	8.0	3.0	9.3	10.3	13.9	7.7	4.2	9.1	0	8.8	11.3	5.8	7.3	5.8	9.1
Basic, almost aporonotic palaeopterogasterine	7.1	0	2.6	0	5.3	10.3	3.7	12.5	3.6	0	3.8	8.0	9.0	7.0	0	5.6	0	5.0	13.6	0	7.4	11.3	7.0	9.1	11.6	6.1
Basic poronotic palaeopterogasterine	2.6	2.7	2.6	2.0	1.8	5.1	3.7	0	8.9	0	0	0	0	2.3	3.4	8.3	0	4.2	4.5	0	1.5	2.8	1.2	1.8	0	1.5
Peripheric poronotic palaeopterogasterine	5.8	2.7	5.3	5.9	10.5	7.7	9.3	8.3	7.1	13.3	7.7	4.0	4.5	7.0	6.9	5.6	23.1	12.5	4.5	25.0	11.8	8.5	9.3	1.8	10.1	9.1
Protoppiid	5.8	8.1	5.3	2.0	5.3	0	0	4.2	0	13.3	11.5	0	4.5	0	3.4	0	0	5.8	0	0	7.4	2.8	3.5	9.1	2.9	6.1
Total	28.4	18.9	28.9	17.6	28.1	42.3	31.5	33.3	32.1	40.0	30.8	20.0	20.9	25.6	24.1	33.3	30.8	31.7	31.8	25.0	36.8	36.6	26.7	29.1	30.4	31.8
Terminal level D																										
(Derived Higher Oribatida)																										
Derived Apheredermata (Poronota)	20.6	24.3	18.4	25.5	22.8	6.4	13.0	29.2	5.4	26.7	26.9	32.0	19.4	7.0	27.6	8.3	46.2	28.3	31.8	25.0	16.2	12.7	14.0	16.4	10.1	12.1
Derived Apheredermata (Oppioidea)	11.6	8.1	15.8	17.6	3.5	0	0	0	3.6	0	3.8	16.0	20.9	9.3	10.3	2.8	0	12.5	4.5	0	11.8	15.5	9.3	20.0	7.2	10.6
Total	32.3	32.4	34.2	43.1	26.3	6.4	13.0	29.2	8.9	26.7	30.8	48.0	40.3	16.3	37.9	11.1	46.2	40.8	36.4	25.0	27.9	28.2	23.3	36.4	17.4	22.7
Total Higher Oribatida	78.7	56.7	68.4	76.5	71.9	74.4	79.6	75.0	75.0	73.3	73.1	76.0	73.1	55.8	69.0	66.7	76.9	88.3	90.9	100.0	83.8	80.3	73.3	83.6	69.6	68.2

* Relative amount of species within the total number of species.

Extra-organismal DNA METABARCODING



PRIMER DESIGN

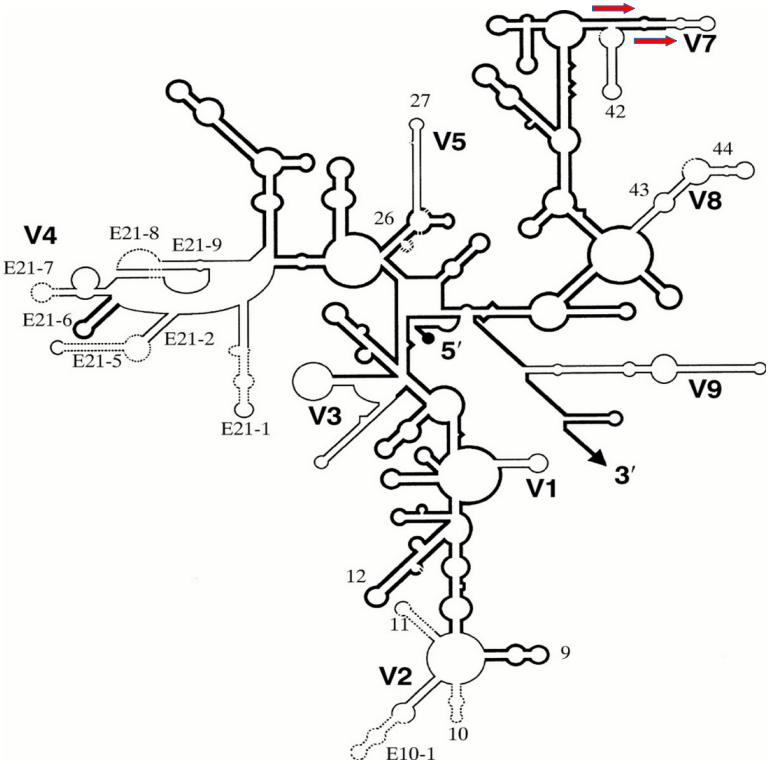
METABARCODING MARKER SELECTION:

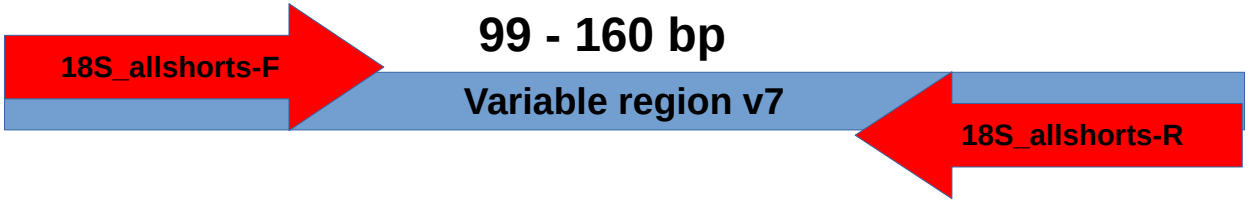
- A region of enough variability, surrounded by universally conserved regions

- Must be the right size for High-Throughput sequencing (<400 bp in Illumina)

- Ideally a multiple-copy, high-abundance genomic region

EXAMPLE: 18S rRNA v7 region





18S_allshorts-F: 5'-TTTGTCTGSTTAATTSCG-3'
18S_allshorts-R: 5'-GCAATAACAGGTCTGTG-3

Designed by P. Taberlet & E. Coissac for Eukarya (Guardiola et al. 2015 PLoS ONE 10, e0139633)

METABARCODING MARKERS

METABARCODING MARKER SELECTION:

Important issues to consider:

- Variability in length (coding vs non-coding regions)
- Taxon resolution (enough variability to distinguish among taxa?)
- Primer specificity (do the primers amplify only the target taxa?)
- Primer universality (do the primers amplify ALL the target taxa?)
- Intracellular localization (nuclear, mitochondrial, chloroplastidic?)
- Genomic copy number (is it conserved among taxa?)

