Metabarcoding : the main steps - Part 1

Clarisse Lemonnier

BIOLAWEB

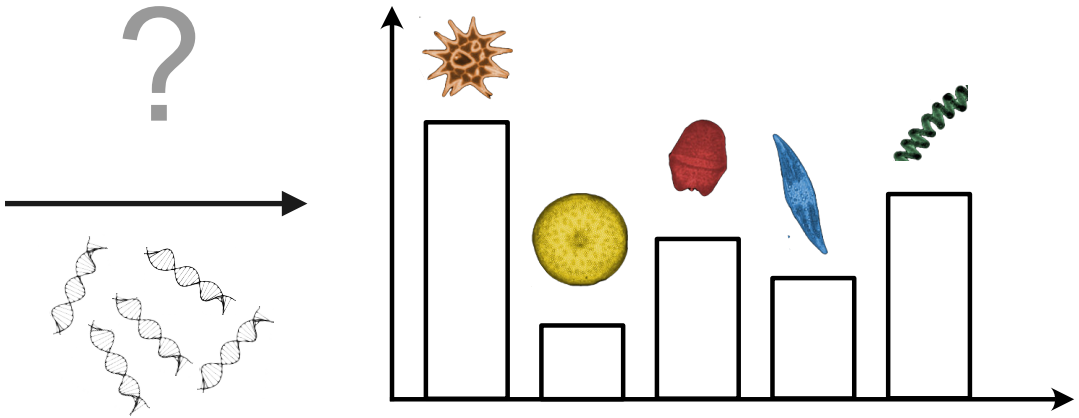
Summary

The main steps of metabarcoding

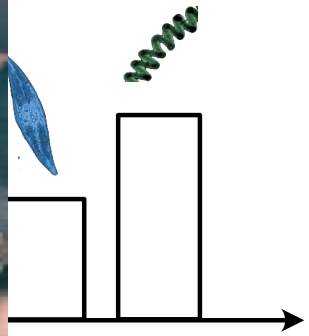
Selection of the barcode

Reference databases

Metabarcoding steps



Metabarcoding steps



Metabarcoding steps

Sampling 	?	?	?	?

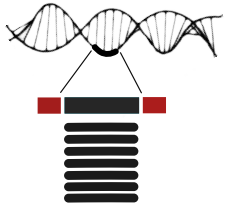
Funded by European Union

Metabarcoding steps

Sampling	DNA extraction	?	?	?
				

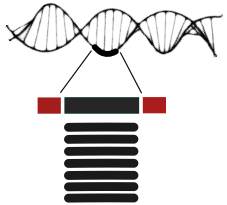
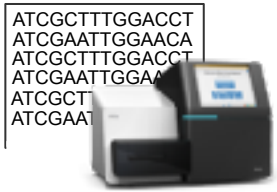
Funded by European Union

Metabarcoding steps

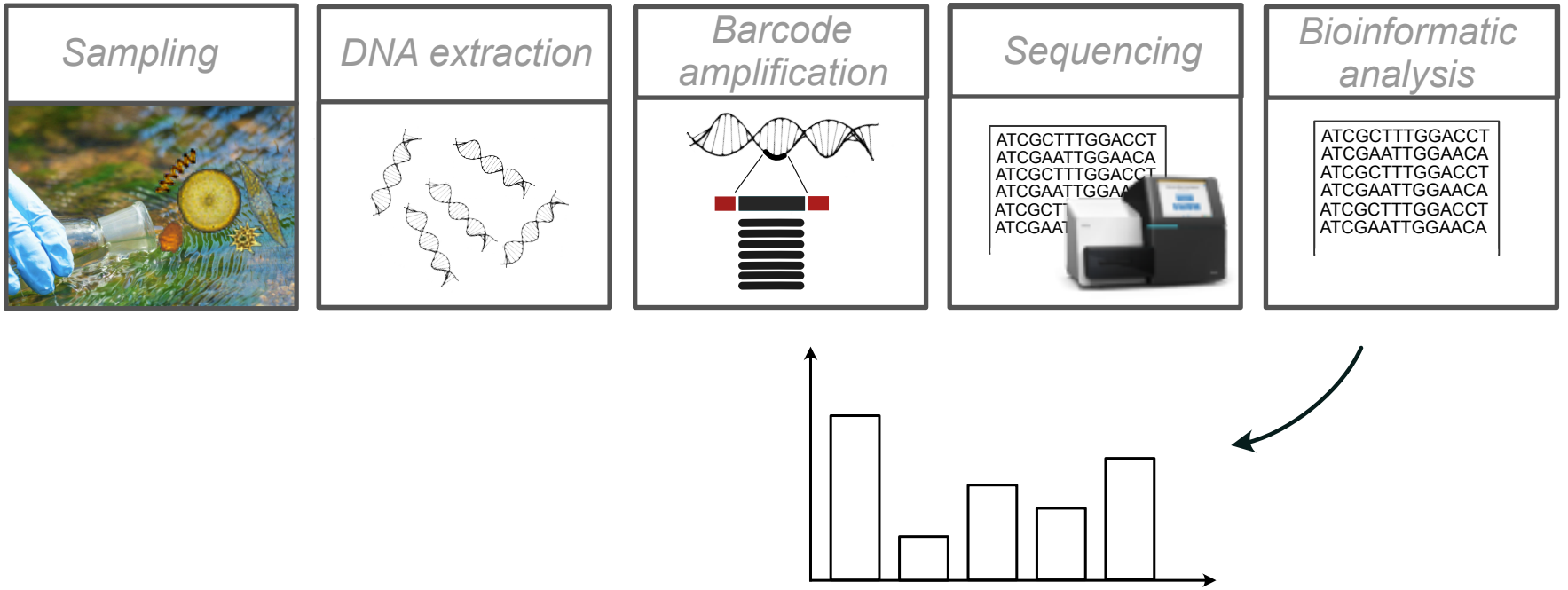
Sampling	DNA extraction	Barcode amplification	?	?
				

Funded by European Union

Metabarcoding steps

Sampling	DNA extraction	Barcode amplification	Sequencing	?
				

