Metabarcoding and data processing pipelines

Presenter: Mafalda Galhardo

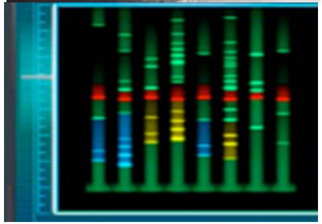
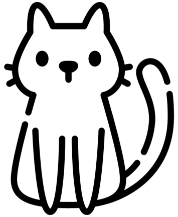
EnvMetaGen team

EnvMetaGen
CAPACITY BUILDING AT INBIO FOR RESEARCH AND INNOVATION USING
ENVIRONMENTAL METAGENOMICS

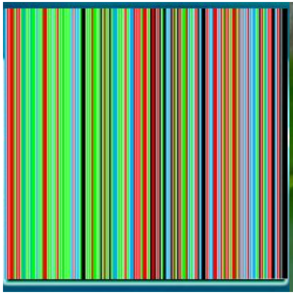
CIBIO

InBIO REDE DE INVESTIGAÇÃO
em Biodiversidade e Biologia Evolutiva

What's in your sample?

- Metabarcoding allows to identify species from complex DNA mixtures
- not dependent on taxonomy expertise


https://sciencetrends.com/butterfly-cocoon-stages-life-cycle-butterfly/

[Image credits: Sónia Ferreira and João Nunes]

Sample types

Many different sources!

-Faeces / gastrointestinal tract contents

-Animal tissues/parts (bone, hair, skin, feathers, blood,...)

-Bulks (e.g. invertebrates), museum specimens

-Waters / sludges

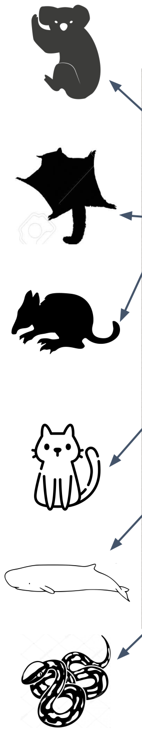
-Soils / sediments

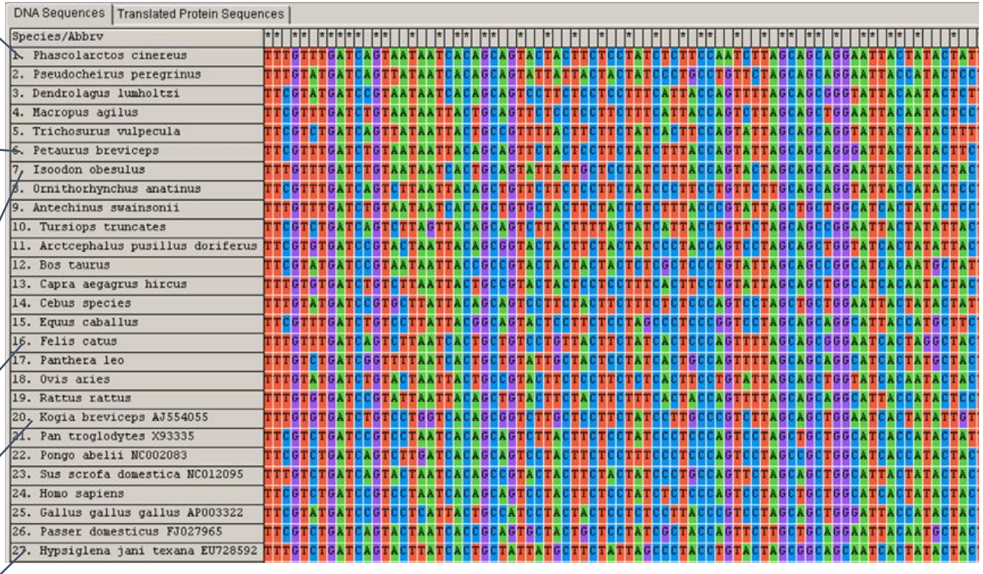
-Whatever you can think of...

[image credits: Bastian Egeter, Joana Paupério, Sónia Ferreira, Pedro Pires, Joana Veríssimo, Filipa Martins]

Metabarcoding

Species identification based on DNA sequence markers





Species/Abbrv

1. Phascolarctos cinereus
2. Pseudocheirus peregrinus
3. Dendrolagus lumholtzi
4. Macropus agilus
5. Trichosurus vulpecula
6. Petaurus breviceps
7. Isoodon obesulus
8. Ornithorhynchus anatinus
9. Antechinus swainsonii
10. Tursiops truncates
11. Arctcephalus pusillus doriferus
12. Bos taurus
13. Capra aegagrus hircus
14. Cebus species
15. Equus caballus
16. Felis catus
17. Panthera leo
18. Ovis aries
19. Rattus rattus
20. Kogia breviceps AJ554055
21. Pan troglodytes X93335
22. Pongo abelii NC002083
23. Sus scrofa domestica NC012095
24. Homo sapiens
25. Gallus gallus gallus AP003322
26. Passer domesticus FJ027965
27. Hypsiglena jani texana EU728592

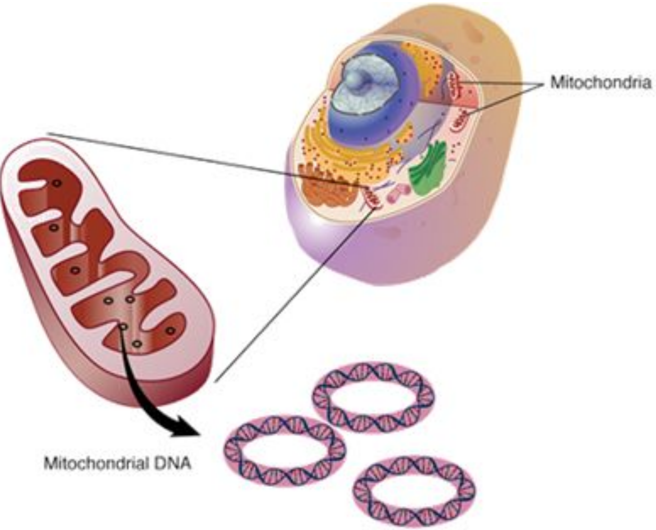
[Sequences image: https://doi.org/10.1016/j.fsigen.2013.12.007]

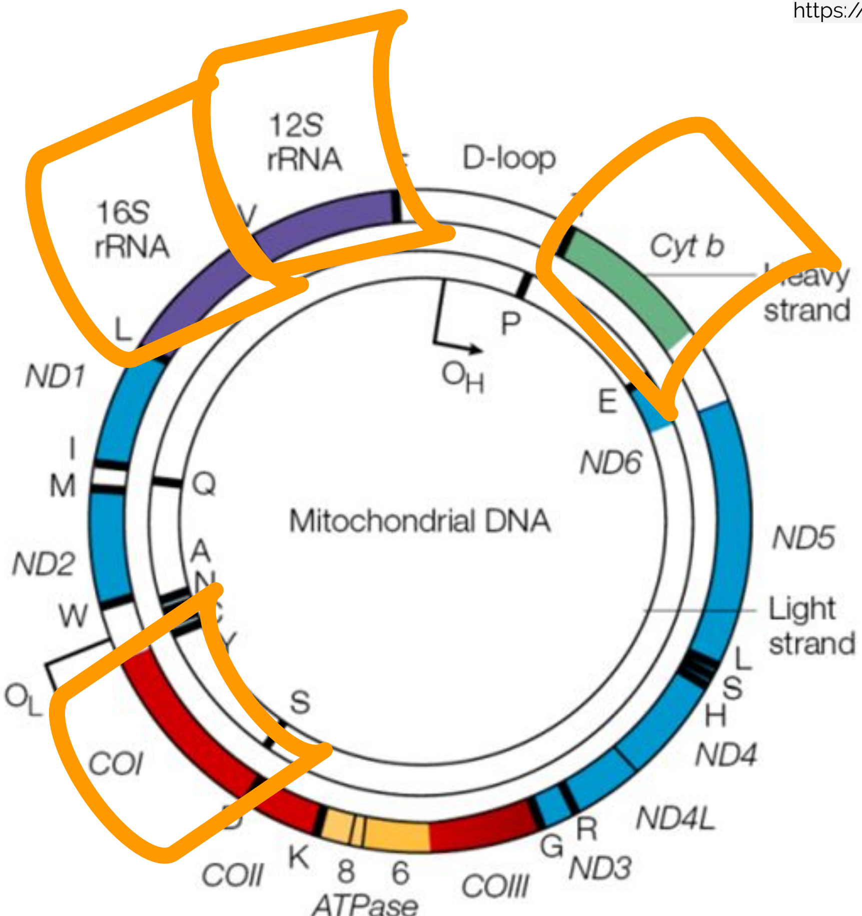
Relies on the existence of different DNA barcodes per species!