PRIMER DESIGN AND PCR

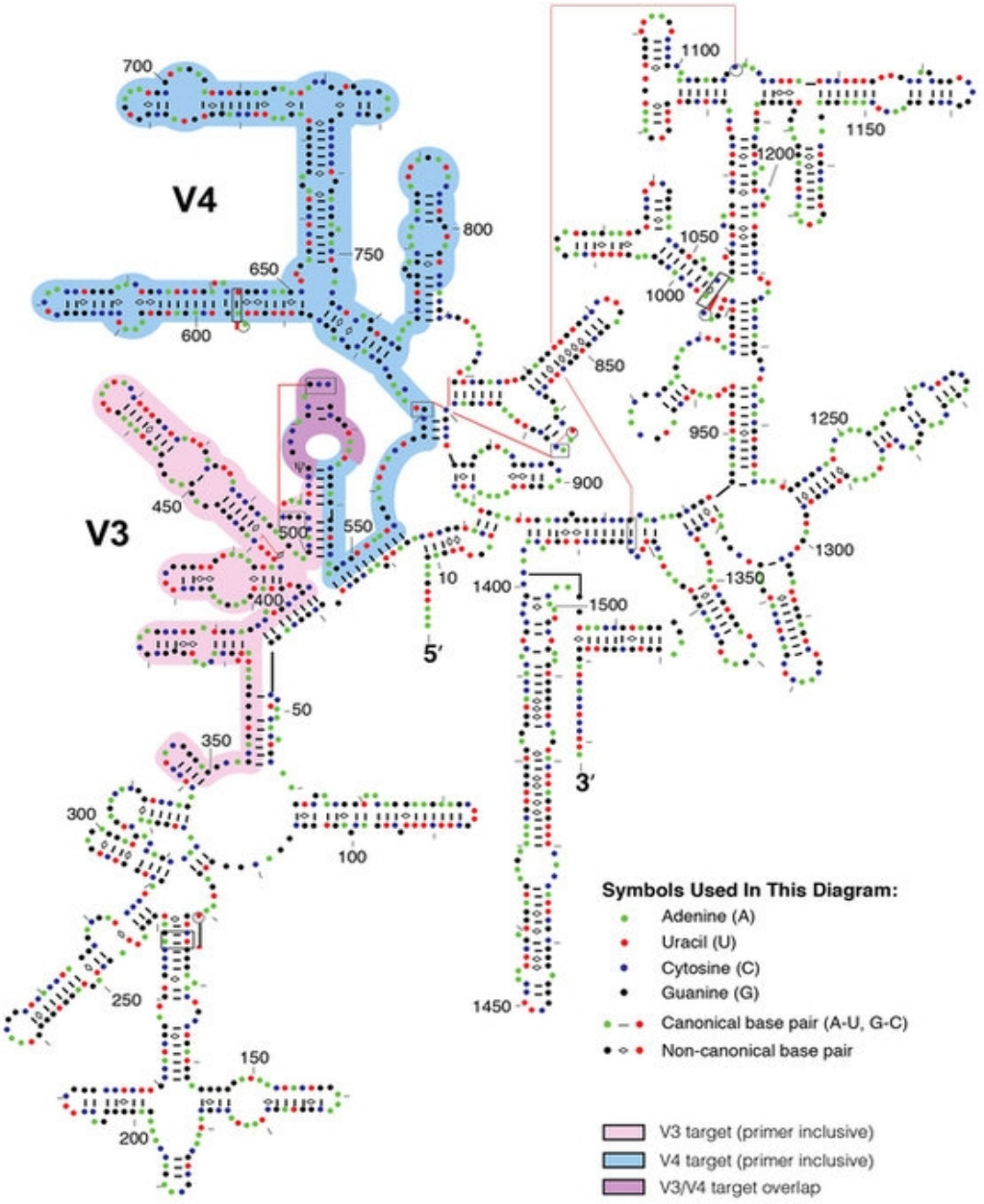
METABARCODING MARKER SELECTION

Different metabarcoding markers may be used for different taxonomic groups

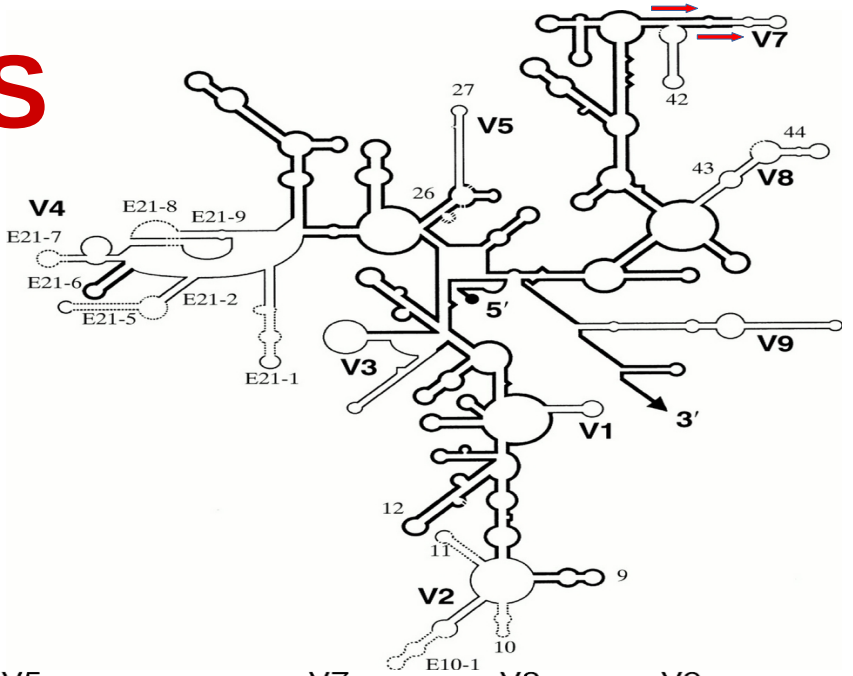
The taxonomic target depends on the universality/specificity of the primers

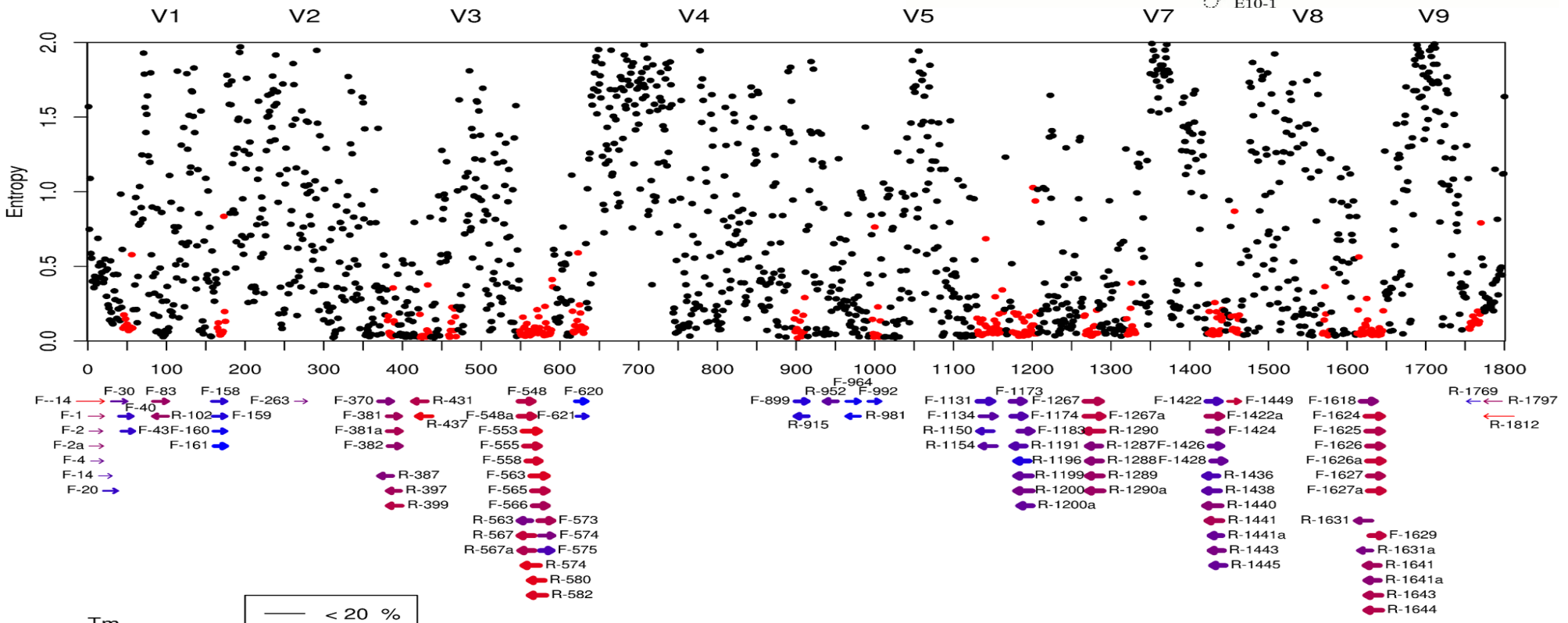
PROKARYOTA:	**rRNA 16S**
METAZOA:	**nuclear rRNA18S, mitochondrial COI**
Vertebrates:	**mitochondrial rRNA12S**
“Invertebrates”:	**mitochondrial rRNA16S**
Arthropoda:	**mitochondrial COI**
FUNGI:	**nuclear ITS**
PHOTOSYNTHETIC ORGANISMS:	**choroplastic trnL, matK,rbcL, nuclear ITS**