Metabarcoding steps



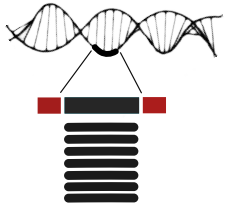
Funded by European Union

Metabarcoding steps

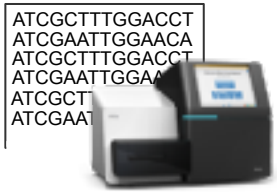
Sampling

DNA extraction

Barcode amplification

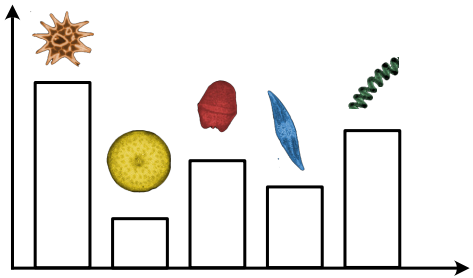


Sequencing



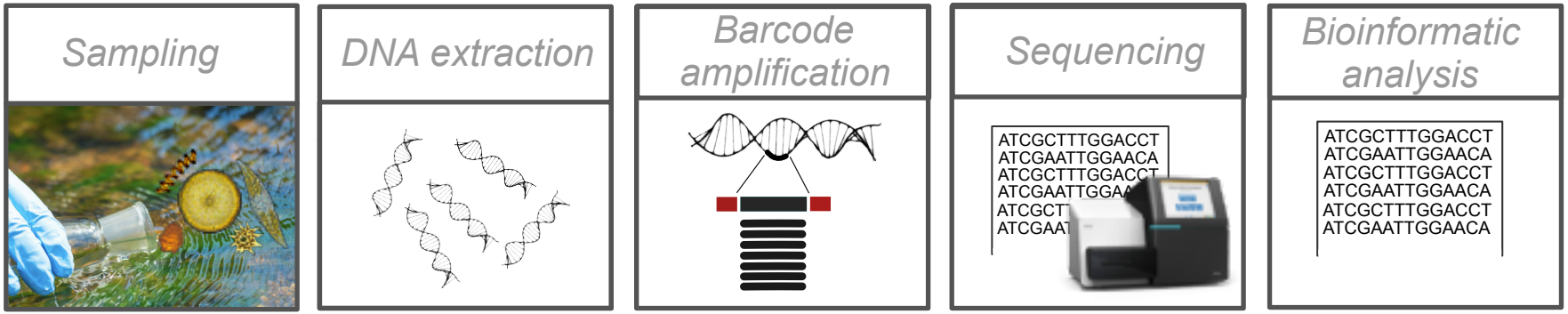
Bioinformatic analysis

ATCGCTTTGGACCT
ATCGAATTGGAACA
ATCGCTTTGGACCT
ATCGAATTGGAACA
ATCGCTTTGGACCT
ATCGAATTGGAACA

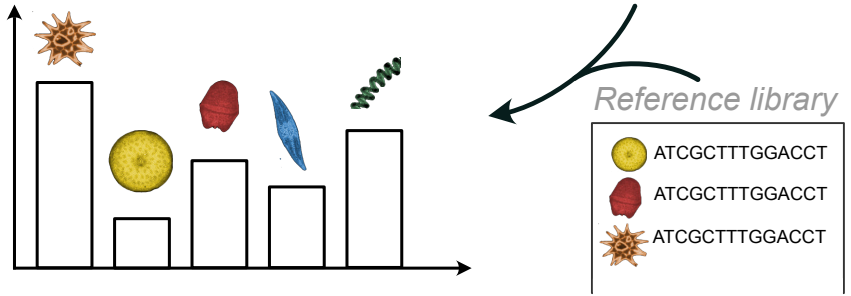


Funded by European Union

Metabarcoding steps



An informed choice at each step is required to get a reliable result at the end

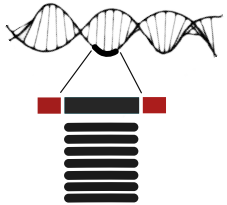


Metabarcoding steps

Sampling

DNA extraction

Barcode amplification



Sequencing

Bioinformatic analysis

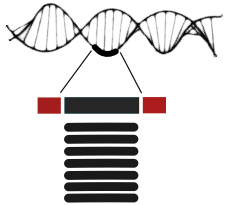
Barcode selection

Metabarcoding steps

Sampling

DNA extraction

Barcode amplification



Sequencing

Bioinformatic analysis

Recovery of all species

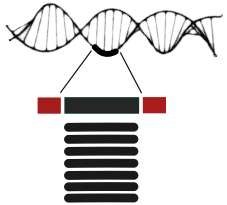
Barcode selection

Metabarcoding steps

Sampling

DNA extraction

Barcode amplification



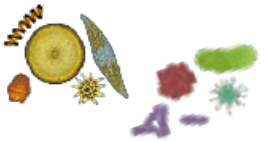
Sequencing

Bioinformatic analysis

Recovery of all species

Barcode selection

Specificity

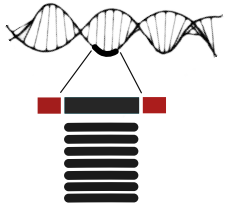


Metabarcoding steps

Sampling

DNA extraction

Barcode amplification



Sequencing

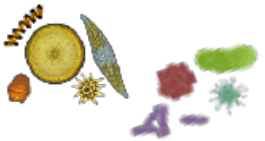
Bioinformatic analysis

Recovery of all species

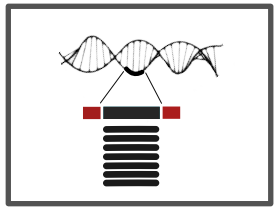
Barcode selection

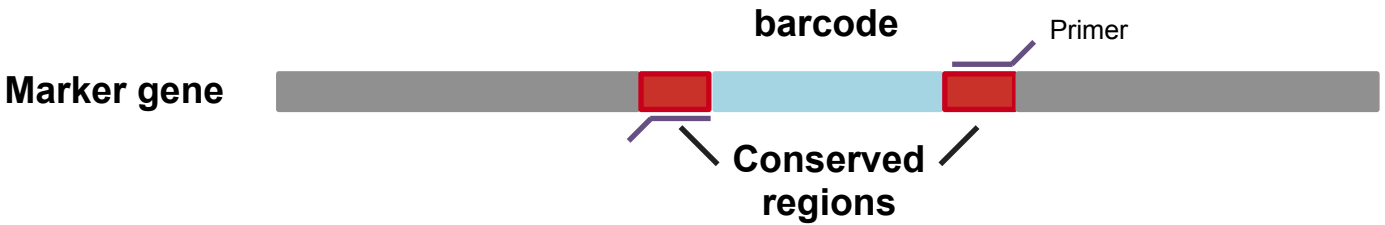
Taxonomic resolution and identification

Specificity

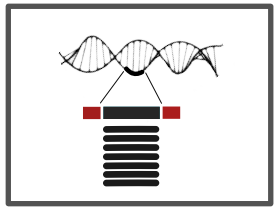


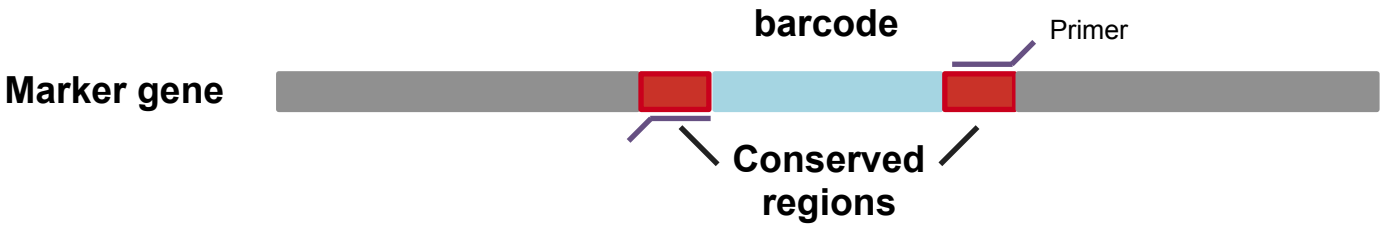
Barcode selection - a fundamental criteria





Barcode selection - a fundamental criteria

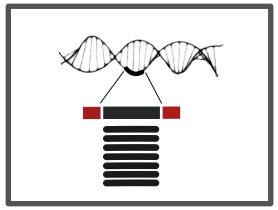


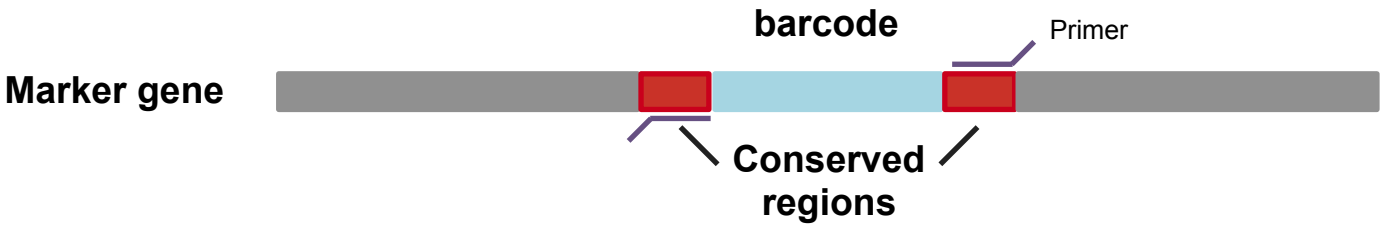


1 Universal

Funded by European Union

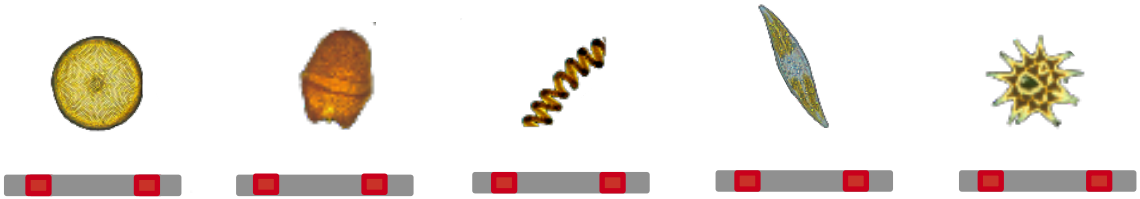
Barcode selection - a fundamental criteria