InBIO barcode initiative:
>7k specimens, >6k identified to species level
>2500 species, > 240 families

A common marker source: mitochondria

[Sources: https://en.wikipedia.org/wiki/Mitochondrial_DNA; Medical Cell Biology, Chapter 4, Figure 4–34, https://doi.org/10.1016/C2009-0-00212-8]

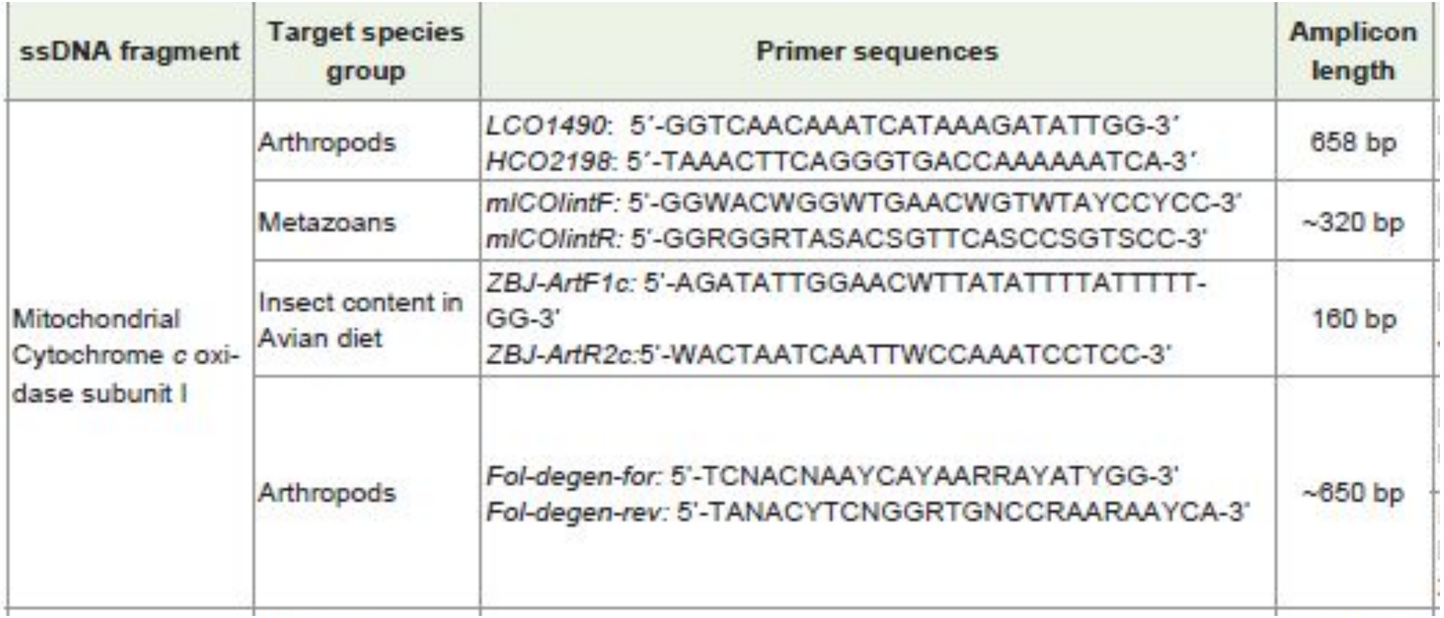




COI – cytochrome c oxidase I

(fragment of 648 bp)

Markers and their applications

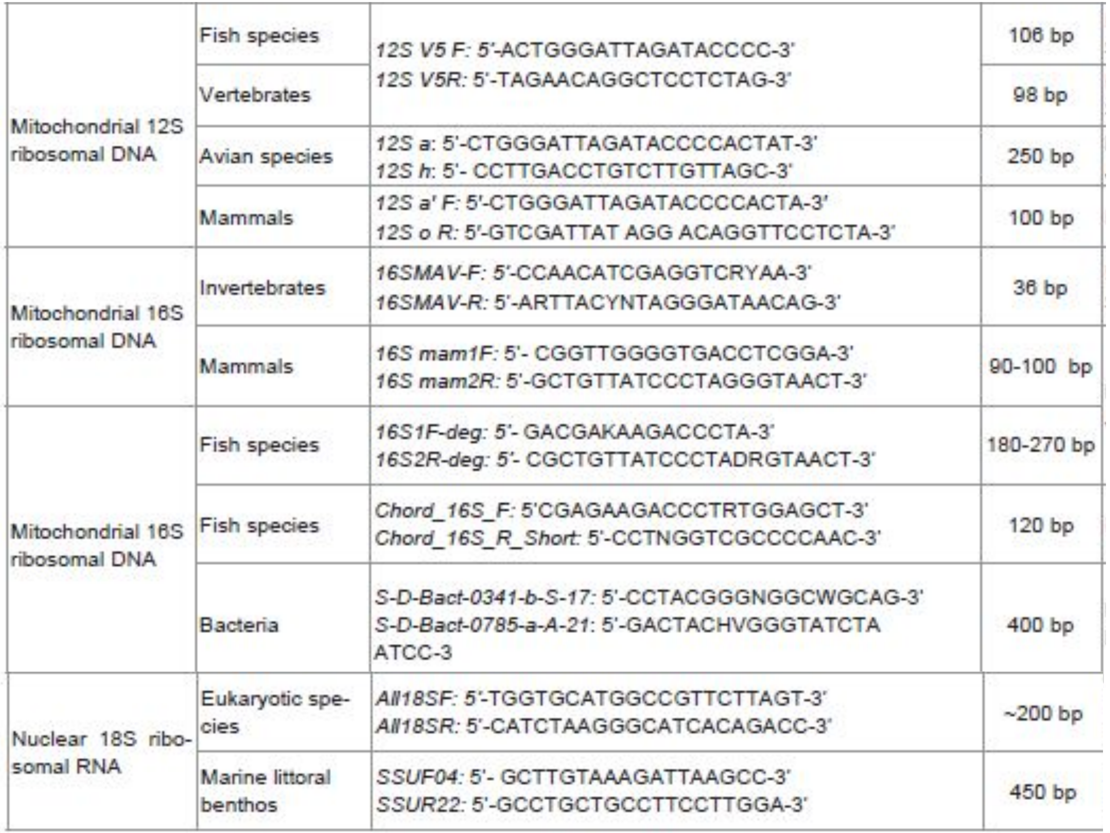


ssDNA fragment	Target species group	Primer sequences	Amplicon length
Mitochondrial Cytochrome *c* oxidase subunit I	Arthropods	*LCO1490*: 5′-GGTCAACAAATCATAAAGATATTGG-3′ *HCO2198*: 5′-TAAACTTCAGGGTGACCAAAAAATCA-3′	658 bp
	Metazoans	*mlCOlintF*: 5'-GGWACWGGWTGAACWGTWTAYCCYCC-3' *mlCOlintR*: 5'-GGRGGRTASACSGTTCASCCSGTSCC-3'	~320 bp
	Insect content in Avian diet	*ZBJ-ArtF1c*: 5'-AGATATTGGAACWTTATATTTTATTTTTGG-3' *ZBJ-ArtR2c*:5'-WACTAATCAATTWCCAAATCCTCC-3'	160 bp
	Arthropods	*Fol-degen-for*: 5'-TCNACNAAYCAYAARRAYATYGG-3' *Fol-degen-rev*: 5'-TANACYTCNGGRTGNCCRAARAAYCA-3'	~650 bp

Marker choice depends on target taxa → different taxa recovery

[Pavan-Kumar et al., J Cell Sci Molecul Biol., 2015]