PROKARYOTIC rRNA 16S



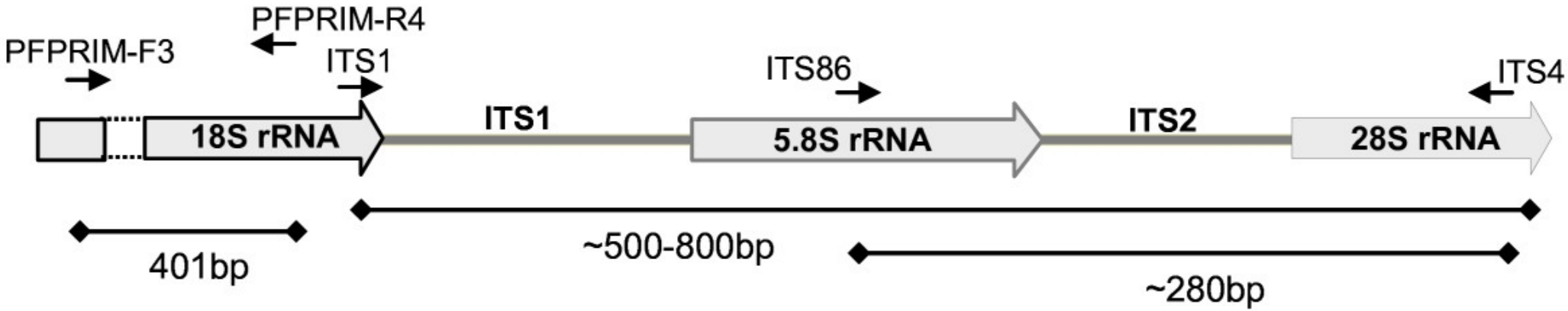
EUKARYOTIC rRNA 18S



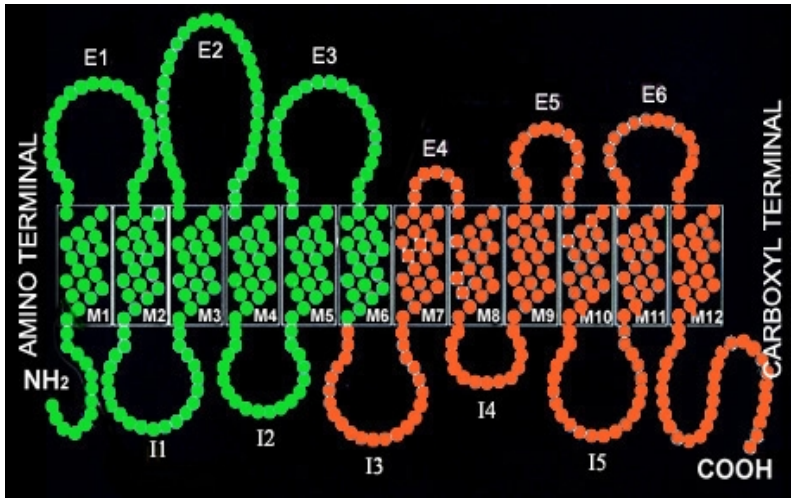


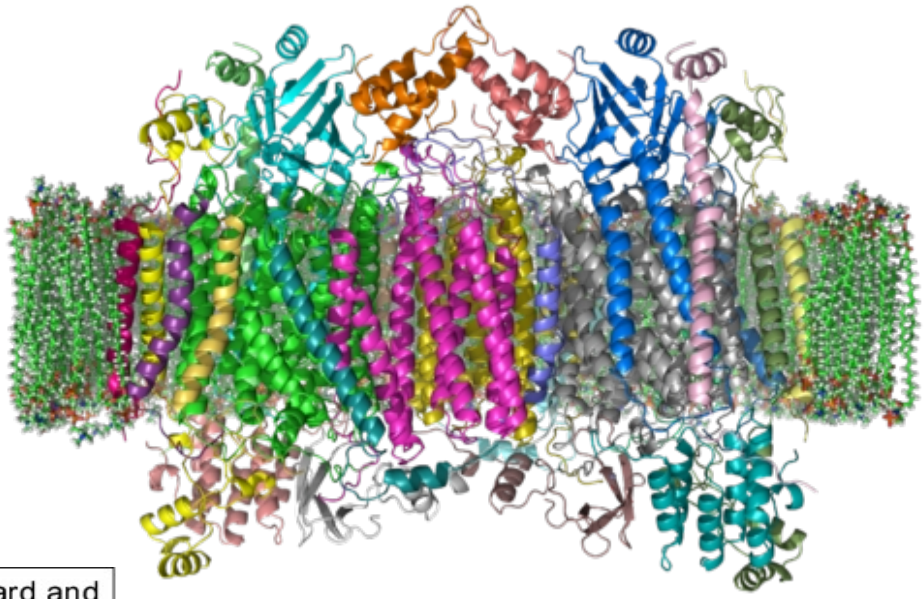
Hadziavdic et al. 2014 PloS ONE 9, e87624