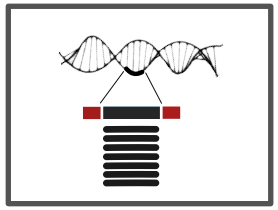


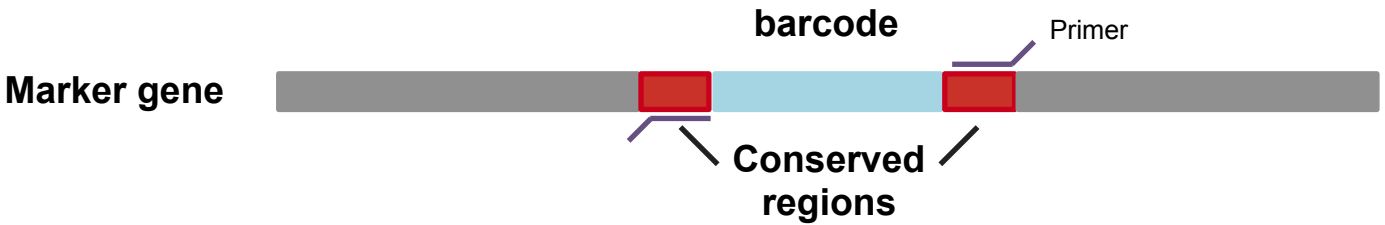
1 Universal

2 Conserved regions



Barcode selection - a fundamental criteria

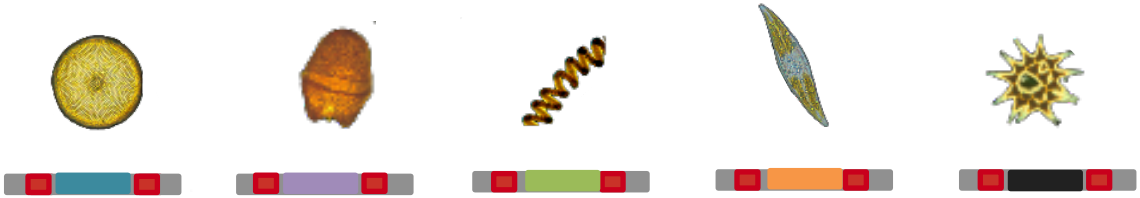




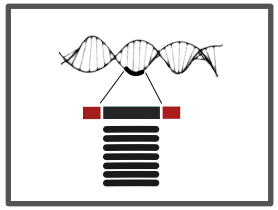
1 Universal

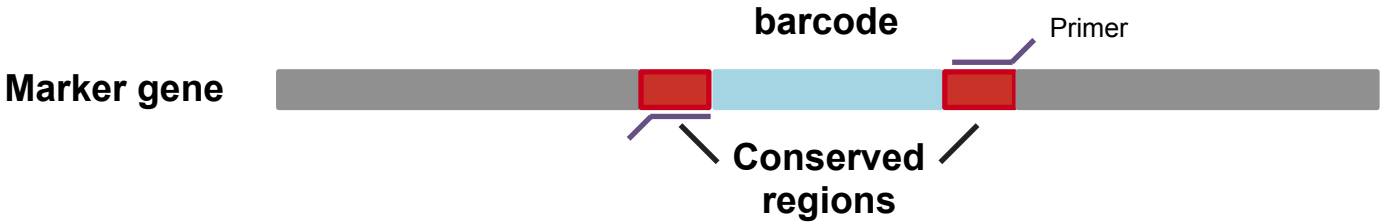
2 Conserved regions

3 Variable enough

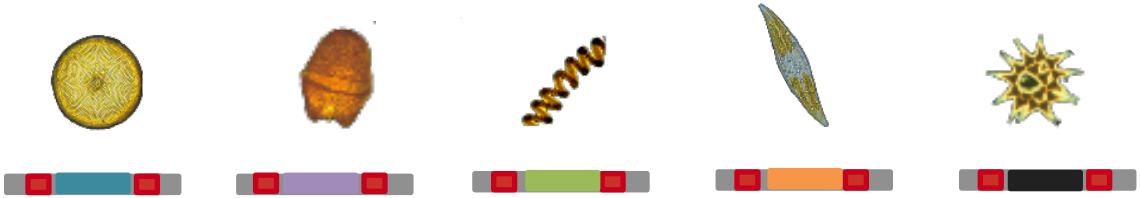


Barcode selection - a fundamental criteria

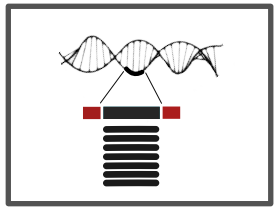


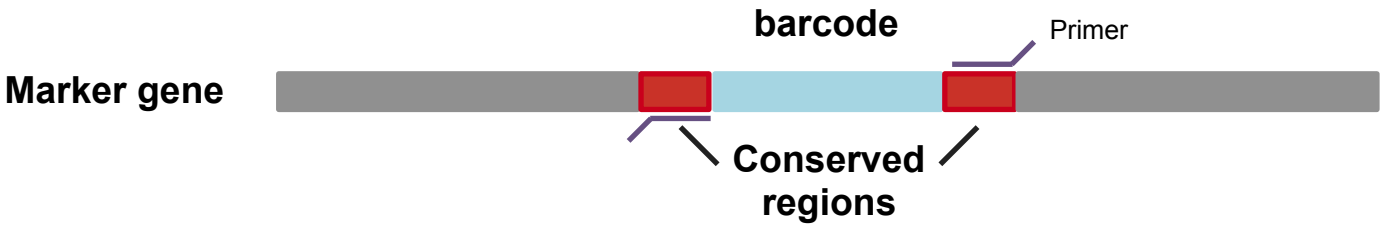


1. Universal
2. Conserved regions
3. Variable enough
4. Match sequencing technology size

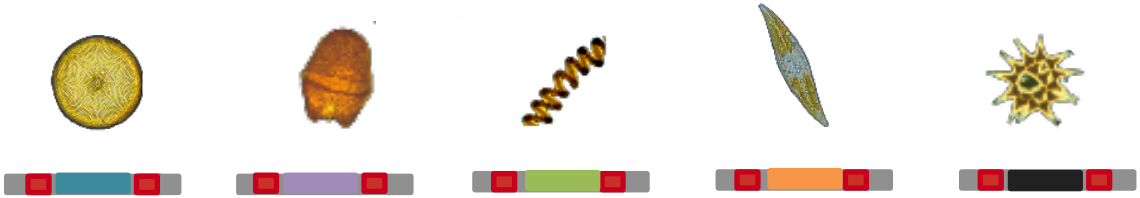


Barcode selection - a fundamental criteria

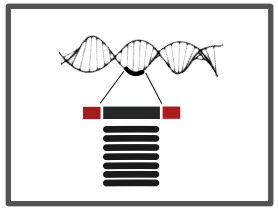


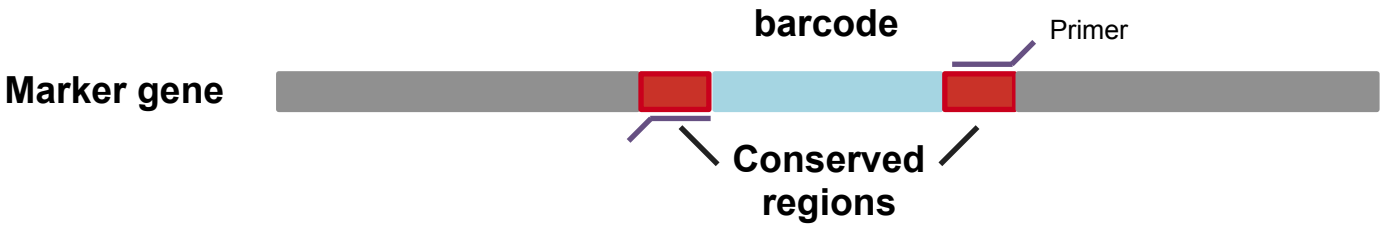


1. Universal
2. Conserved regions
3. Variable enough
4. Match sequencing technology size
5. Represented in reference libraries



Barcode selection - a fundamental criteria



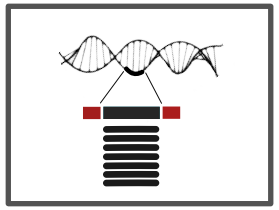


1. Universal
2. Conserved regions
3. Variable enough
4. Match sequencing technology size
5. Represented in reference libraries

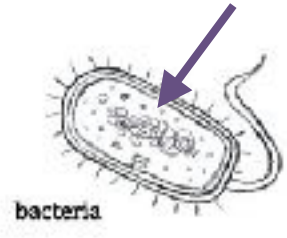
What are the classical marker genes used in metabarcoding?