Markers and their applications

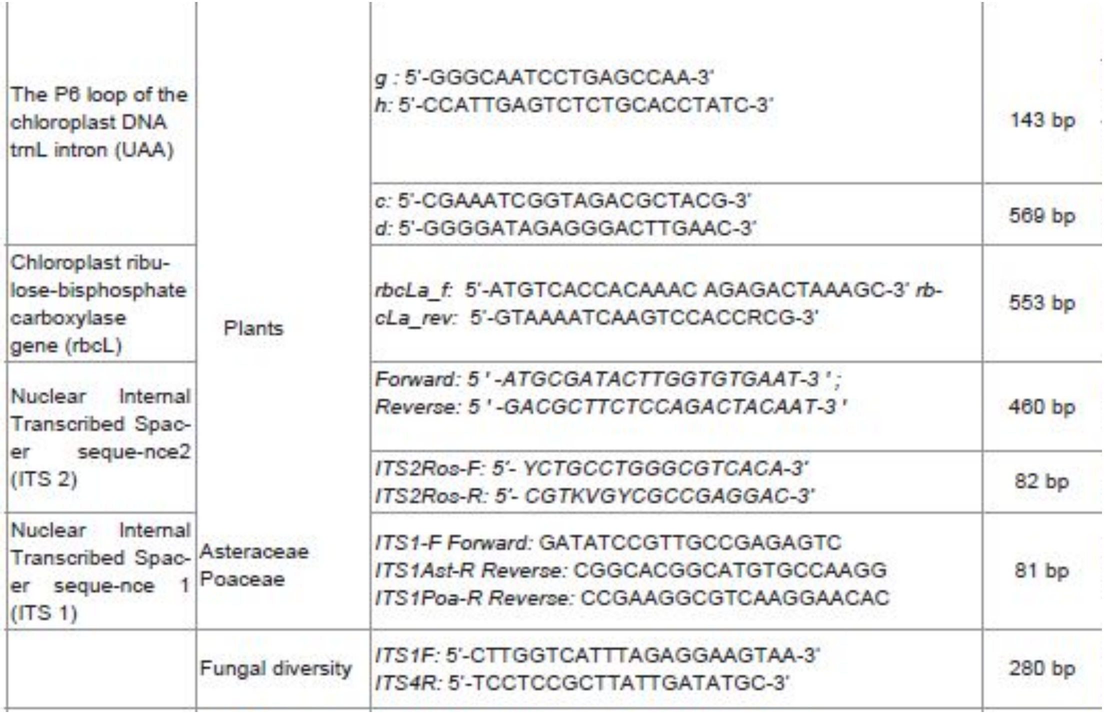


Mitochondrial 12S ribosomal DNA	Fish species	*12S V5 F:* 5'-ACTGGGATTAGATACCCC-3' *12S V5R:* 5'-TAGAACAGGCTCCTCTAG-3'	106 bp
	Vertebrates		98 bp
	Avian species	*12S a:* 5'-CTGGGATTAGATACCCCACTAT-3' *12S h:* 5'- CCTTGACCTGTCTTGTTAGC-3'	250 bp
	Mammals	*12S a' F:* 5'-CTGGGATTAGATACCCCACTA-3' *12S o R:* 5'-GTCGATTAT AGG ACAGGTTCCTCTA-3'	100 bp
Mitochondrial 16S ribosomal DNA	Invertebrates	*16SMAV-F:* 5'-CCAACATCGAGGTCRYAA-3' *16SMAV-R:* 5'-ARTTACYNTAGGGATAACAG-3'	36 bp
	Mammals	*16S mam1F:* 5'- CGGTTGGGGTGACCTCGGA-3' *16S mam2R:* 5'-GCTGTTATCCCTAGGGTAACT-3'	90-100 bp
Mitochondrial 16S ribosomal DNA	Fish species	*16S1F-deg:* 5'- GACGAKAAGACCCTA-3' *16S2R-deg:* 5'- CGCTGTTATCCCTADRGTAACT-3'	180-270 bp
	Fish species	*Chord_16S_F:* 5'CGAGAAGACCCTRTGGAGCT-3' *Chord_16S_R_Short:* 5'-CCTNGGTCGCCCCAAC-3'	120 bp
	Bacteria	*S-D-Bact-0341-b-S-17:* 5'-CCTACGGGNGGCWGCAG-3' *S-D-Bact-0785-a-A-21:* 5'-GACTACHVGGGTATCTA ATCC-3	400 bp
Nuclear 18S ribosomal RNA	Eukaryotic species	*All18SF:* 5'-TGGTGCATGGCCGTTCTTAGT-3' *All18SR:* 5'-CATCTAAGGGCATCACAGACC-3'	~200 bp
	Marine littoral benthos	*SSUF04:* 5'- GCTTGTAAAGATTAAGCC-3' *SSUR22:* 5'-GCCTGCTGCCTTCCTTGGA-3'	450 bp

Marker choice depends on target taxa → different taxa recovery

[Pavan-Kumar et al., J Cell Sci Molecul Biol., 2015]

Markers and their applications



The P6 loop of the chloroplast DNA trnL intron (UAA)	Plants	*g* : 5'-GGGCAATCCTGAGCCAA-3' *h*: 5'-CCATTGAGTCTCTGCACCTATC-3'	143 bp
		c: 5'-CGAAATCGGTAGACGCTACG-3' *d*: 5'-GGGGATAGAGGGACTTGAAC-3'	569 bp
Chloroplast ribulose-bisphosphate carboxylase gene (rbcL)		*rbcLa_f*: 5'-ATGTCACCACAAAC AGAGACTAAAGC-3' *rbcLa_rev*: 5'-GTAAAATCAAGTCCACCRCG-3'	553 bp
Nuclear Internal Transcribed Spacer seque-nce2 (ITS 2)		*Forward: 5'-ATGCGATACTTGGTGTGAAT-3';* *Reverse: 5'-GACGCTTCTCCAGACTACAAT-3'*	460 bp
		ITS2Ros-F: 5'- YCTGCCTGGGCGTCACA-3' *ITS2Ros-R: 5'- CGTKVGYCGCCGAGGAC-3'*	82 bp
Nuclear Internal Transcribed Spacer seque-nce 1 (ITS 1)	Asteraceae Poaceae	*ITS1-F Forward:* GATATCCGTTGCCGAGAGTC *ITS1Ast-R Reverse:* CGGCACGGCATGTGCCAAGG *ITS1Poa-R Reverse:* CCGAAGGCGTCAAGGAACAC	81 bp
	Fungal diversity	*ITS1F*: 5'-CTTGGTCATTTAGAGGAAGTAA-3' *ITS4R*: 5'-TCCTCCGCTTATTGATATGC-3'	280 bp

Multiple markers provide better resolution!

[Pavan-Kumar et al., J Cell Sci Molecul Biol., 2015]

Protocols matter

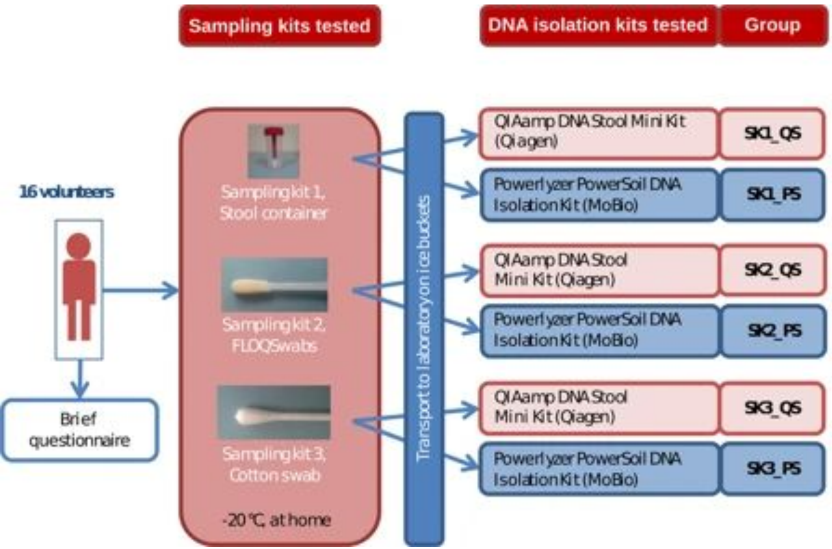
Many different technical variants → impact on results!

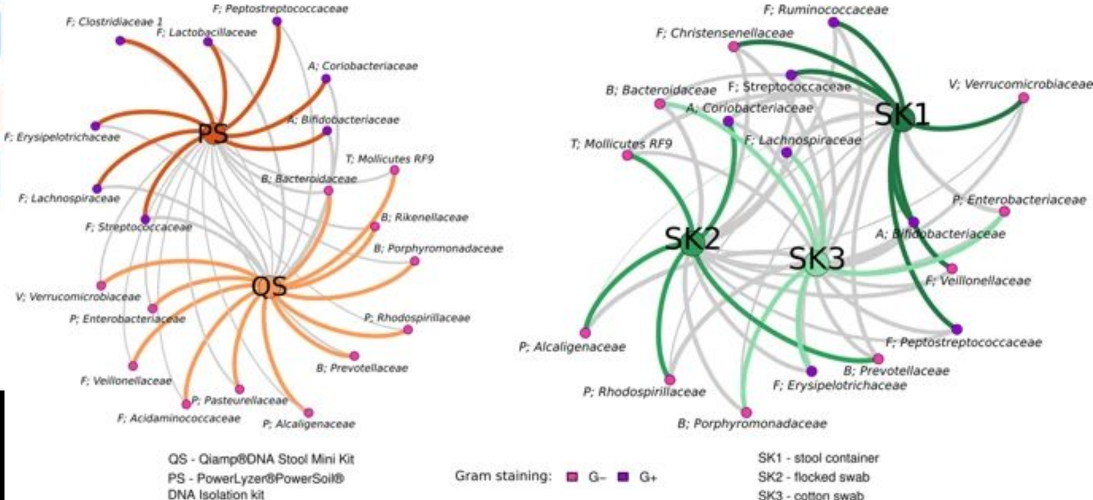
-Sample collection, storage, transport, lab processing, seque

Article | Open Access | Published: 25 September 2019

Stool sampling and DNA isolation kits affect DNA quality and bacterial composition following 16S rRNA gene sequencing using MiSeq Illumina platform

Petra Videnska, Kristyna Smerkova, Barbora Zwinsova, Vlad Popovici, Lenka Micenkova, Karel Sedlar & Eva Budinska