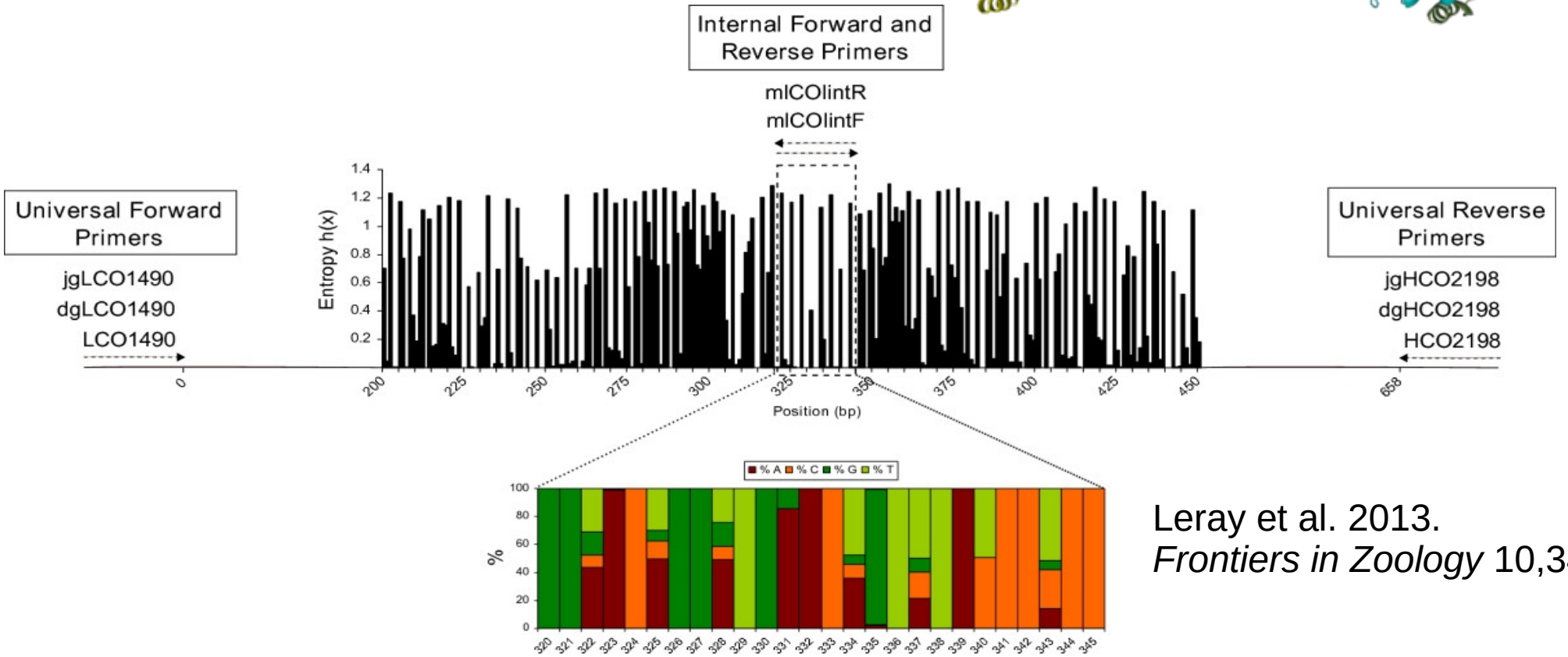
INTERNAL TRANSCRIPT SPACERS ITS1 & ITS2



CYTOCHROME *c* OXIDASE SUBUNIT I







Leray et al. 2013.
Frontiers in Zoology 10,34

PRIMER DESIGN AND PCR

FREQUENTLY USED METABARCODING MARKERS:

Target	Gene/ region	Reference	Data bases
Bacteria	16S	Sogin *et al.* (2006)	RDP, Greengenes, SILVA
Archaea	16S	Sogin *et al.* (2006)	RDP, Greengenes, SILVA
Fungi	ITS	Epp *et al.* (2012), Schoch *et al.* (2012)	UNITE, GenBank, BOLD (few records)
	18S	Not recommended (Schoch *et al.* 2012)	SILVA
Protists	18S	Pawlowski *et al.* (2012)	SILVA
	ITS	Pawlowski *et al.* (2012)	GenBank
	CO1	Pawlowski *et al.* (2012)	BOLD
Meiofauna	CO1	Hebert *et al.* (2003)	BOLD
	18S	Deagle *et al.* (2014)	GenBank
Macrofauna	CO1	Hebert *et al.* (2003)	BOLD
	16S	Epp *et al.* (2012), Deagle *et al.* (2014)	GenBank
	12S	Epp *et al.* (2012), Deagle *et al.* (2014)	GenBank
	18S	Deagle *et al.* (2014)	GenBank
Plants	matK + rbcL	Hollingsworth *et al.* (2009)	GenBank, BOLD (few records)
	ITS	China Plant Barcode of Life Group (2011)	GenBank

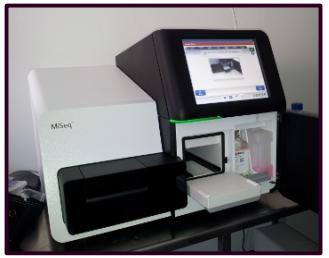
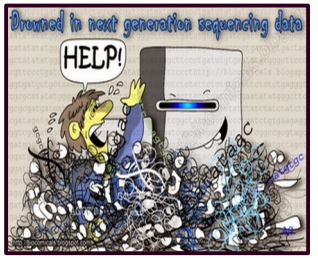
THE METABARCODING WORKFLOW

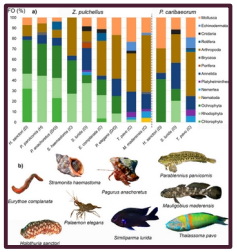
1. Experimental design, sampling & sample preservation
2. Sample pre-processing
3. DNA extraction
4. DNA amplification (metabarcoding primers)
5. High-throughput sequencing
6. Bioinformatic data analysis
    - 6.1. Demultiplexing and quality control
    - 6.2. Clustering sequences into MOTUs
        (Molecular Operational Taxonomic Units)
    - 6.3. Taxonomic assignment
    - 6.4. Refinement of the final dataset
7. Ecological & statistical data treatment

THE METABARCODING WORKFLOW





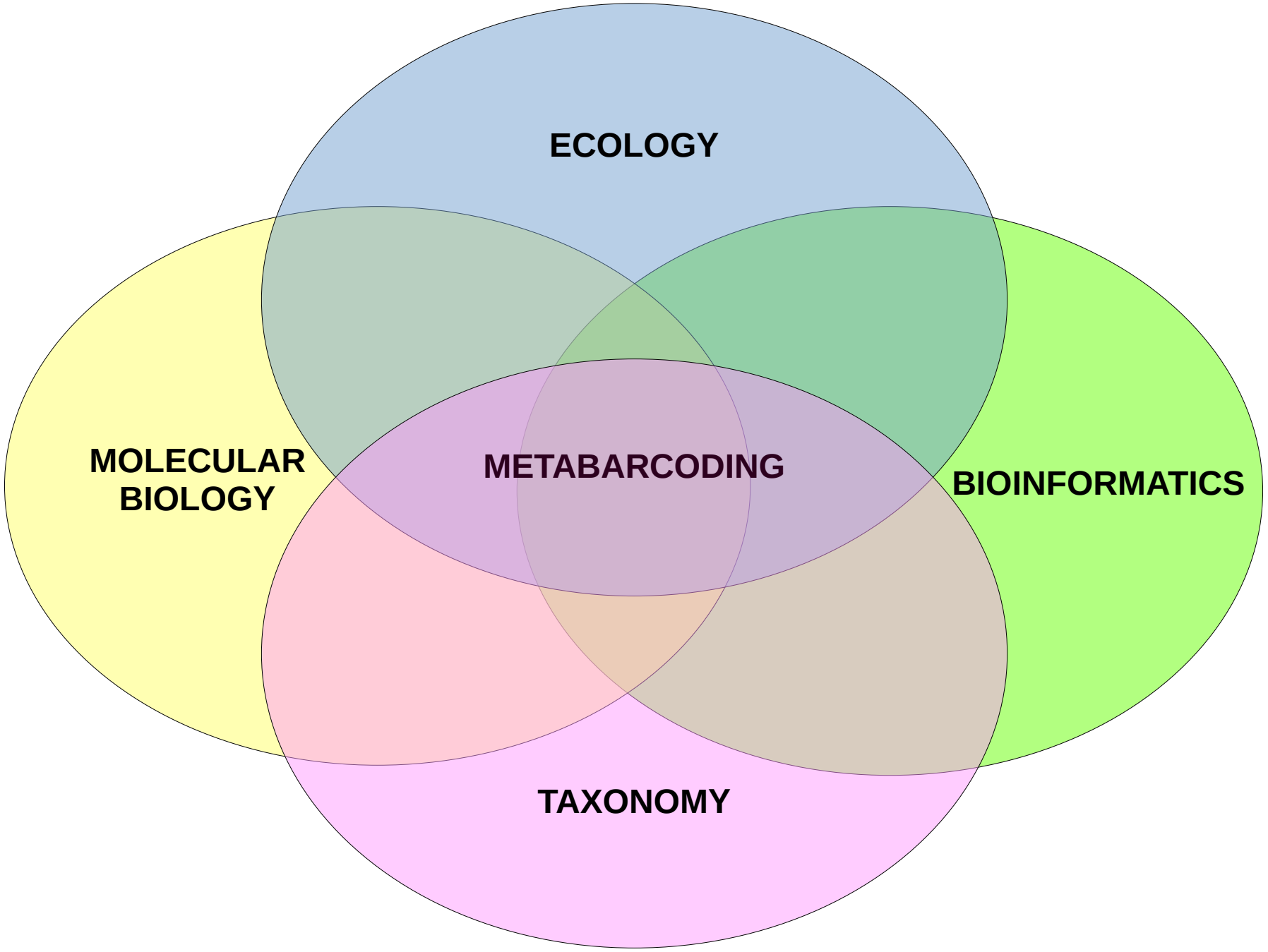






Sampling → Sample preprocessing → DNA extraction → Primer choice & PCR → HT sequencing → Bioinformatic pipeline → Results