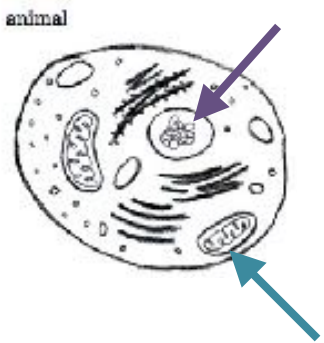
Funded by European Union

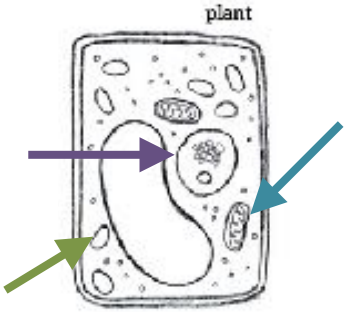
Classical marker genes - an overview



- A marker gene can be present in all organisms genomes







Genomic DNA

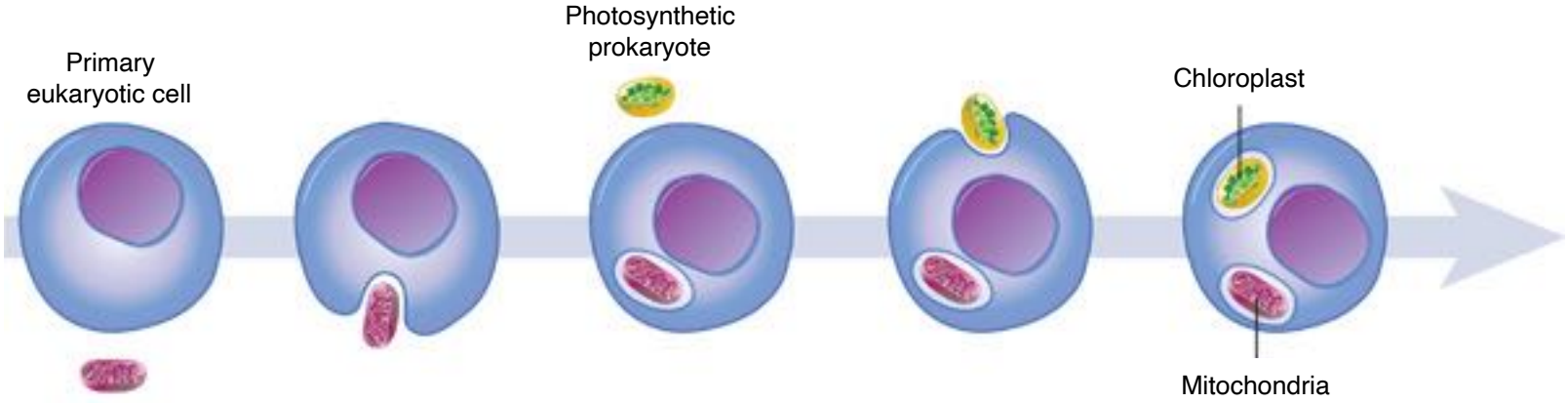
Mitochondrial DNA

Chloroplast DNA

Funded by European Union

Classical marker genes - an overview

- A marker gene can be present in all organisms genomes



Plastids DNA are former DNA of prokaryotes

Funded by European Union

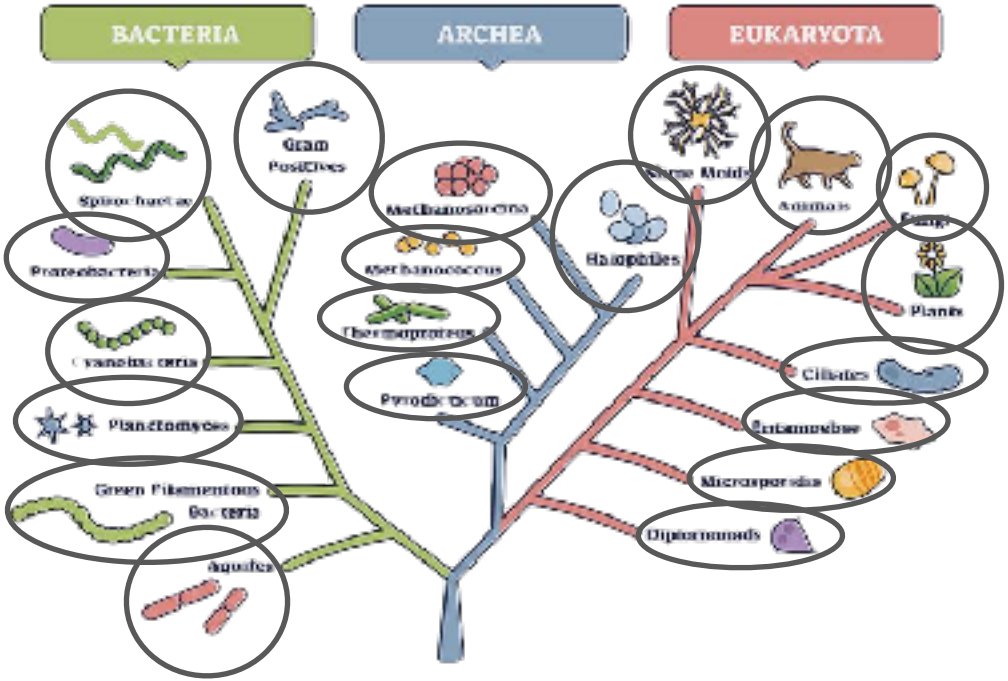
Classical marker genes - an overview

Ribosomal proteins

Nuclear

Mitochondrial

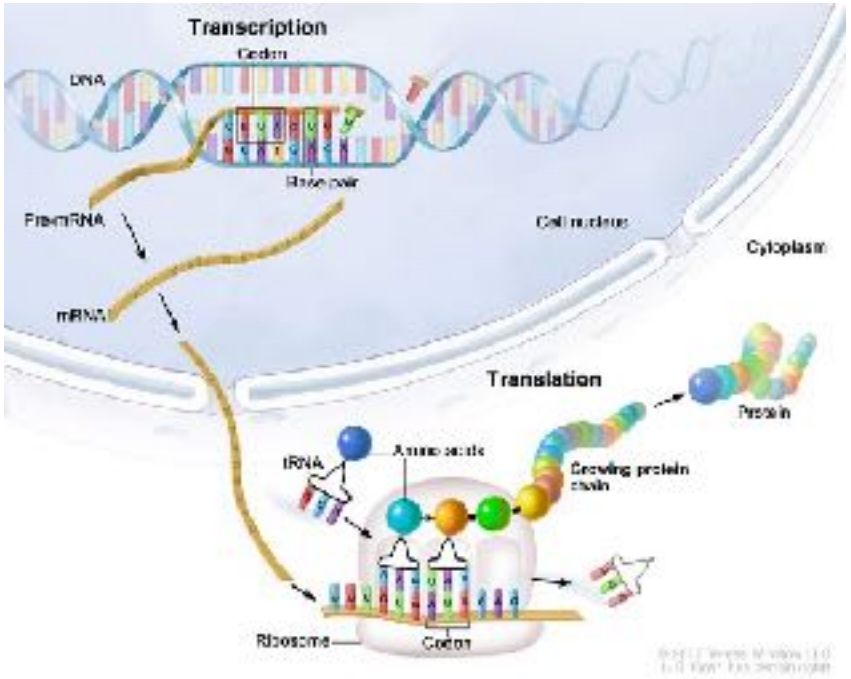
Chloroplast



Classical marker genes - an overview

Ribosomal genes

- Code for the large and small units of ribosomes which are responsible of the translation of mRNA into proteins
- Present in the nuclear DNA of **all living organisms**
- Present in the plastid DNA (mitochondria, chloroplasts)

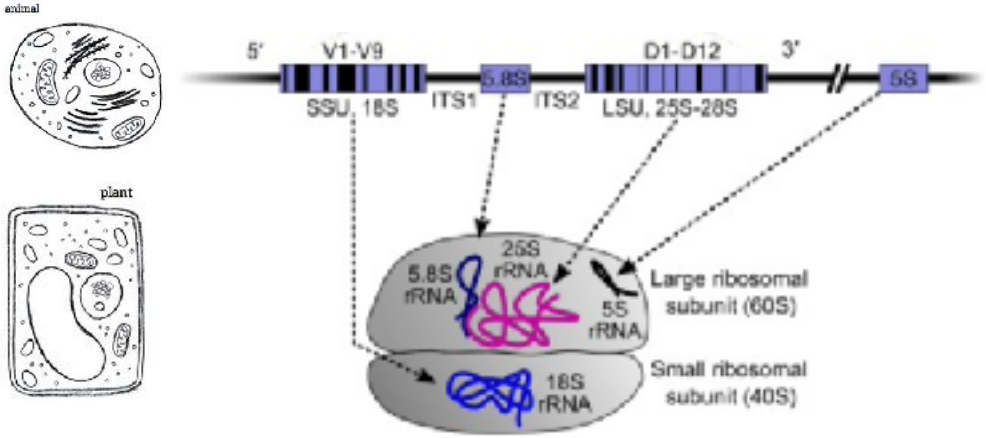


Funded by European Union

Classical marker genes - an overview

Ribosomal genes

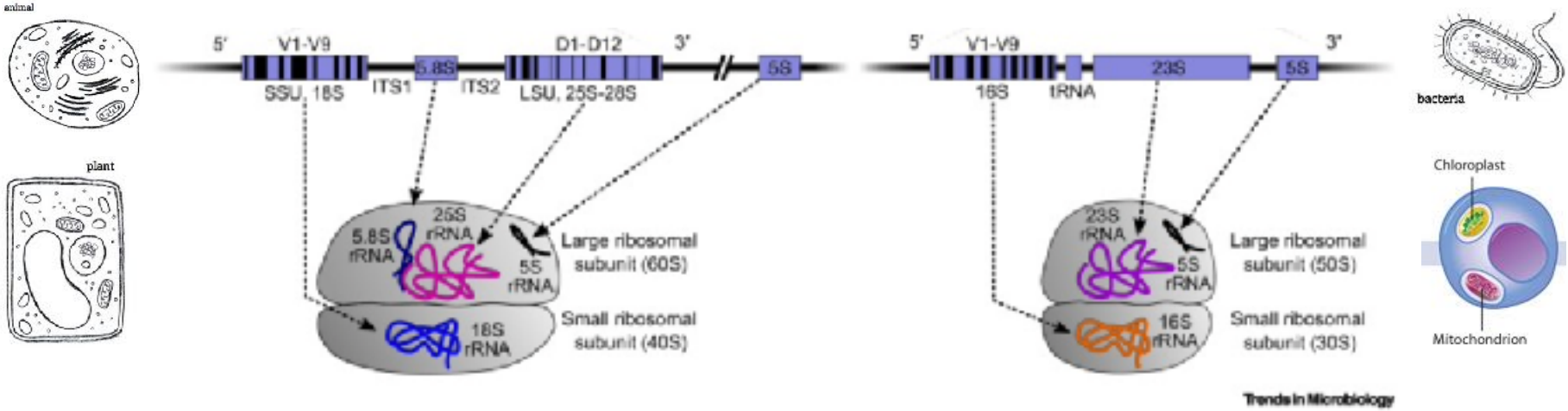
- Genes organized in an operon (*i.e.* clustered in the genome)