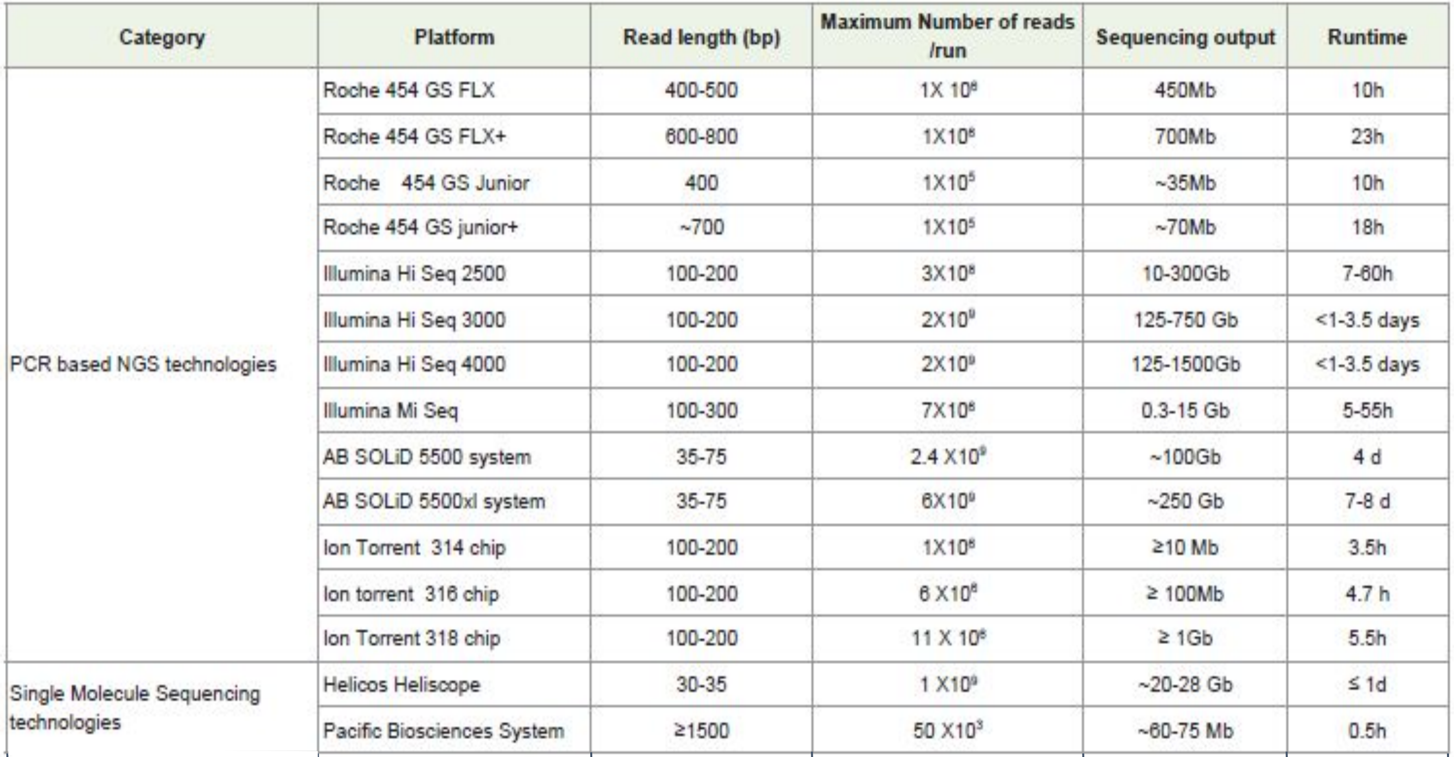


QS - Qiamp®DNA Stool Mini Kit
PS - PowerLyzer®PowerSoil® DNA Isolation kit

Gram staining: ■ G– ■ G+

SK1 - stool container
SK2 - flocked swab
SK3 - cotton swab

Sequencing technologies



Category	Platform	Read length (bp)	Maximum Number of reads /run	Sequencing output	Runtime
PCR based NGS technologies	Roche 454 GS FLX	400-500	1X 10^6	450Mb	10h
	Roche 454 GS FLX+	600-800	1X10^6	700Mb	23h
	Roche 454 GS Junior	400	1X10^5	~35Mb	10h
	Roche 454 GS junior+	~700	1X10^5	~70Mb	18h
	Illumina Hi Seq 2500	100-200	3X10^6	10-300Gb	7-60h
	Illumina Hi Seq 3000	100-200	2X10^9	125-750 Gb	<1-3.5 days
	Illumina Hi Seq 4000	100-200	2X10^9	125-1500Gb	<1-3.5 days
	Illumina Mi Seq	100-300	7X10^6	0.3-15 Gb	5-55h
	AB SOLiD 5500 system	35-75	2.4 X10^9	~100Gb	4 d
	AB SOLiD 5500xl system	35-75	6X10^9	~250 Gb	7-8 d
	Ion Torrent 314 chip	100-200	1X10^6	≥10 Mb	3.5h
	Ion torrent 316 chip	100-200	6 X10^6	≥ 100Mb	4.7 h
	Ion Torrent 318 chip	100-200	11 X 10^6	≥ 1Gb	5.5h
Single Molecule Sequencing technologies	Helicos Heliscope	30-35	1 X10^9	~20-28 Gb	≤ 1d
	Pacific Biosciences System	≥1500	50 X10^3	~60-75 Mb	0.5h
	Oxford Nanopore Technologies (MinIon)	bps-kbs	millions	<=30Gb (DNA)	<= 48h per flow cell

-Different sequencing chemistries

-different throughput

-different suitability

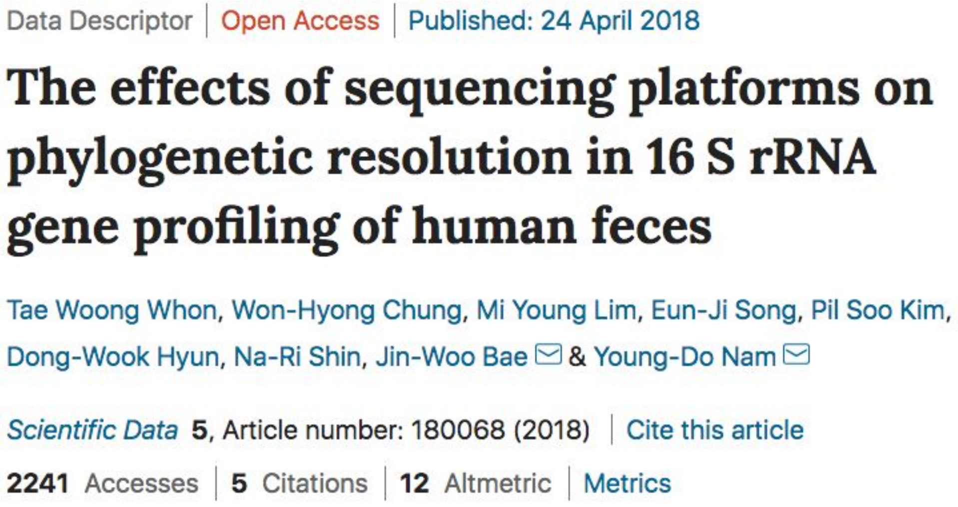
*adapted from [Pavan-Kumar et al., J Cell Sci Molecul Biol., 2015]

Sequencing technologies matter

Mol Ecol Resour. 2018 May;18(3):590-601. doi: 10.1111/1755-0998.12770. Epub 2018 Mar 8.

Disparities in second-generation DNA metabarcoding results exposed with accessible and repeatable workflows.

Divoll TJ[1], Brown VA[2,3], Kinne J[4], McCracken GF[2], O'Keefe JM[1].



Data Descriptor | Open Access | Published: 24 April 2018

The effects of sequencing platforms on phylogenetic resolution in 16 S rRNA gene profiling of human feces

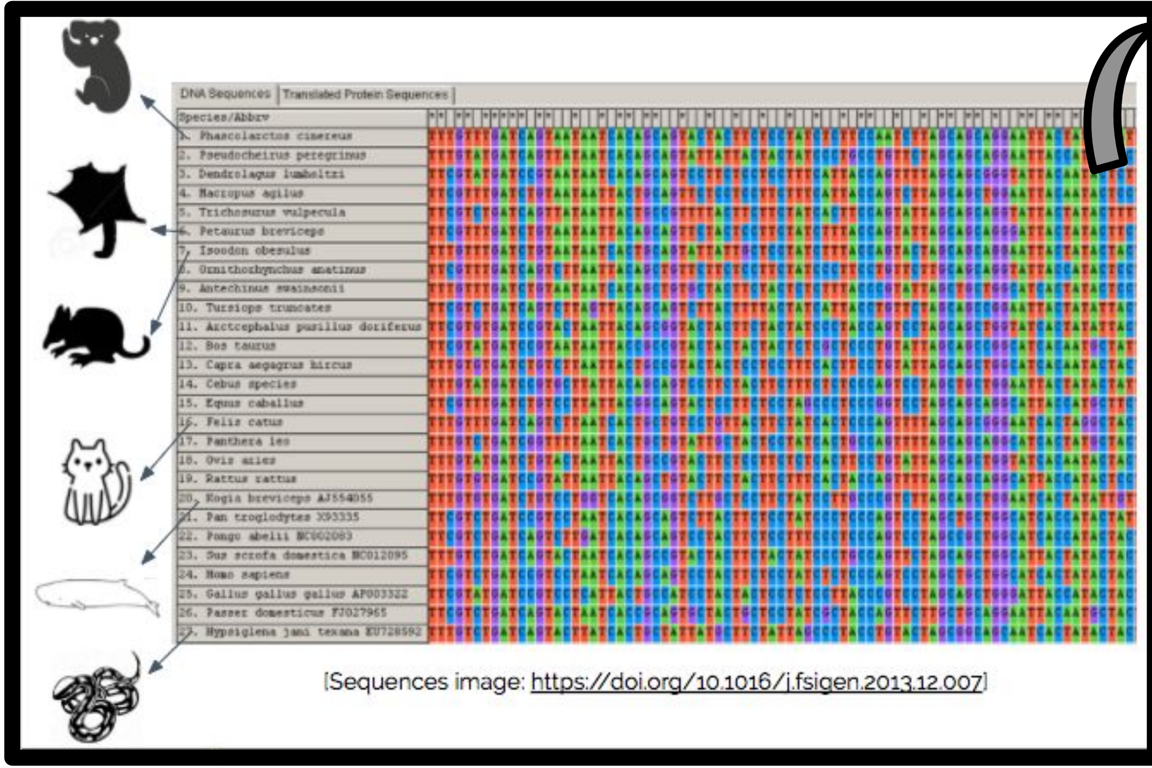
Tae Woong Whon, Won-Hyong Chung, Mi Young Lim, Eun-Ji Song, Pil Soo Kim, Dong-Wook Hyun, Na-Ri Shin, Jin-Woo Bae ✉ & Young-Do Nam ✉