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

DISCIPLINES AND SKILLS



DEMULTIPLEXING AND QUALITY FILTERING

Typical MiSeq output: 15 - 25 million raw reads (sequences)

Bioinformatic tools must:

- **Remove low-quality sequences**
- **Assign every sequence to its sample (demultiplexing)**
- **Group all equal sequences (dereplicating), keeping information of the samples where they occur**
- **Delete chimeric sequences**

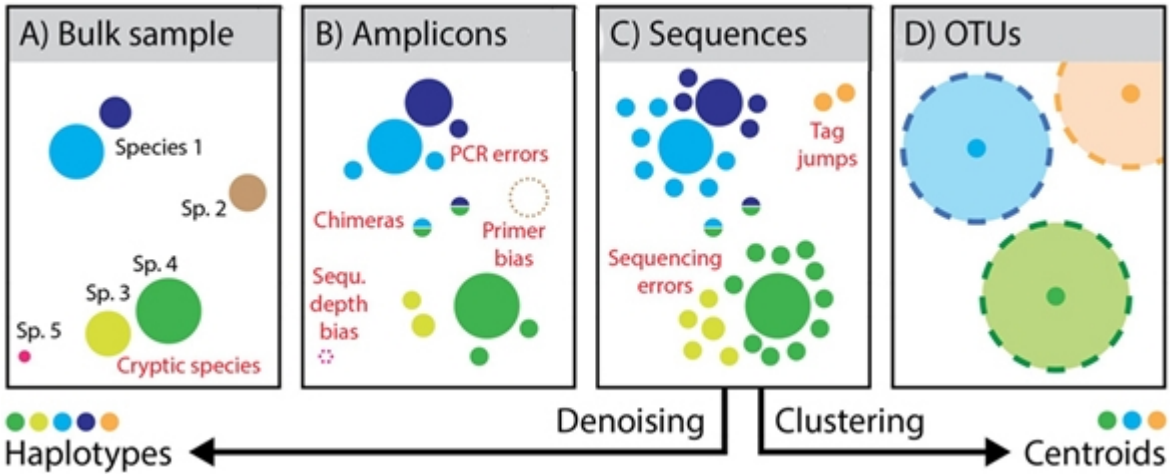
Still a huge amount of unique sequences!!!

Typically: 1 - 4 million unique sequences

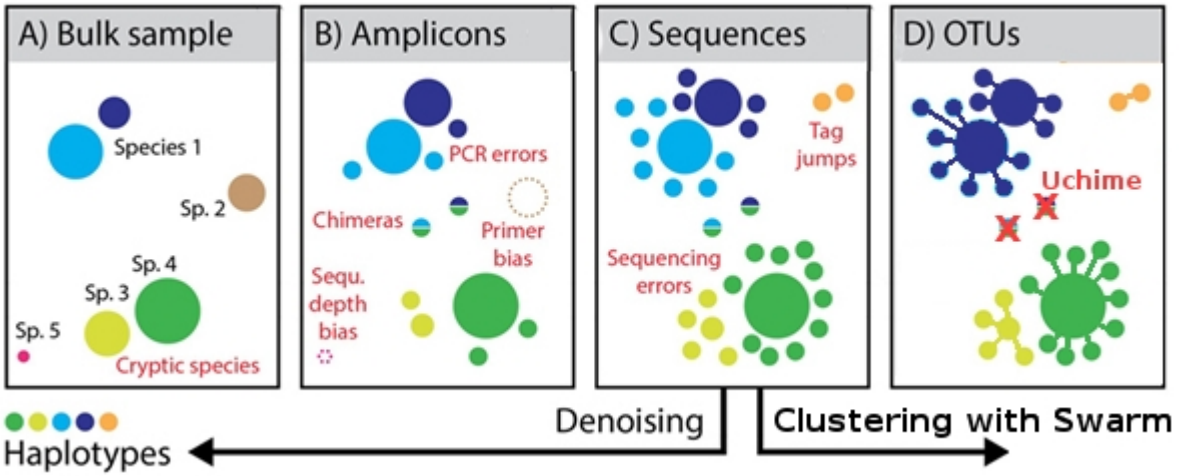
Due to:
- **diversity of species in the samples**
- **natural intraspecific variability**
- **random PCR errors**
- **random sequencing errors**

WHAT TO DO WITH ERROR-CONTAINING READS?