Classical marker genes - an overview

Ribosomal genes

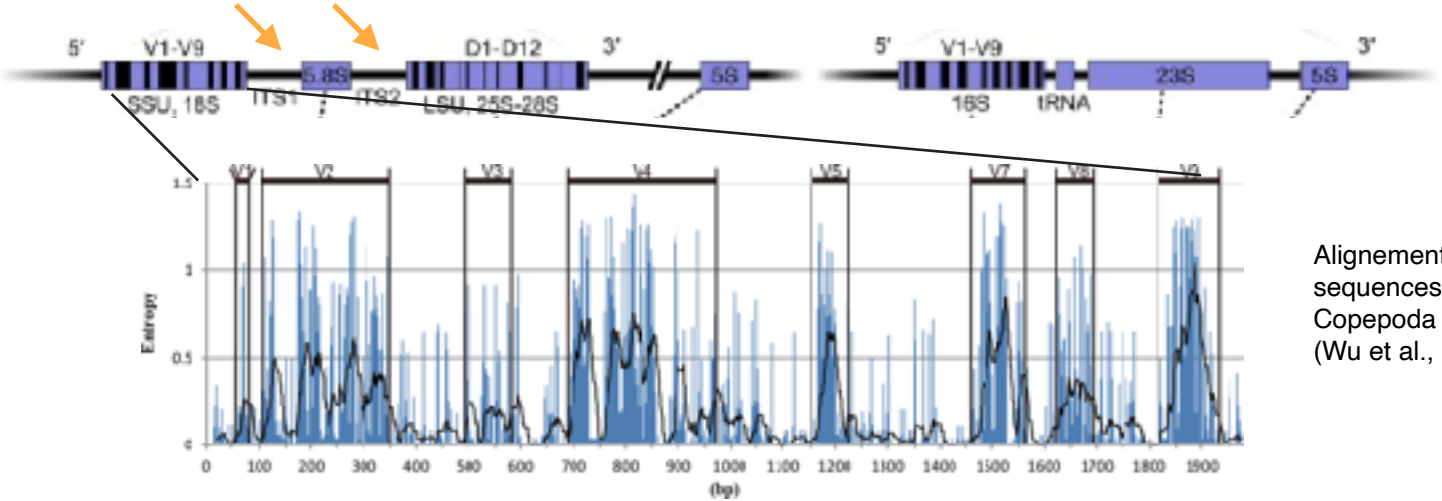
- Genes organized in an operon (*i.e.* clustered in the genome)



Classical marker genes - an overview

Ribosomal genes

- Alternance of hyper variable and conserved regions + Internal Transcribe spacer (ITS)



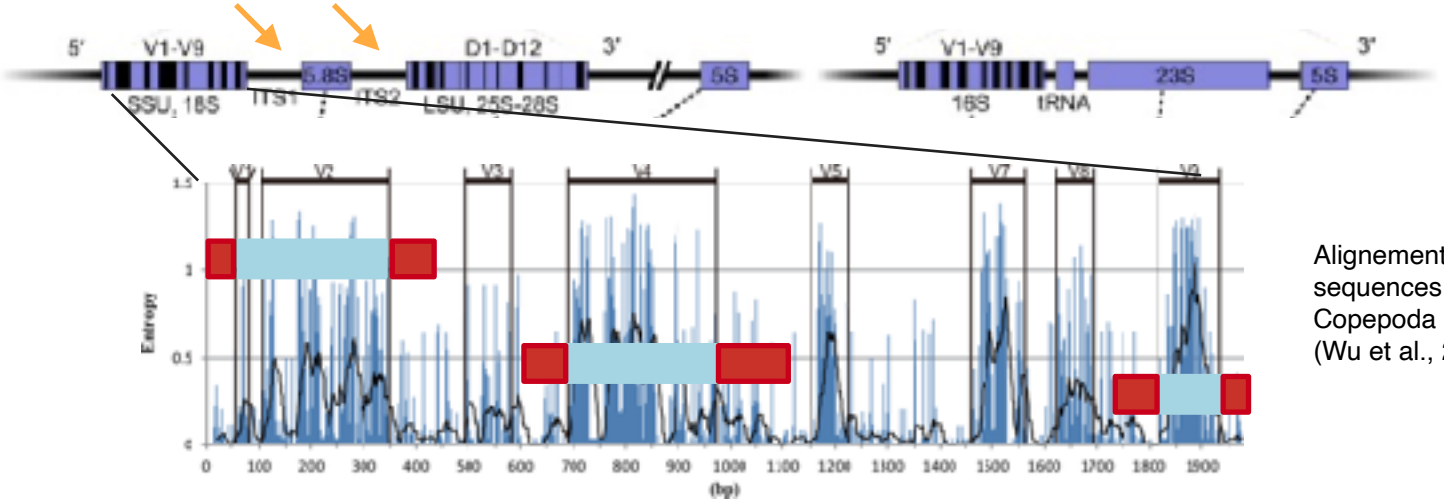
Alignement of 18S sequences from 192 Copepoda species (Wu et al., 2015)

Funded by European Union

Classical marker genes - an overview

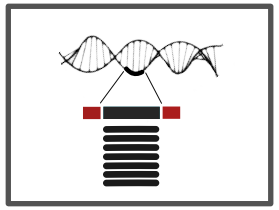
Ribosomal genes

- Alternance of hyper variable and conserved regions + Internal Transcribe spacer (ITS)

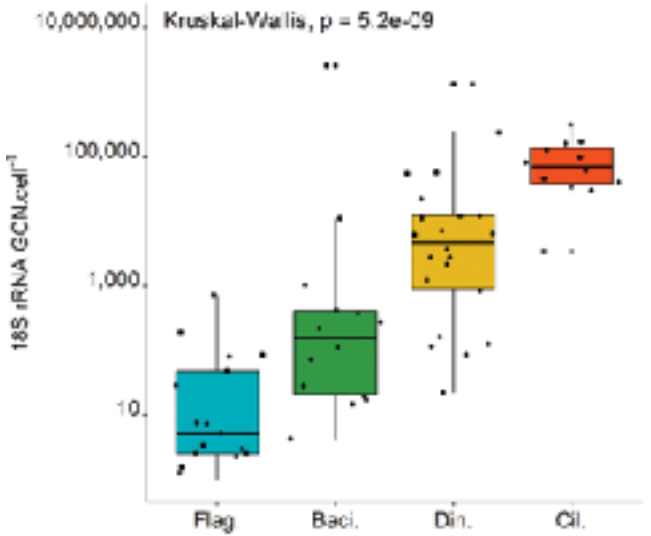


Alignement of 18S sequences from 192 Copepoda species (Wu et al., 2015)

Classical marker genes - an overview

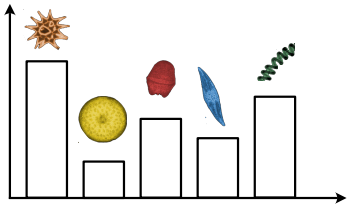


Various copy numbers between taxa



Martin et al., 2022

To meet the huge biosynthetic demand, eukaryotic cells contains hundreds to thousands of copies of ribosomal genes. (Kobayashi et al., 2011)