Scientific Data **5**, Article number: 180068 (2018) | Cite this article

2241 Accesses | **5** Citations | **12** Altmetric | Metrics

Metabarcoding data processing

From thousands/millions of sequences to taxonomic ID

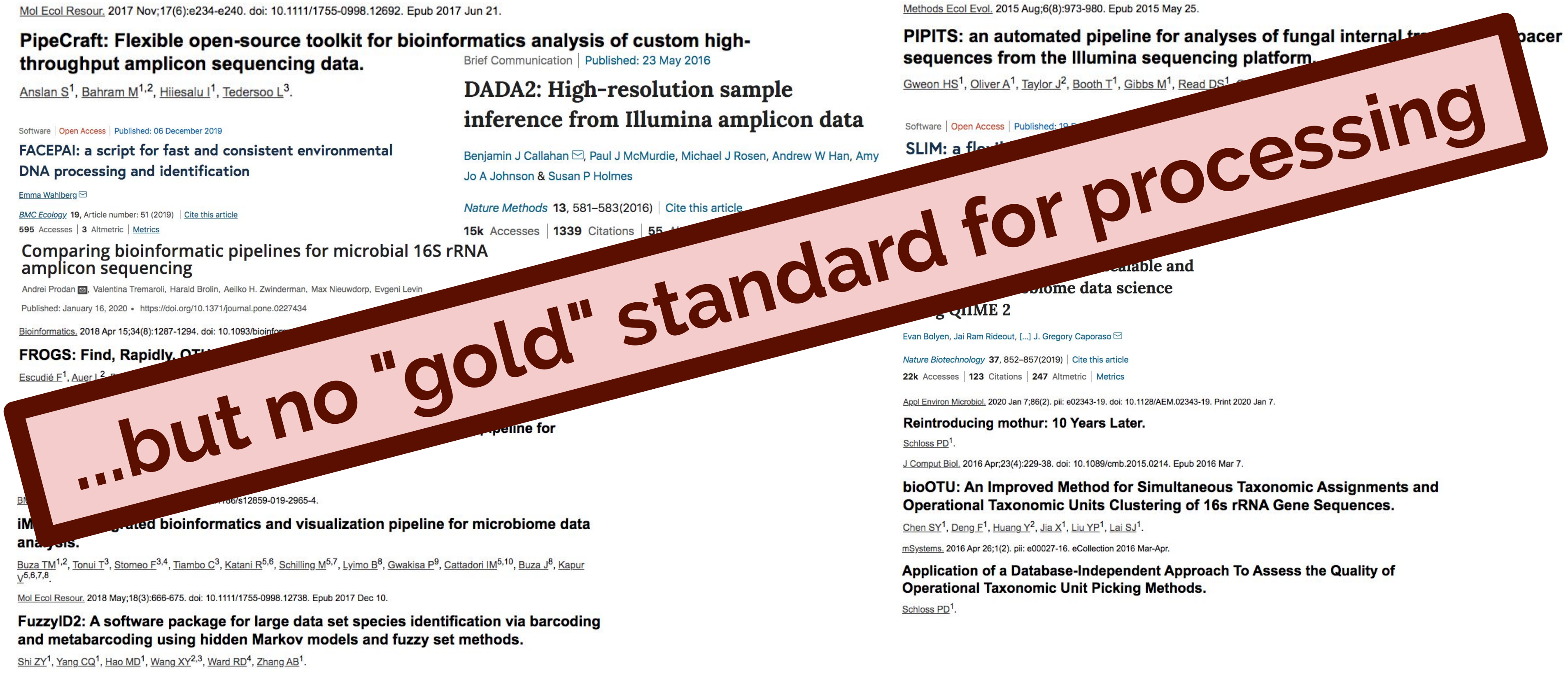


Species/Abbrv
1. Phascolarctos cinereus
2. Pseudocheirus peregrinus
3. Dendrolagus lumholtzi
4. Macropus agilus
5. Trichosurus vulpecula
6. Petaurus breviceps
7. Isoodon obesulus
8. Ornithorhynchus anatinus
9. Antechinus swainsonii
10. Tursiops truncates
11. Arctocephalus pusillus doriferus
12. Bos taurus
13. Capra aegagrus hircus
14. Cebus species
15. Equus caballus
16. Felis catus
17. Panthera leo
18. Ovis aries
19. Rattus rattus
20. Kogia breviceps AJ554055
21. Pan troglodytes X93335
22. Pongo abelii NC002083
23. Sus scrofa domestica NC012095
24. Homo sapiens
25. Gallus gallus gallus AP003322
26. Passer domesticus FJ027965
27. Hypsiglena jani texana EU728592

[Sequences image: https://doi.org/10.1016/j.fsigen.2013.12.007]

- Raw reads
- Filtering steps
- Clustering &/or TaxID
- Species identification

Many available processing pipelines



Mol Ecol Resour. 2017 Nov;17(6):e234-e240. doi: 10.1111/1755-0998.12692. Epub 2017 Jun 21.

PipeCraft: Flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data.

Anslan S[1], Bahram M[1,2], Hiiesalu I[1], Tedersoo L[3].

Software | Open Access | Published: 06 December 2019

FACEPAI: a script for fast and consistent environmental DNA processing and identification

Emma Wahlberg

BMC Ecology 19, Article number: 51 (2019) | Cite this article

595 Accesses | 3 Altmetric | Metrics

Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing

Andrei Prodan, Valentina Tremaroli, Harald Brolin, Aeilko H. Zwinderman, Max Nieuwdorp, Evgeni Levin

Published: January 16, 2020 • https://doi.org/10.1371/journal.pone.0227434

Bioinformatics. 2018 Apr 15;34(8):1287-1294. doi: 10.1093/bioinfor

FROGS: Find, Rapidly, OT

Escudié F[1], Auer L[2],

ipeline for

BM 186/s12859-019-2965-4.

iM grated bioinformatics and visualization pipeline for microbiome data analysis.

Buza TM[1,2], Tonui T[3], Stomeo F[3,4], Tiambo C[3], Katani R[5,6], Schilling M[5,7], Lyimo B[8], Gwakisa P[9], Cattadori IM[5,10], Buza J[8], Kapur V[5,6,7,8].

Mol Ecol Resour. 2018 May;18(3):666-675. doi: 10.1111/1755-0998.12738. Epub 2017 Dec 10.

FuzzyID2: A software package for large data set species identification via barcoding and metabarcoding using hidden Markov models and fuzzy set methods.

Shi ZY[1], Yang CQ[1], Hao MD[1], Wang XY[2,3], Ward RD[4], Zhang AB[1].

Brief Communication | Published: 23 May 2016

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson & Susan P Holmes

Nature Methods 13, 581–583(2016) | Cite this article

15k Accesses | 1339 Citations | 55

Methods Ecol Evol. 2015 Aug;6(8):973-980. Epub 2015 May 25.

PIPITS: an automated pipeline for analyses of fungal internal tr pacer sequences from the Illumina sequencing platform.

Gweon HS[1], Oliver A[1], Taylor J[2], Booth T[1], Gibbs M[1], Read DS[1],

Software | Open Access | Published: 19

SLIM: a fl

calable and biome data science QIIME 2

Evan Bolyen, Jai Ram Rideout, [...] J. Gregory Caporaso

Nature Biotechnology 37, 852–857(2019) | Cite this article

22k Accesses | 123 Citations | 247 Altmetric | Metrics

Appl Environ Microbiol. 2020 Jan 7;86(2). pii: e02343-19. doi: 10.1128/AEM.02343-19. Print 2020 Jan 7.

Reintroducing mothur: 10 Years Later.

Schloss PD[1].

J Comput Biol. 2016 Apr;23(4):229-38. doi: 10.1089/cmb.2015.0214. Epub 2016 Mar 7.

bioOTU: An Improved Method for Simultaneous Taxonomic Assignments and Operational Taxonomic Units Clustering of 16s rRNA Gene Sequences.

Chen SY[1], Deng F[1], Huang Y[2], Jia X[1], Liu YP[1], Lai SJ[1].

mSystems. 2016 Apr 26;1(2). pii: e00027-16. eCollection 2016 Mar-Apr.

Application of a Database-Independent Approach To Assess the Quality of Operational Taxonomic Unit Picking Methods.

Schloss PD[1].

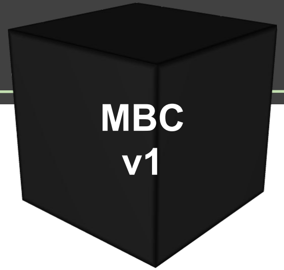
...but no "gold" standard for processing

Some possible pipelines

MBC
v1

step1	Error correction		spades		none	
step2	Pairwise alignment	obitools		vsearch		pear
step3	Pre-filter (q-score)	obitools		vsearch + cutadapt		none
step4	dereplication		obitools		vsearch	
step5	Attribute filter	obitools		vsearch		none
step6	clustering	obitools	sumaclust	vsearch	swarm	none
step7	Db search (tax id)		blastn		bowtie2	
step8	Taxa table					

Some possible pipelines



step1	Error correction		spades		none	
step2	Pairwise alignment	obitools		vsearch		pear
step3	Pre-filter (q-score)	obitools		vsearch + cutadapt		none
step4	dereplication		obitools		vsearch	
step5	Attribute filter	obitools		vsearch		none
step6	clustering	obitools	sumaclust	vsearch	swarm	none
step7	Db search (tax id)		blastn		bowtie2	
step8	Taxa table					

Incompatible combination of tools

MBC
v1

step1	Error correction		spades		none	
step2	Pairwise alignment	obitools		vsearch		pear
step3	Pre-filter (q-score)	obitools		vsearch + cutadapt		none
step4	dereplication		obitools		vsearch	
step5	Attribute filter	obitools		vsearch		none
step6	clustering	obitools	sumaclust	vsearch	swarm	none
step7	Db search (tax id)		blastn			
step8	Taxa table					

Needs “merged_sample” field from obitools processing

MBC benchmarking

- Which pipeline to use?
    - Real data and mock samples (known ground truth)
    - Comparing different pipelines
        - Computational resources: memory/cpu time
        - Consistency (are results similar)
        - Accuracy