CLUSTERING vs DENOISING



Constant vs variable cutoff clustering



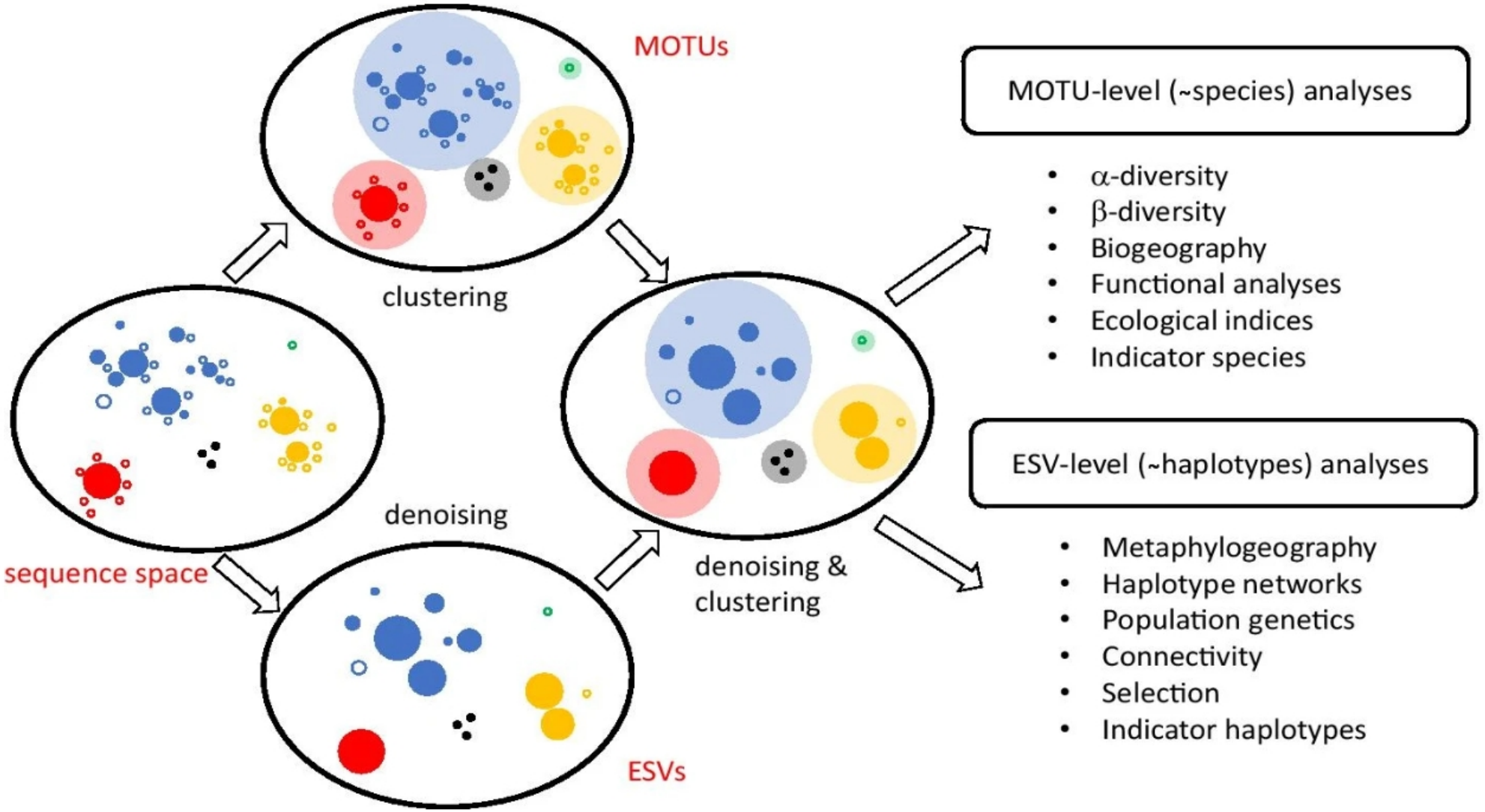
INTEGRATING CLUSTERING & DENOISING

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

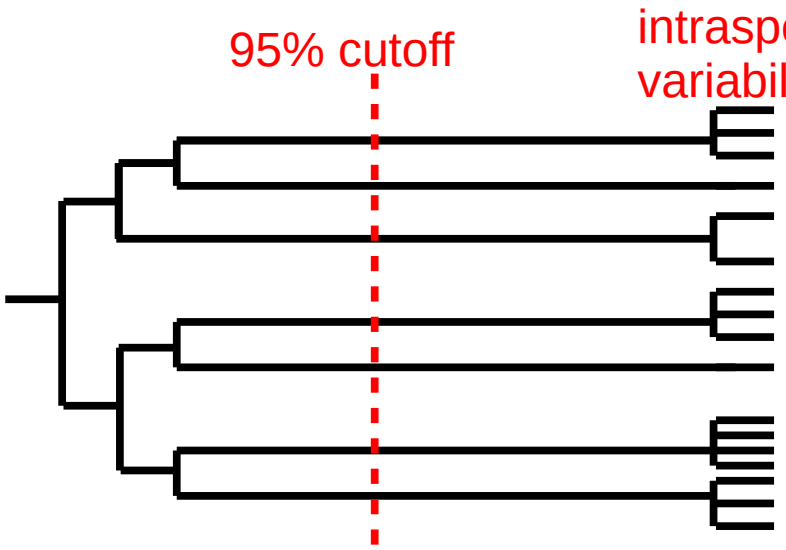
https://doi.org/10.1101/2021.01.08.425760

- Denoising AND clustering
(not just denoising OR just clustering)



MOTU CLUSTERING

Theory: Constant ID % cutoff

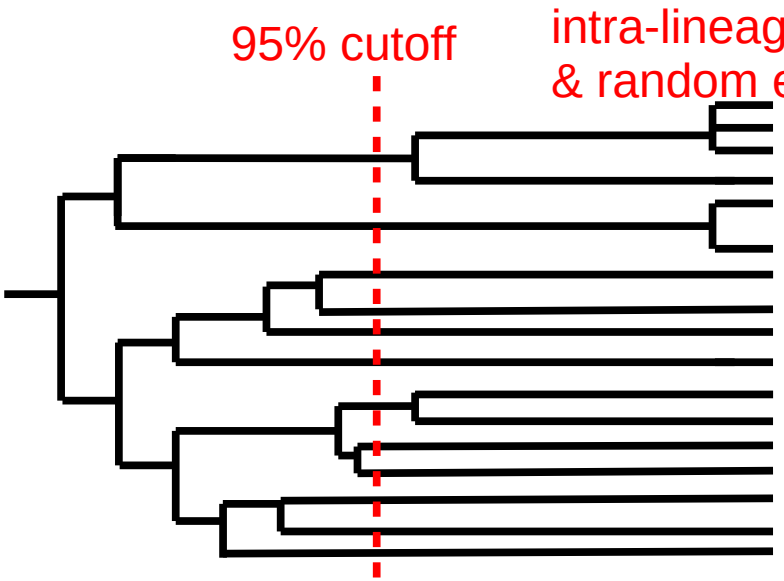


17 different sequences

7 MOTUs (species)

Constant cutoff clustering software:
- **UCLUST** (Edgar, 2010)
- **VSEARCH** (Rognes et al 2016)
- **SUMACLUST** (Mercier et al., 2013)

Reality: Variable cutoff



12 MOTUs (species?)

Bayesian Clustering Algorithms:
- **CROP** (Hao, Jiang & Chen, 2011)

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

BIOINFORMATICS PIPELINE: TAXONOMIC ASSIGNMENT OF MOTUs

Ideal: assignment by direct match (ID > 97%)

Marthasterias glacialis voucher LMBRUI7-001 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial
Sequence ID: gb|KF369150.1| **Length:** 658 **Number of Matches:** 1

Range 1: 346 to 658 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
556 bits(616)	7e-155	311/313(99%)	0/313(0%)	Plus/Plus

```
Query  1    TCTCTCTAGCGGATTAGCCCATGCTGGAGGATCAGTGGACCTCGCTATATTCTCTCTTCA  60
            ||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||
Sbjct  346  TCTCTCTAGCGGATTAGCCCATGCTGGAGGATCAGTAGACCTCGCTATATTCTCTCTTCA  405

Query  61   CTTAGCTGGTGCCTCATCAATCCTTGCCTCCATAAAATTTATCACAACAATAATTAATAT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  406  CTTAGCTGGTGCCTCATCAATCCTTGCCTCCATAAAATTTATCACAACAATAATTAATAT  465

Query  121  GCGAACTCCTGGCATGTCCTTTGACCGCTTACCCCTATTTGTTTGATCAGTCTTTGTAAC  180
            ||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||
Sbjct  466  GCGAACTCCTGGCATGTCCTTTGACCGCTTACCCCTATTTGTTTGATCGGTCTTTGTAAC  525
```

- BOLD identification
- BLAST against Genbank
- Use local databases (obtained from *in silico* PCR)

What to do with the Unmatched? Phylogenetic Assignment

- **Statistical Assignment Package (SAP)** Bayesian phylogenetics or constrained-NJ (Munch et al., 2008)
- **OBItools - Ecotag** Supervised classification algorithm (Boyer et al., 2016)
- **PROTAX** Probabilistic taxonomic assignment (Somervuo et al. 2016)


```
+Echinodermata (phylum) 98%
    +Echinoidea (class) 8%
    |   +Echinoida (order) 3%
    |   |   +Echinometridae (family) 3%
    |   |       +Echinometra (genus) 3%
    |   +Clypeasteroida (order) 2%
    |   |   +Clypeasteroida_incertaesedis (family) 2%
    |   |       +Clypeasteroida_incertaesedis_incertaesedis (genus) 2%
    |   +Echinothurioida (order) 1%
    |   |   +Echinothuriidae (family) 1%
    |   |       +Araeosoma (genus) 1%
    |   +Cidaroida (order) 2%
    |       +Cidaroida_incertaesedis (family) ~0%
    |       |   +Cidaroida_incertaesedis_incertaesedis (genus) ~0%
    |       +Cidaridae (family) 2%
    |           +Eucidaris (genus) 2%
    +Holothuroidea (class) 62%
    |   +Aspidochirotida (order) 45%
    |   |   +Holothuriidae (family) 45%
    |   |       +Holothuria (genus) ~0%
    |   |       +Actinopyga (genus) 2%
    |   |       +Bohadschia (genus) 38%
    |   +Elasipodida (order) 3%
    |   |   +Elpidiidae (family) 3%
    |   |       +Peniagone (genus) 3%
    |   +Dendrochirotida (order) 14%
    |       +Phyllophoridae (family) 6%
    |       |   +Pentamera (genus) 3%
    |       |   +Phyrella (genus) 1%
    |       +Dendrochirotida_incertaesedis (family) ~0%
    |       |   +Dendrochirotida_incertaesedis_incertaesedis (genus) ~0%
    |       +Cucumariidae (family) 4%
    |       |   +Staurocucumis (genus) 4%
    |       +Psolidae (family) 3%
    |           +Psolus (genus) 1%
    |           +Psolidium (genus) 2%
    +Ophiuroidea (class) 18%
    |   +Ophiurida (order) 12%
    |   |   +Ophiacanthidae (family) 10%
    |   |   |   +Ophiacantha (genus) 10%
    |   |   +Ophiotrichidae (family) 1%
    |   |       +Ophiomaza (genus) 1%
    |   +Euryalida (order) 5%
    |       +Euryalida_incertaesedis (family) 5%
    |           +Euryalida_incertaesedis_incertaesedis (genus) 5%
    +Asteroidea (class) ~0%
```