Influence on abundance

Funded by European Union

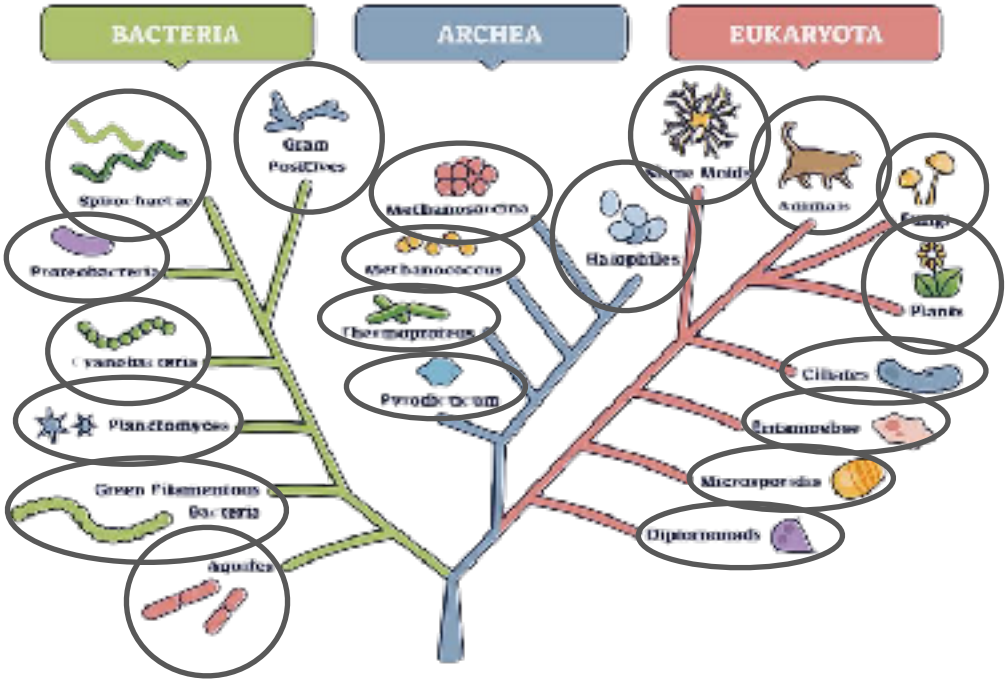
Classical marker genes - an overview

Ribosomal proteins

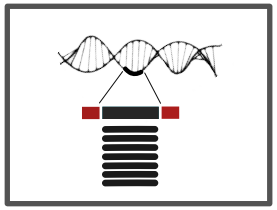
Nuclear

Mitochondrial

Chloroplast



Classical marker genes - an overview

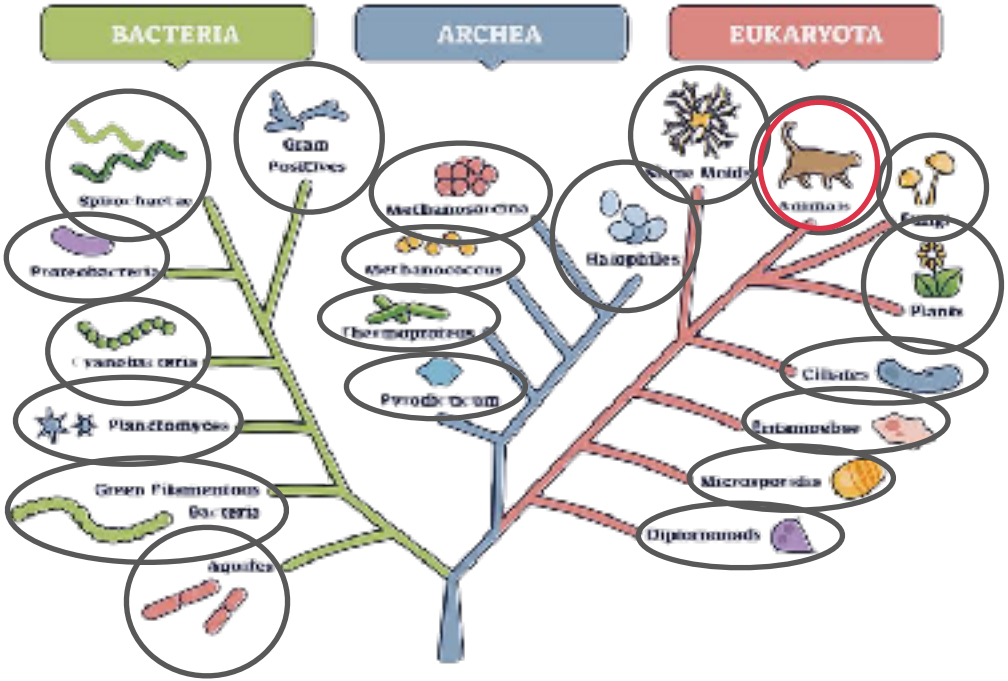


Ribosomal proteins

Nuclear

Mitochondrial

Chloroplast

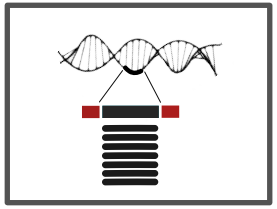


COI

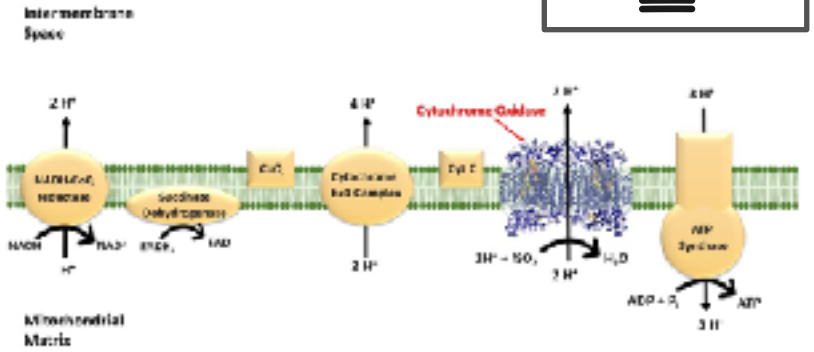
Mitochondrial

Funded by European Union

Classical marker genes - COI gene

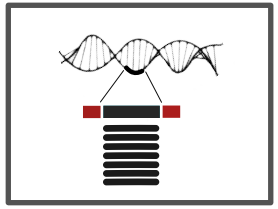


COI

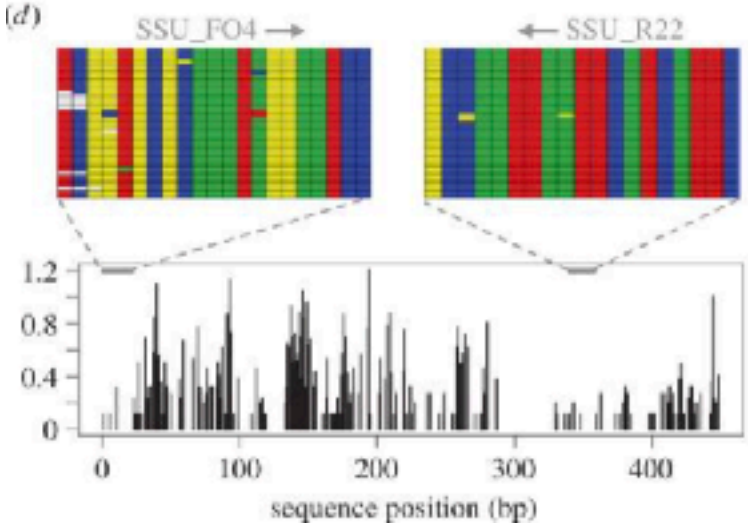


- COI codes for the Cytochrome Oxidase subunit I
  It is a key mitochondrial enzyme for respiration
- Historically it is used for barcoding of animals. lot of primers has been designed for various animals group
- COI often provides a better taxonomic resolution than nuclear rRNA (18S)
- But contrary to ribosomal genes, it is complex to define universal primers.

Classical marker genes - COI gene



Nuclear 18S

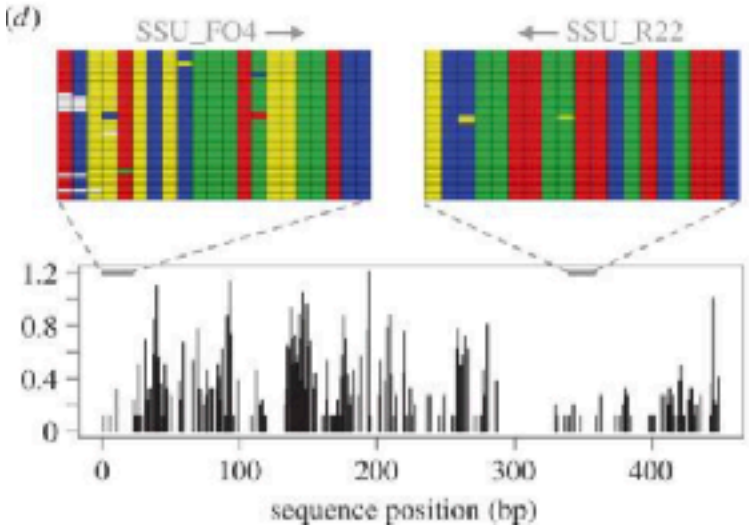


From more than 40 species of insects among 25 different orders (Deagle et al. 2014)

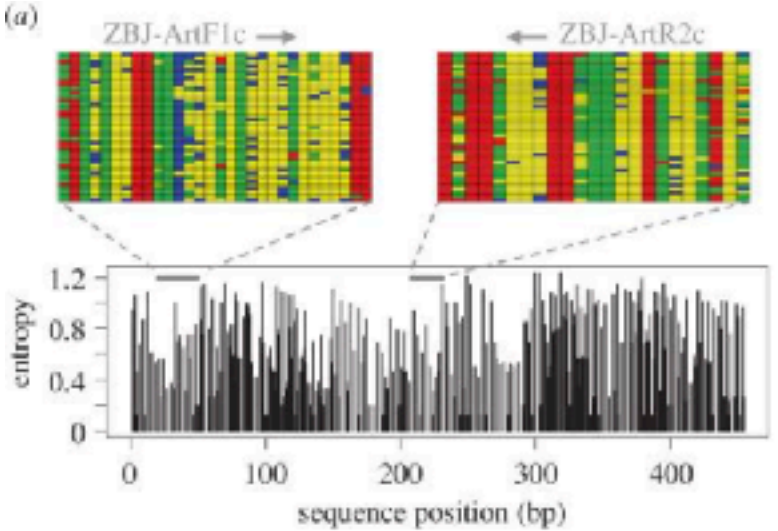
Funded by European Union

Classical marker genes - COI gene

Nuclear 18S

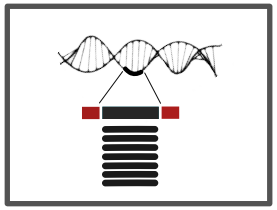


COI



From more than 40 species of insects among 25 different orders (Deagle et al. 2014)