➢ Pipelines have very different computational requirements

➢ Consistency/Accuracy varies

MBC report

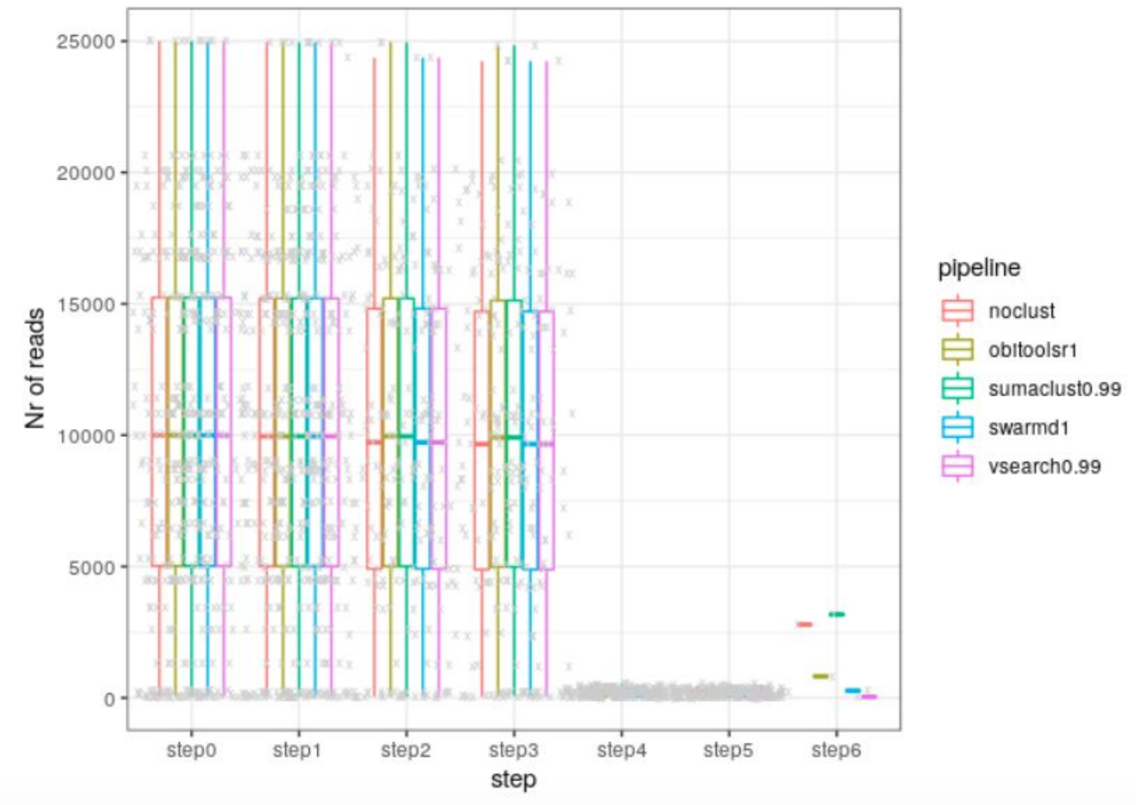
-mbc: 5 main pipelines - obitools, sumaclust, vsearch, swarm, noclust

-mbc report: Rmarkdown / html file - plots summarizing stats per processing step and pipeline

-automated generation by the end of a processing run

-main purpose: facilitate the choice of a pipeline / set of pipelines for analyzing bigger dataset; compare pipelines; understand if experiment worked as expected; ...

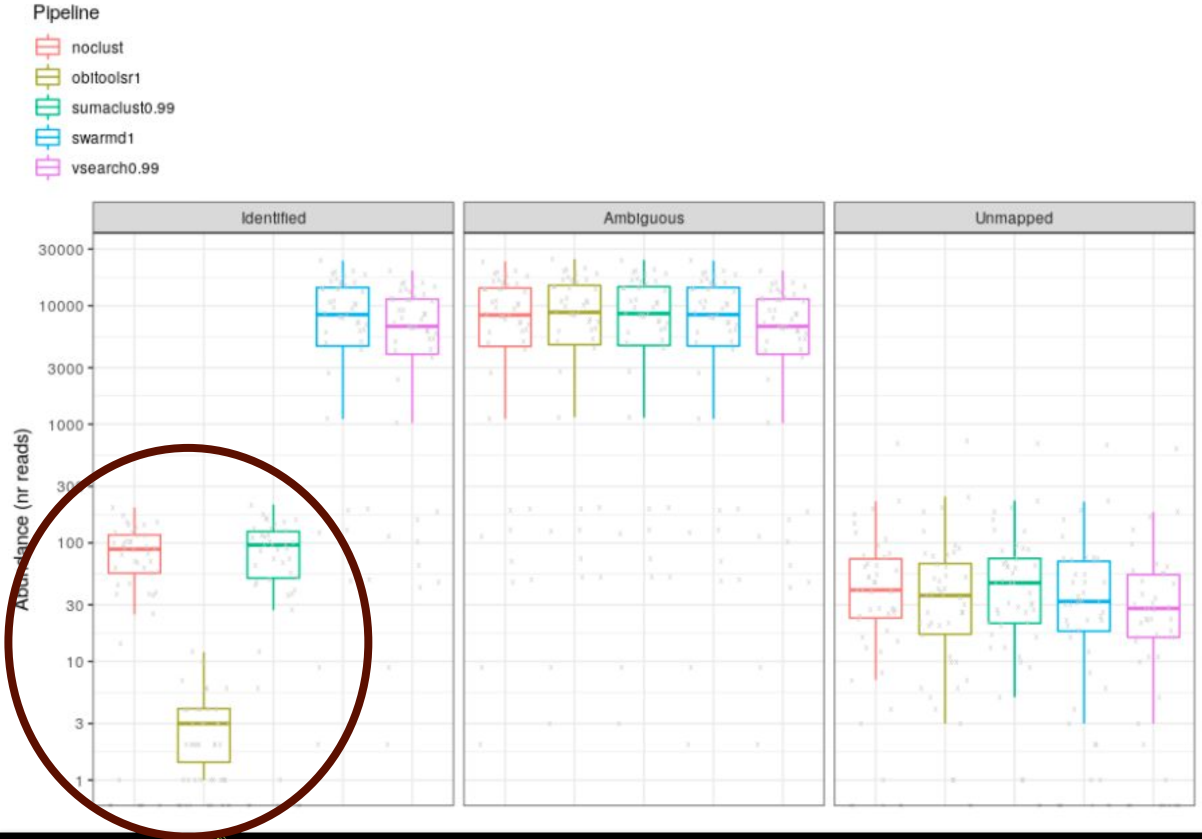
Sequence stats



N=50

- number of reads per step / pipeline / sample

Sequence identification stats

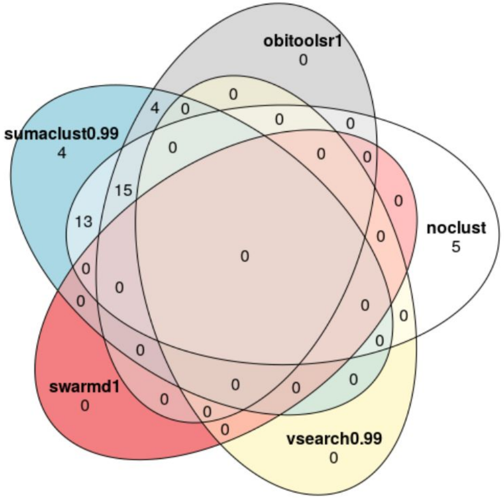


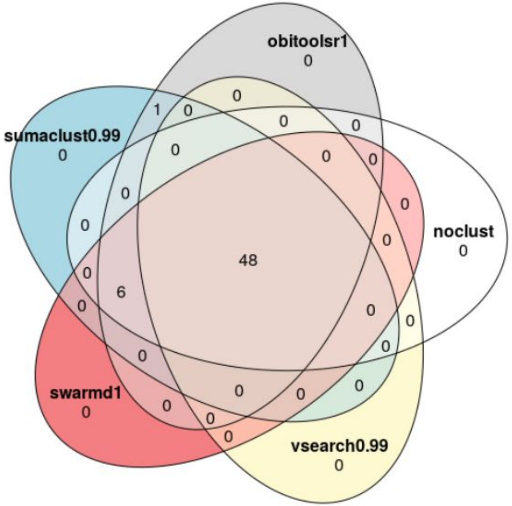
N=50

- number of "identified", "ambiguous" and "unidentified" sequences per pipeline (and sample)

- 3 pipelines with much less "identified" sequences

Pipeline agreement on detected taxa



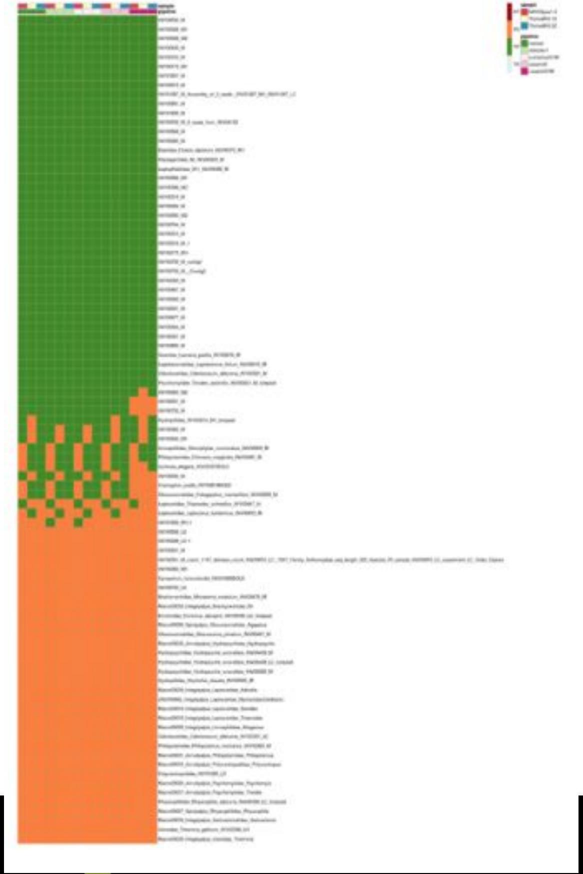


- Number of shared taxa detected per pipeline

-quick way to check pipeline agreement

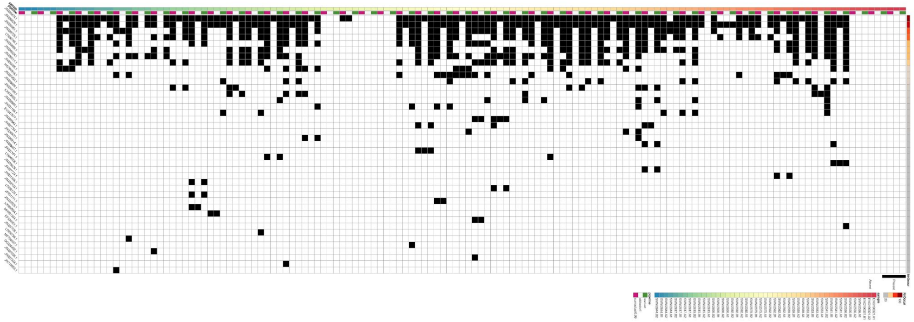
-check whether 1 / some pipelines working much better / worse than others