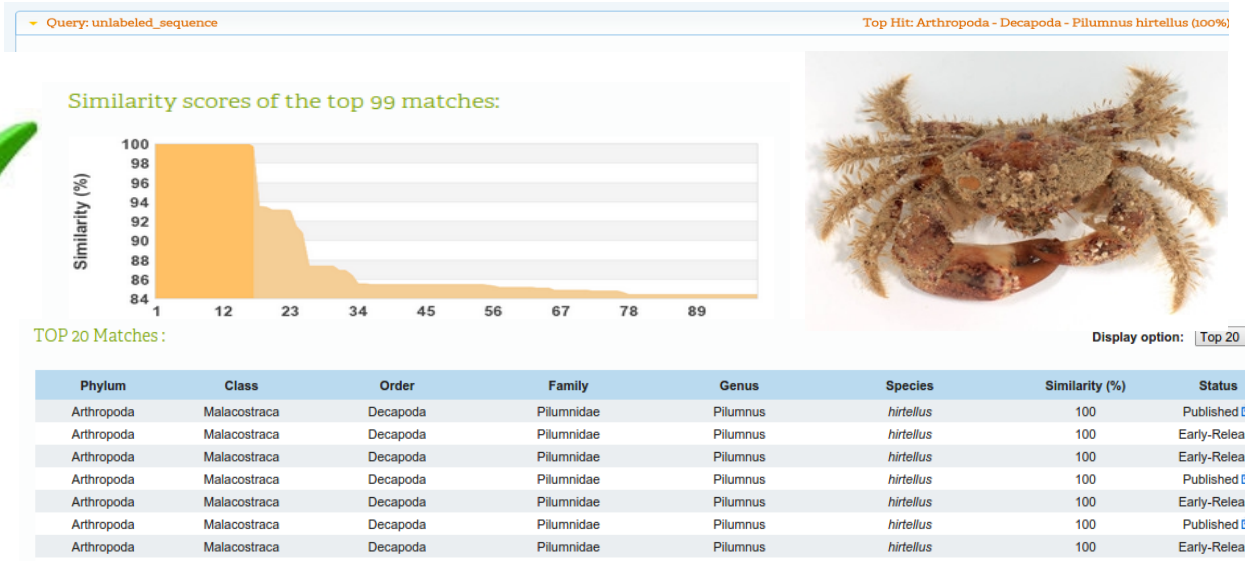
REMOTE vs LOCAL DATABASES

BLAST against Genbank

- Slow!
- Bad taxonomy!
- Misidentified entries!

BOLD identification

- Good, curated, taxonomy!
- Only relevant sequences
- Slow for too many seqs
- Only for conventional barcodes!



Phylum	Class	Order	Family	Genus	Species	Similarity (%)	Status
Arthropoda	Malacostraca	Decapoda	Pilumnidae	Pilumnus	*hirtellus*	100	Published
Arthropoda	Malacostraca	Decapoda	Pilumnidae	Pilumnus	*hirtellus*	100	Early-Release
Arthropoda	Malacostraca	Decapoda	Pilumnidae	Pilumnus	*hirtellus*	100	Early-Release
Arthropoda	Malacostraca	Decapoda	Pilumnidae	Pilumnus	*hirtellus*	100	Published
Arthropoda	Malacostraca	Decapoda	Pilumnidae	Pilumnus	*hirtellus*	100	Early-Release
Arthropoda	Malacostraca	Decapoda	Pilumnidae	Pilumnus	*hirtellus*	100	Published
Arthropoda	Malacostraca	Decapoda	Pilumnidae	Pilumnus	*hirtellus*	100	Early-Release

ECOTAG using a local database

- Usually fast (although the speed depends on the reference database size!)
- Errors and non-relevant sequences may be removed from the database
- Custom sequences may be added immediately
- Can be used for non-conventional markers and novel primer sets

OTHER TAXONOMIC ASSIGNMENT ALGORITHMS

BOLDigger / Boldigger II (Buchner & Leese 2020)

https://github.com/DominikBuchner/BOLDigger2

The BOLD identification tool allows only batches of 100 sequences that can be identified in one run.
BOLD's API does not grant access to private and early release data.

BOLDigger aims to solve this problem. It gives automated access to the identification engine and can also be used to download additional metadata for each sequence as well as helping to choose the top hit from the returned results.

mkLTG (Meglécz 2024)

Faster than ecotag, also based in the LCA approach to phylogenetic assignment, however with arbitrary thresholds for taxonomic ranks

It comes with an automatedly-generated database: COInr

› Biol Futur. 2023 Dec;74(4):369-375. doi: 10.1007/s42977-024-00201-x. Epub 2024 Feb 1.

mkLTG: a command-line tool for taxonomic assignment of metabarcoding sequences using variable identity thresholds

Emese Meglécz [1]

https://github.com/meglecz/mkLTG

https://github.com/meglecz/mkCOInr

OTHER TAXONOMIC ASSIGNMENT ALGORITHMS

MOLECULAR ECOLOGY RESOURCES WILEY

Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes

Jose S. Hleap[1,2,3] | Joanne E. Littlefair[1,4] | Dirk Steinke[5] | Paul D. N. Hebert[5] | Melania E. Cristescu[1]