Classical marker genes - an overview

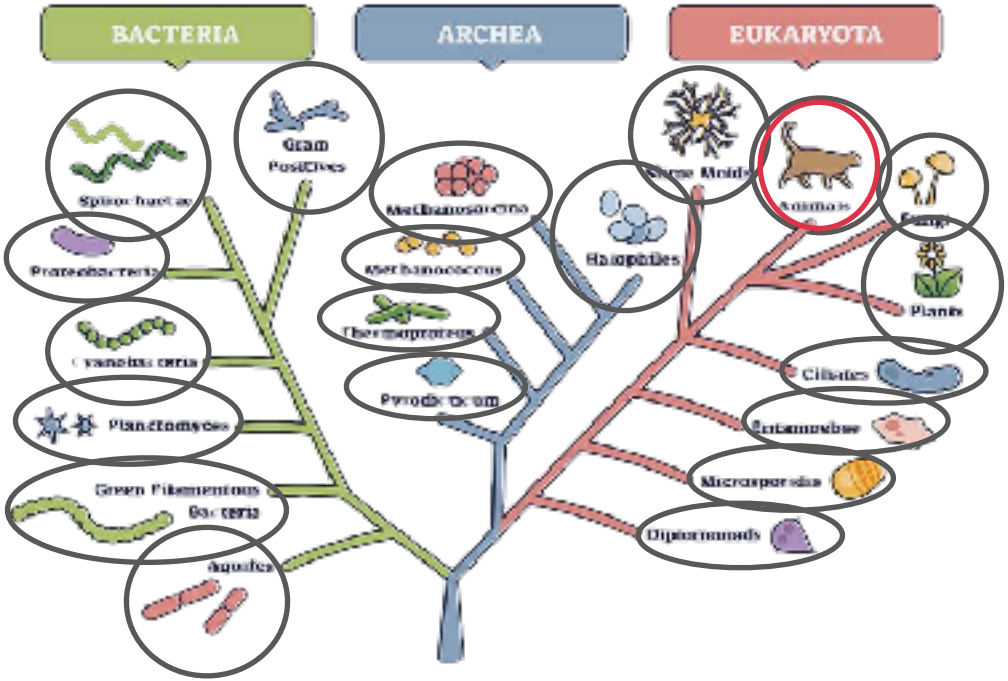


Ribosomal proteins

Nuclear

Mitochondrial

Chloroplast



COI

Mitochondrial

Funded by European Union

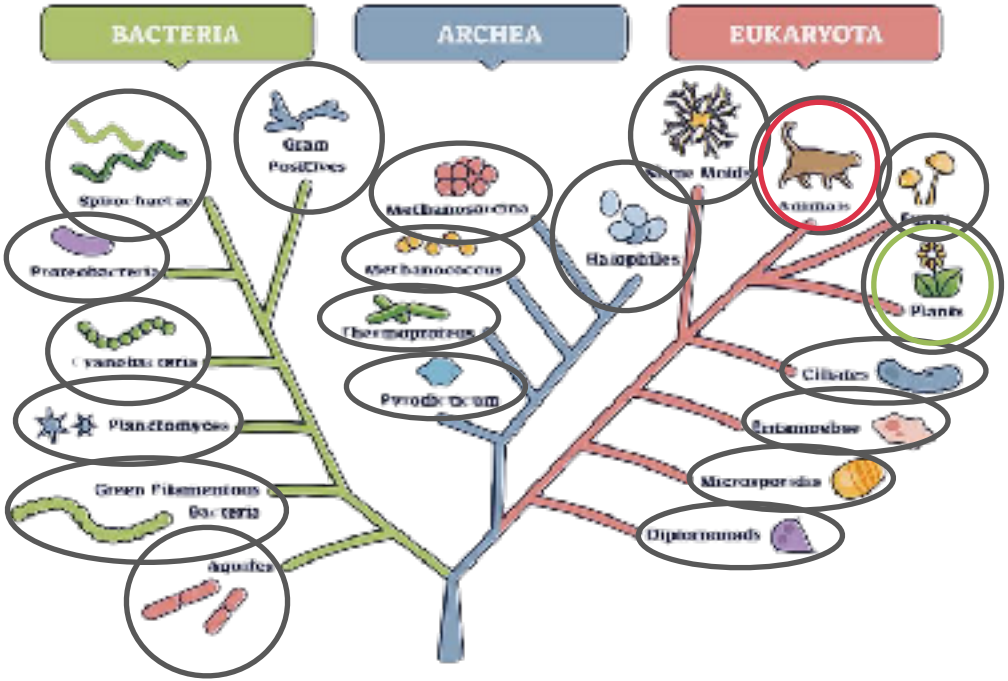
Classical marker genes - an overview

Ribosomal proteins

Nuclear

Mitochondrial

Chloroplast



COI

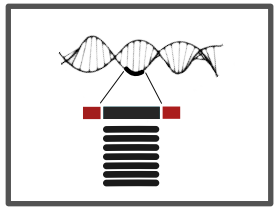
Mitochondrial

Rbcl/matK

Chloroplast

Funded by European Union

Classical marker genes



Rbcl/MatK - chloroplastic genes to target photosynthetic organisms

- For plant, low substitution rates of mitochondrial DNA - COI is not a good target
- Genes present in chloroplasts - target specifically photosynthetic organisms

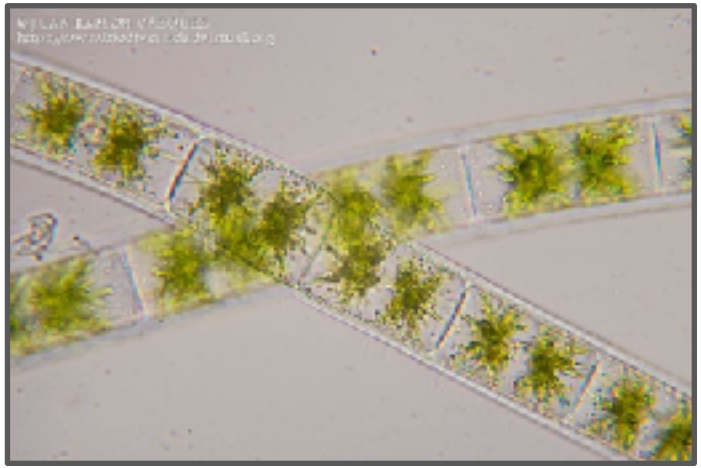
MatK : plastid genes that codes for an intron maturase

Rbcl : codes for the large subunit of RuBisCo, a key enzyme of photosynthesis.
Difficulties to get target all the phytoplanktonic diversity

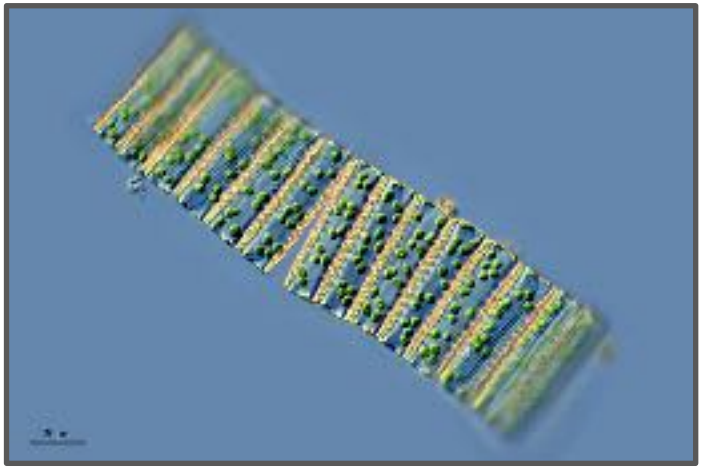
Ribosomal genes (16S)

Classical marker genes - an overview

- A marker gene in chloroplasts and mitochondria can be present in variable numbers in the cell



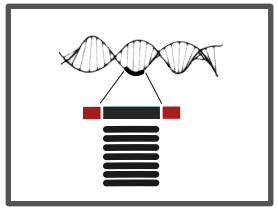
Zygnema sp. 1-2 chloroplasts/cell

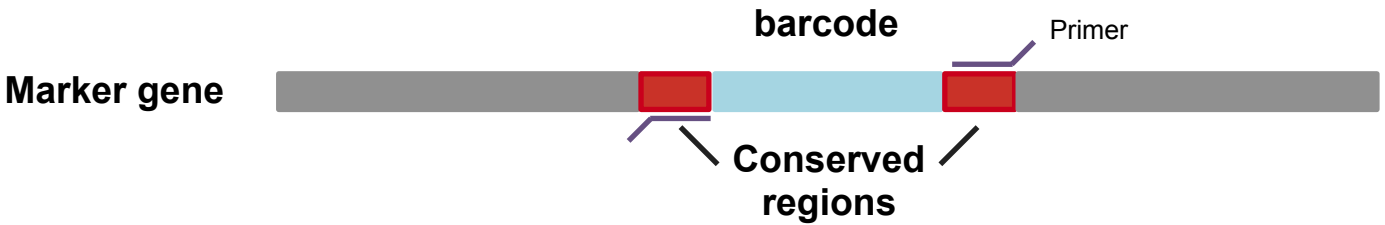


Diatoma sp. > 10 chloroplasts/cell

Funded by European Union

Barcode selection - a fundamental criteria



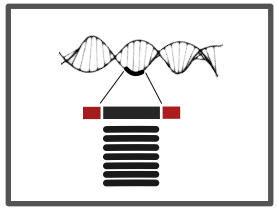


1. Universal
2. Conserved regions
3. Variable enough
4. Match sequencing technology size
5. Represented in reference libraries

How primers are designed to hybride only in the DNA of all targeted organisms?