Detection success | presence /absence

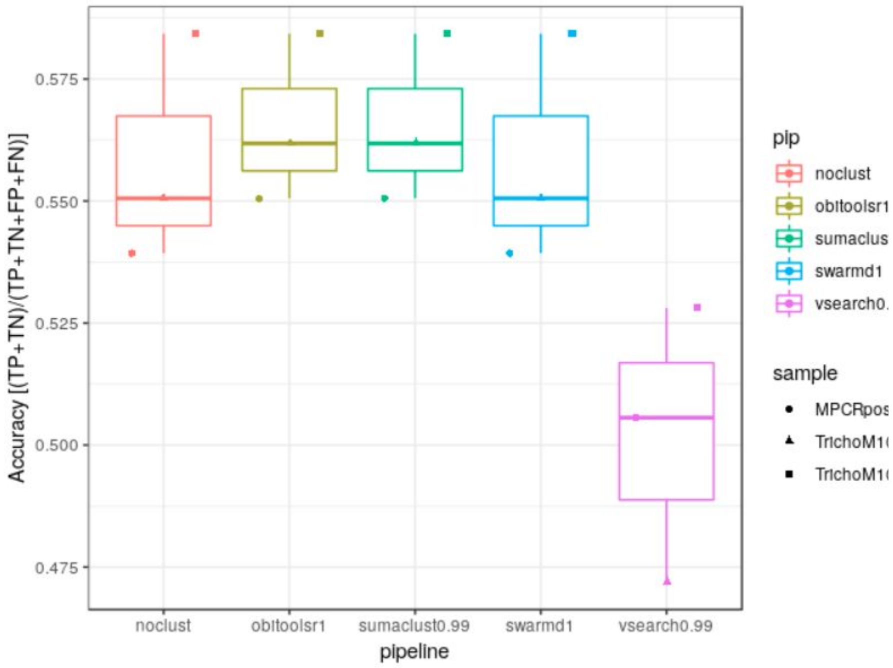


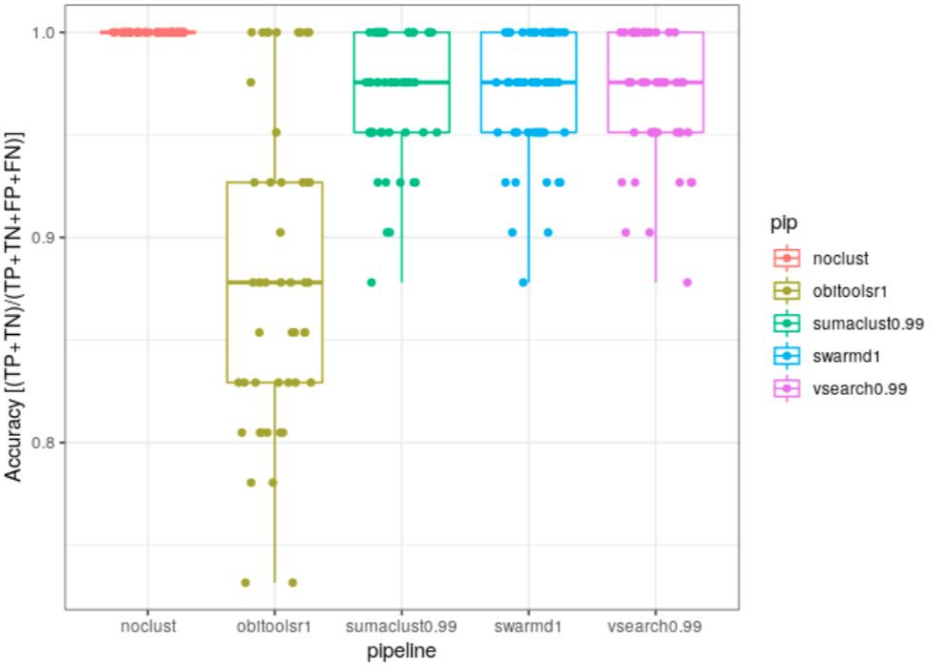
- Known sample composition (ref/mock) → TP, FP, FN
- Unknown composition → similar taxa appearance per sample? → taxa presence/absence heatmap

-invasive species alert, …

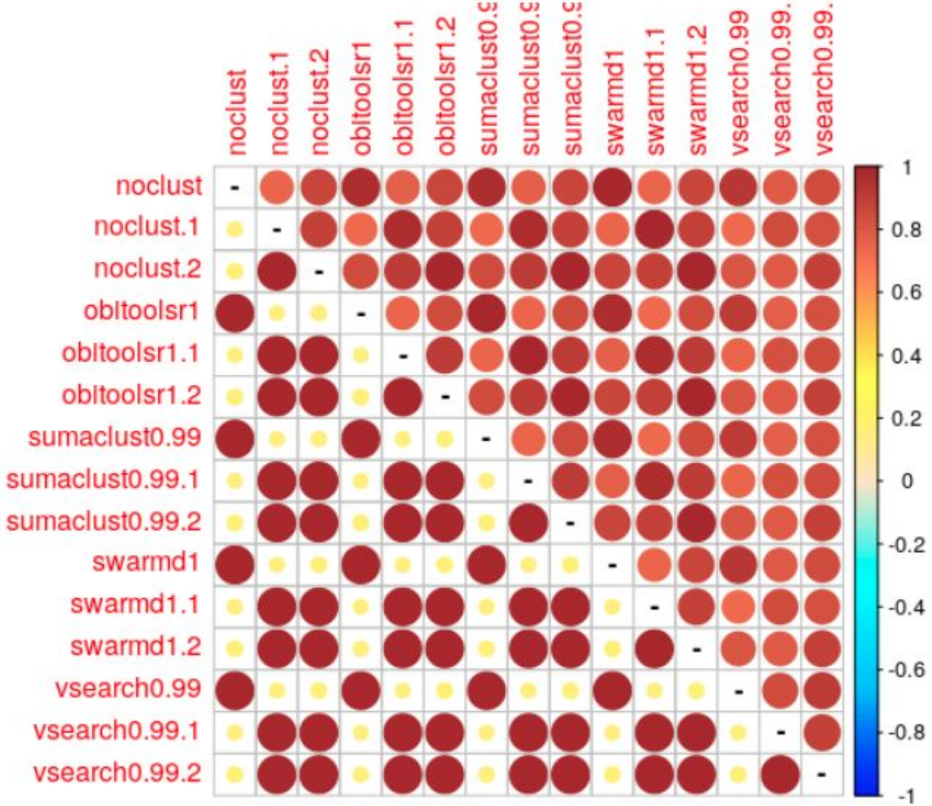


Detection accuracy





Sample correlation

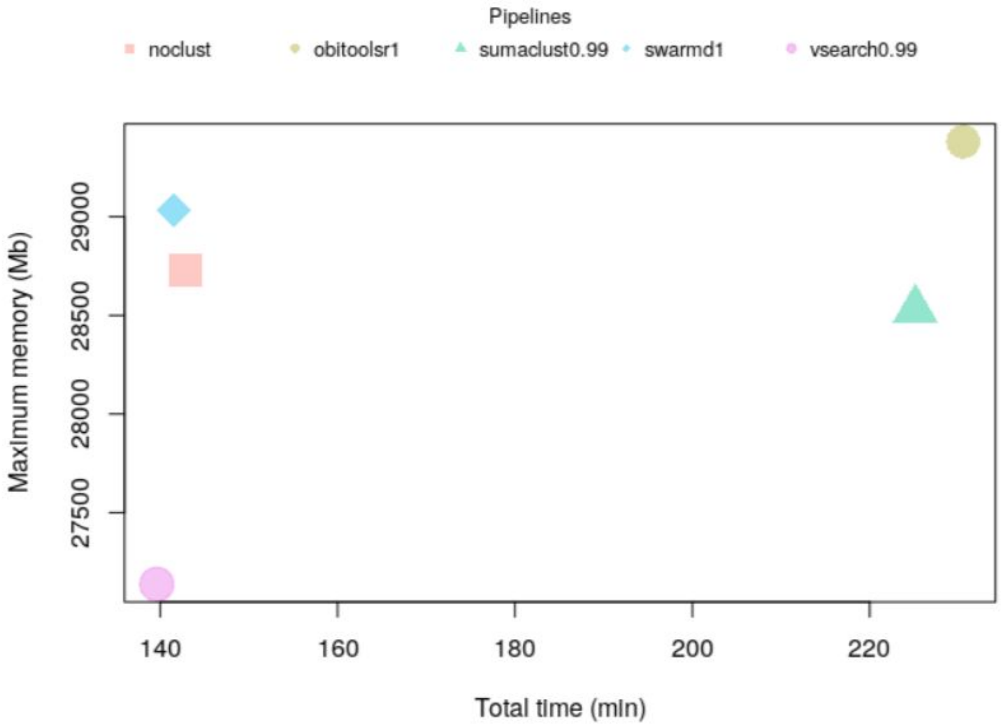


Correlation among samples:
-presence/absence (upper triangle)
-number of reads (lower triangle)