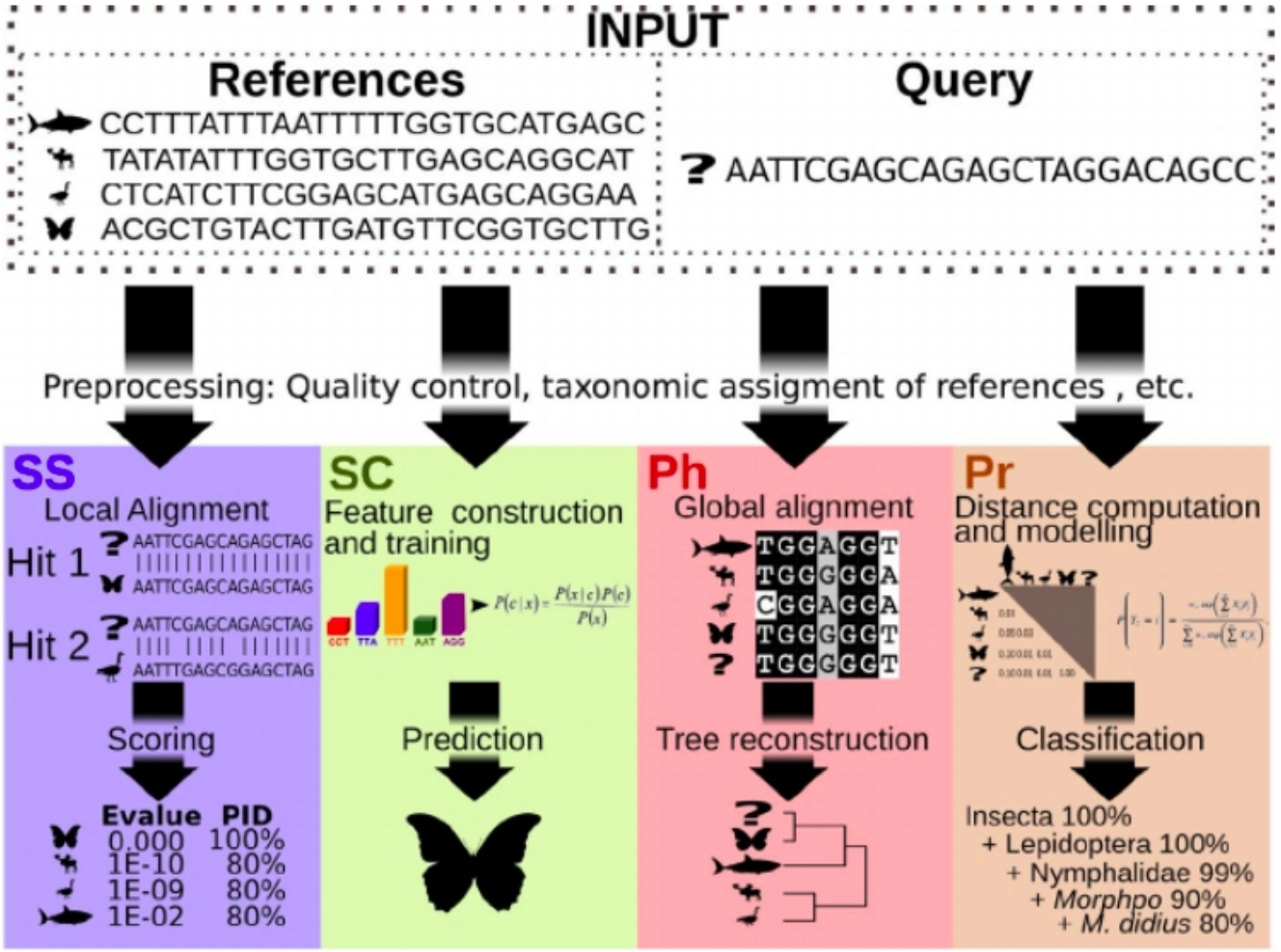


FIGURE 1 Overview of methodological approaches. Sequence similarity (SS) methods use local alignments to search for similarity between each query and the reference sequences. Sequence composition (SC) methods are trained by computing a k-mer frequency profile for each reference sequence, and then matching each query to this profile. Phylogenetic (Ph) methods use global alignments including (or placing) the query in a phylogenetic tree. Probabilistic (Pr) methods use a distance metric and then perform a hierarchical multinomial regression to estimate the certainty in the classification of each query at each taxonomic rank

BIOINFORMATICS PIPELINE: REFINING THE FINAL DATASET

- Removal of false positives from tag-switching

- Taxonomic clustering (cluster MOTUs assigned to same taxon)

- Removal of sequences derived from pseudogenes: LULU

- Blank corrections

- Removal of sequences derived from contaminations

- Rarefaction and re-sampling techniques for controlling variability in total reads

COMMONLY USED METABARCODING PIPELINES

- QIIME / QIIME2: https://qiime2.org/
- Mothur: https://www.mothur.org/
- OBITools: https://metabarcoding.org/obitools
- JAMP: https://github.com/VascoElbrecht/JAMP
- USEARCH: https://www.drive5.com/usearch/
- VSEARCH: https://github.com/torognes/vsearch
- DADA2: https://benjjneb.github.io/dada2/
- APSCALE: https://github.com/DominikBuchner/apscale
- MJOLNIR: https://github.com/uit-metabarcoding/MJOLNIR

SOME CURRENT APPLICATIONS

STUDIES ON INDIVIDUALS

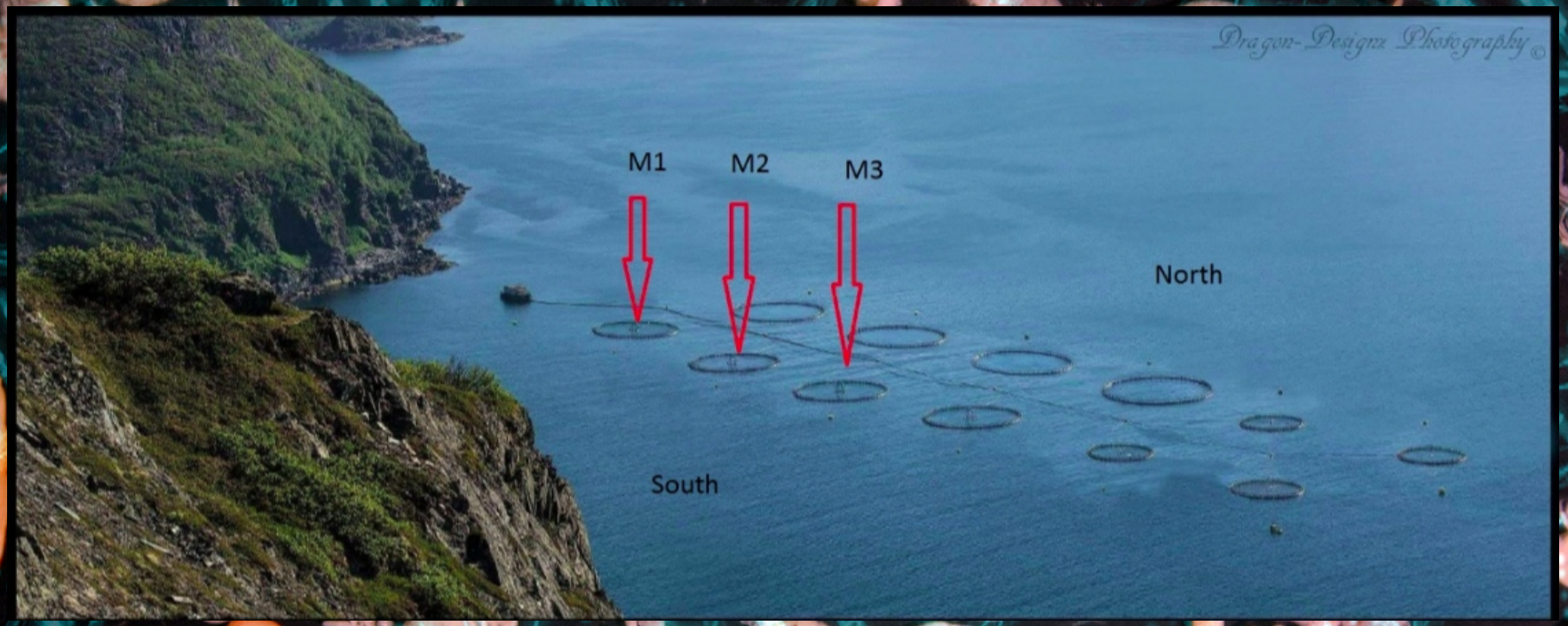
- Trophic relationships: gut contents / analyses of faeces
- Symbiomics: host/pathogen/symbiont interactions

STUDIES ON COMMUNITIES: SPATIAL PATTERNS

- Molecular characterization of biomes
- Population metagenomics / Metaphylogeography

STUDIES ON COMMUNITIES: TEMPORAL PATTERNS

- Community dynamics / annual cycles
- Settlement events detection / reproductive cycles
- Long-term shifts in community structure (including genetic structure)
- Distribution shifts / assessment of effects of Climate Change
- Detection / monitoring of introduced species
- Studies on palaeoecological records: palaeometabarcoding



AND EXAMPLE USING THE MJOLNIR PIPELINE