For higher number of samples → boxplots

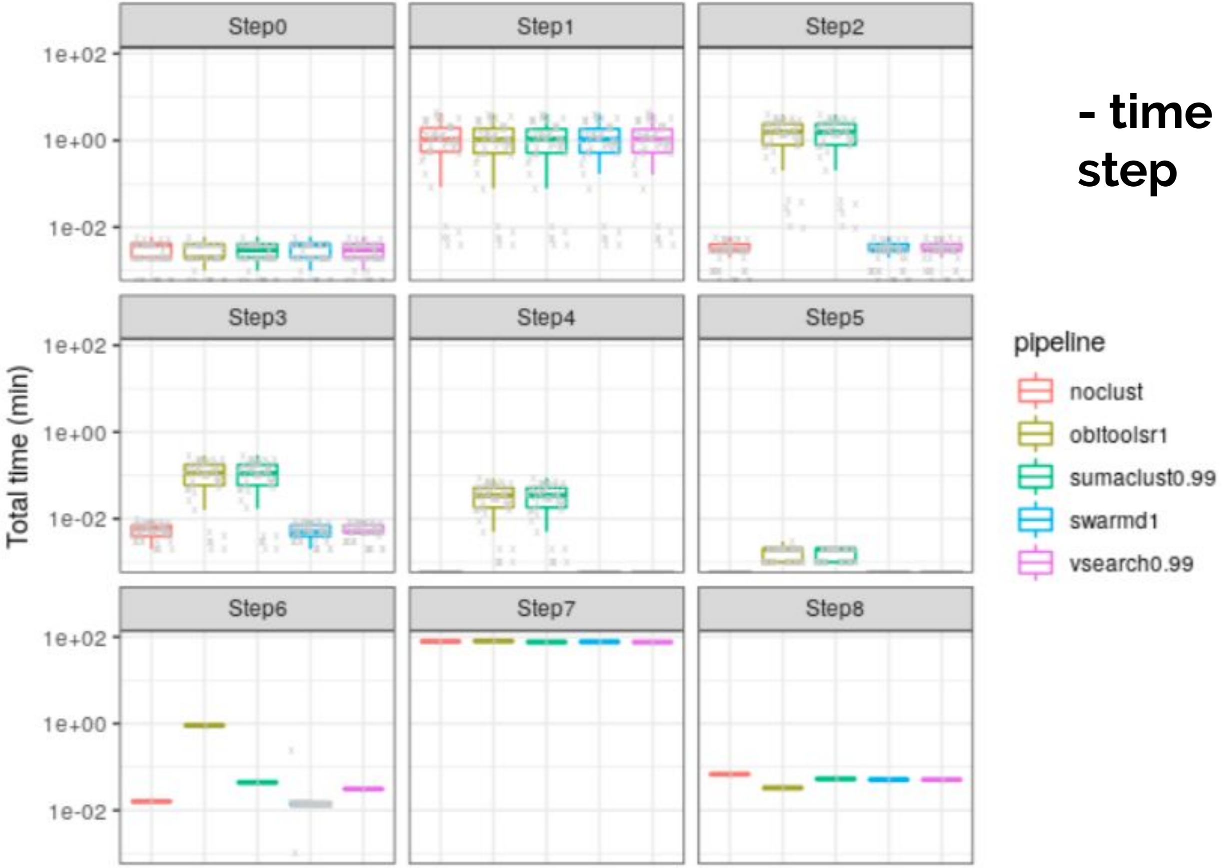
Total time & memory usage

N=50



- General overview of time & memory resources used by each pipeline

Detailed time usage



N=50

- time each pipeline took per step

Detailed memory usage

N=50

- memory each pipeline took per step

MBC pipeline

- Easy, flexible and reproducible data processing
- Toolbox connecting "all" steps of metabarcoding data processing
- Efficient computing
- Automated generation of a summary report for easy data inspection
- Possibility of adding processing tools
- Will be publicly available on GitHub

Metabarcoding

Different biological questions!

-Biomonitoring, biodiversity assessment,...

Several DNA markers!

-COI, 12S, 16S, 18S, ITS, chloroplast, ...

Several sequencing technologies!

-Illumina, IonTorrent, PacBio, Nanopore, ...

Many processing tools!

-obitools, v/usearch, qiime, mothur, dada2, sumaclust,...

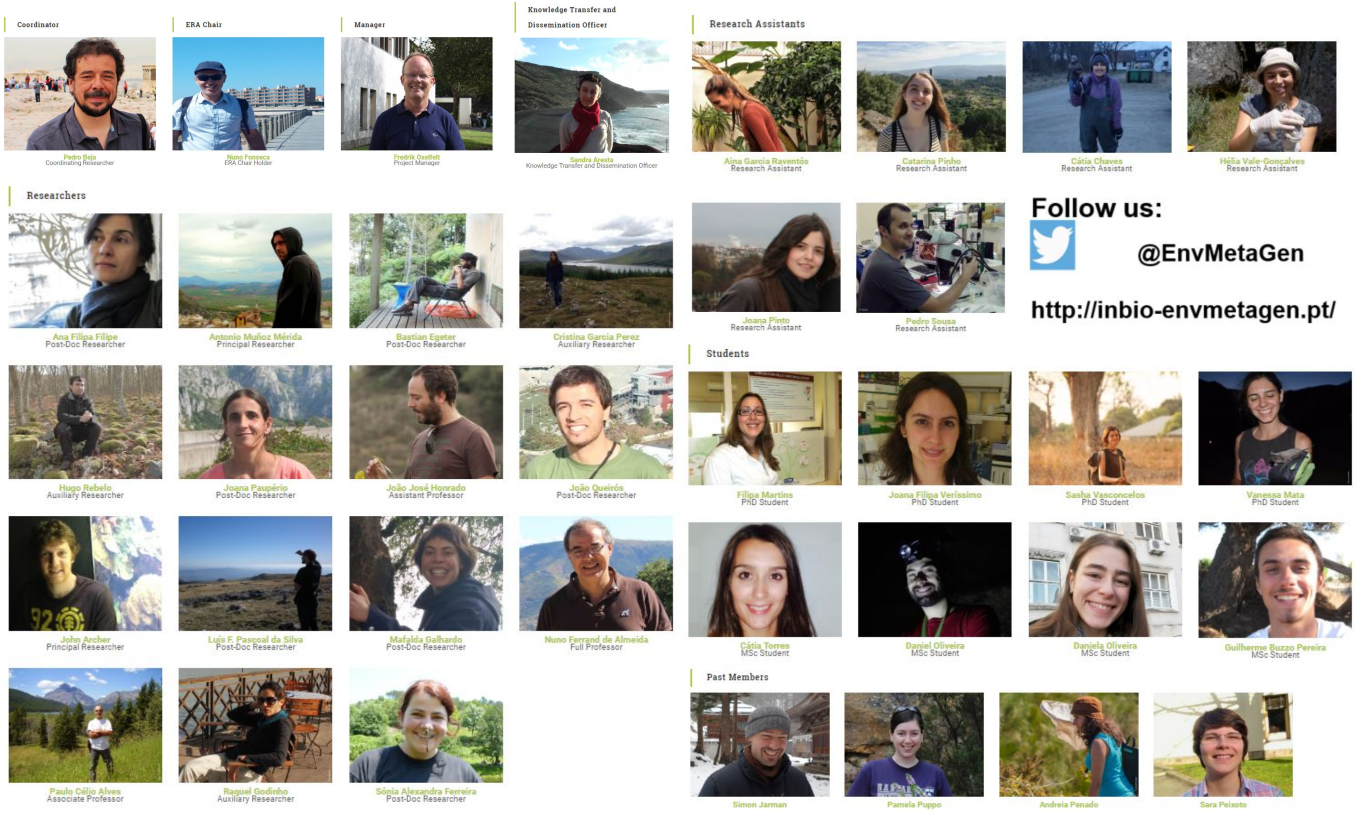
Reference database paucity!

-Uneven taxa presence in databases, DNA barcodes unavailable for many taxa

Lack of standards!

-technical procedures, metadata, ...

Team



Coordinator

Pedro Beja
Coordinating Researcher

ERA Chair

Nuno Fonseca
ERA Chair Holder

Manager

Fredrik Oxelfelt
Project Manager

Knowledge Transfer and Dissemination Officer

Sandra Aresta
Knowledge Transfer and Dissemination Officer

Researchers

Ana Filipa Filipe
Post-Doc Researcher

Antonio Muñoz Mérida
Principal Researcher

Bastian Egeter
Post-Doc Researcher

Cristina Garcia Perez
Auxiliary Researcher

Hugo Rebelo
Auxiliary Researcher

Joana Paupério
Post-Doc Researcher

João José Honrado
Assistant Professor

João Queirós
Post-Doc Researcher

John Archer
Principal Researcher

Luís F. Pascoal da Silva
Post-Doc Researcher

Mafalda Galhardo
Post-Doc Researcher

Nuno Ferrand de Almeida
Full Professor

Paulo Célio Alves
Associate Professor

Raquel Godinho
Auxiliary Researcher

Sónia Alexandra Ferreira
Post-Doc Researcher

Research Assistants

Aina Garcia Raventós
Research Assistant

Catarina Pinho
Research Assistant

Cátia Chaves
Research Assistant

Hélia Vale-Gonçalves
Research Assistant

Joana Pinto
Research Assistant

Pedro Sousa
Research Assistant

Students

Filipa Martins
PhD Student

Joana Filipa Veríssimo
PhD Student

Sasha Vasconcelos
PhD Student

Vanessa Mata
PhD Student

Cátia Torres
MSc Student

Daniel Oliveira
MSc Student

Daniela Oliveira
MSc Student

Guilherme Buzzo Pereira
MSc Student

Past Members

Simon Jarman

Pamela Puppo

Andreia Penado

Sara Peixoto

Follow us:

@EnvMetaGen

http://inbio-envmetagen.pt/

Questions?