Specificity

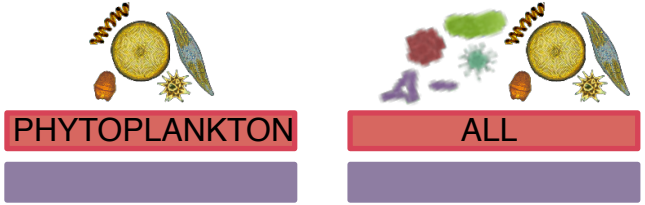
Funded by European Union

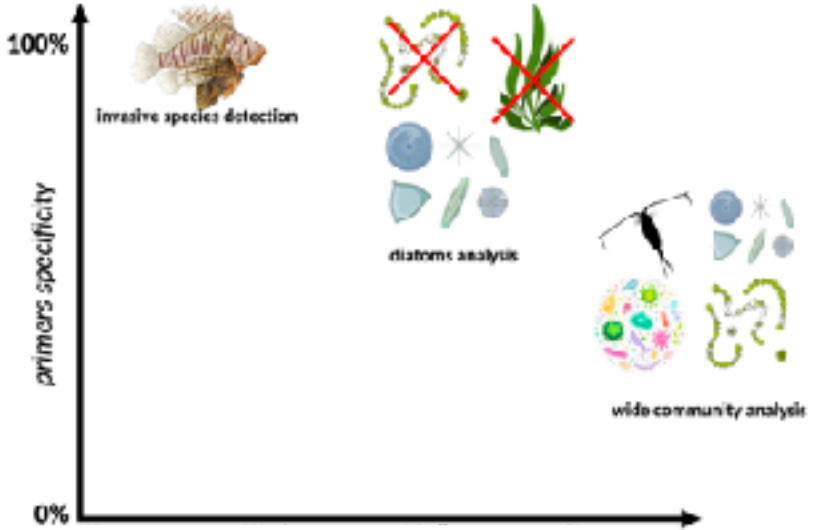


Conserved regions : specificity

Conserved regions : specificity

Primer specificity :

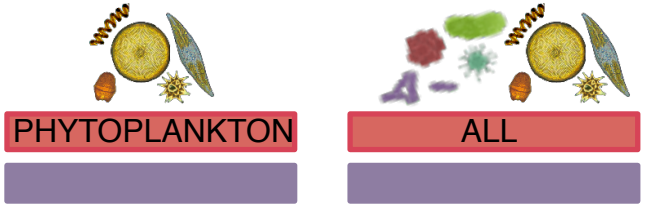




Funded by European Union

Conserved regions : specificity

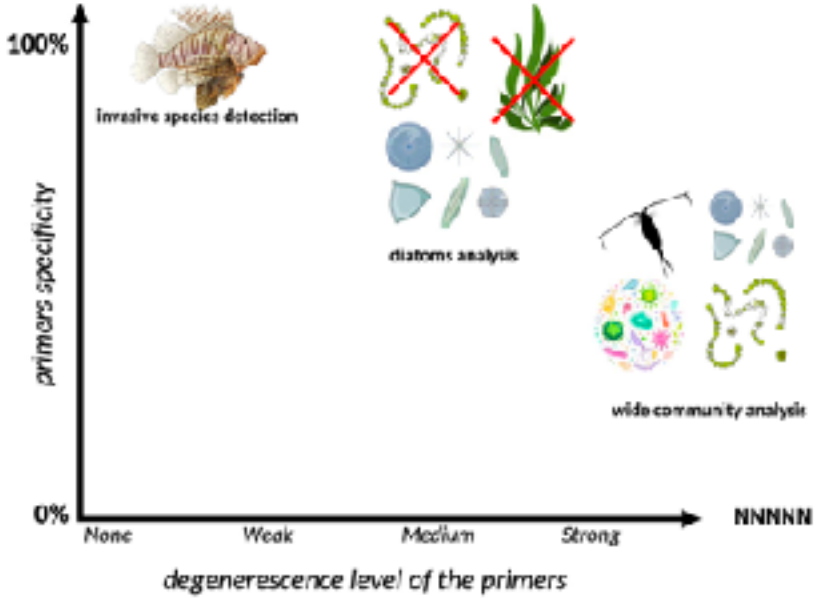
Primer specificity :



Primer degenerescence :

GAGCCTAATAACC
GAACCTCATAACC
GAACCTAATAACC
CTYGGAKTATTGG

Code	Description
M	AC
R	AG
W	AT
S	CG
Y	CT
K	GT
V	ACG
H	ACT
D	AGT
B	CGT
N	ACGT

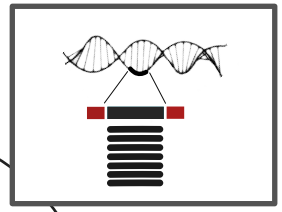


Funded by European Union

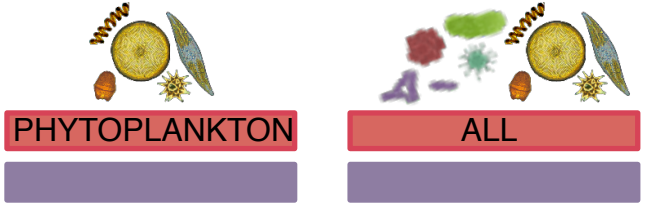
BIOLAWEB

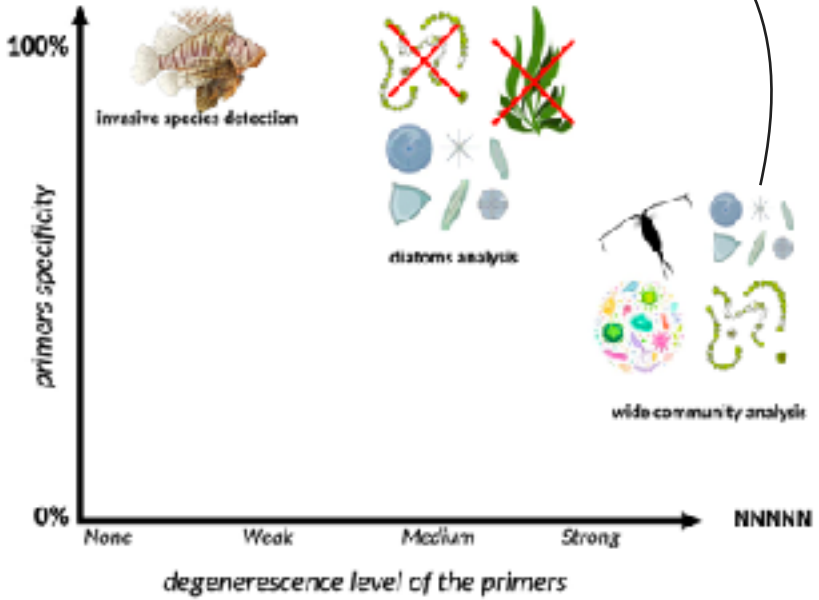
Conserved regions : specificity

515F GTGYTCAGCMGCCGCGGTAA
926R CCGYCAATTYMTTTRAGTTT

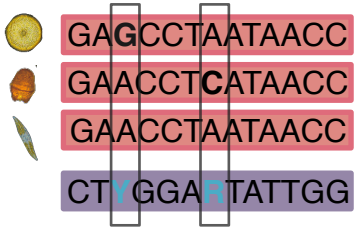


Primer specificity :





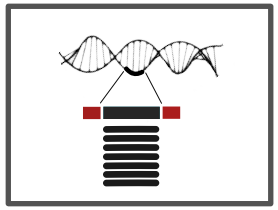
Primer degenerescence :

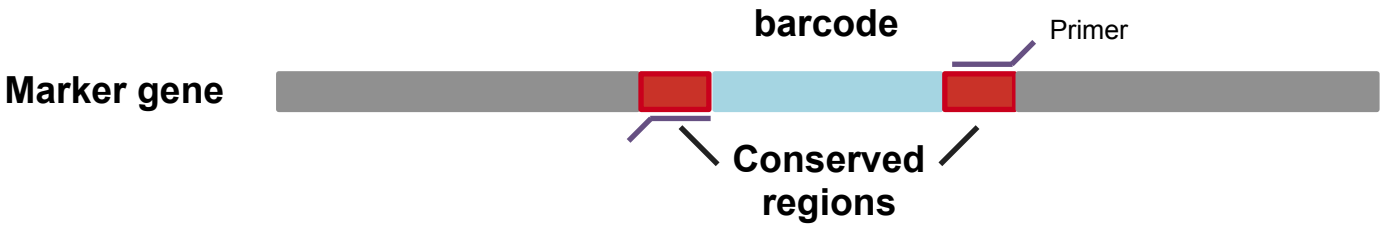


Code	Description
M	AC
R	AG
W	AT
S	CG
Y	CT
K	GT
V	ACG
H	ACT
D	AGT
B	CGT
N	ACGT

Funded by European Union

Barcode selection - a fundamental criteria



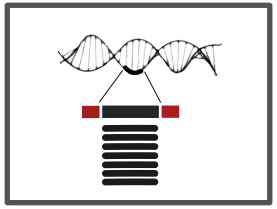


1. Universal
2. Conserved regions
3. Variable enough
4. Match sequencing technology size
5. Represented in reference libraries

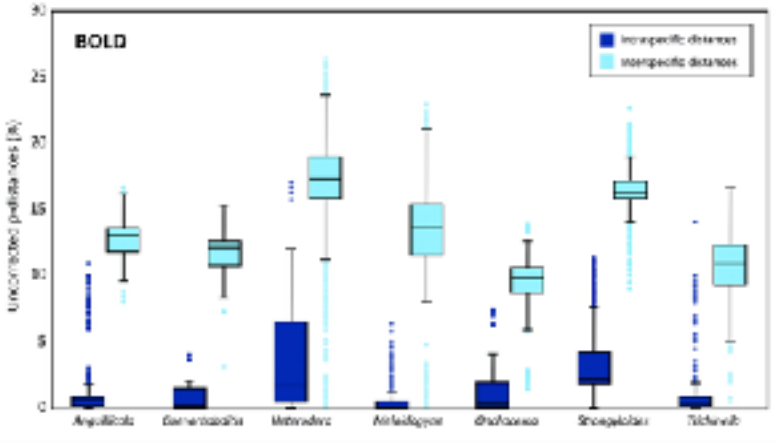
Barcode-gap efficiency for a great diversity of organisms in metabarcoding

Funded by European Union

Variable enough

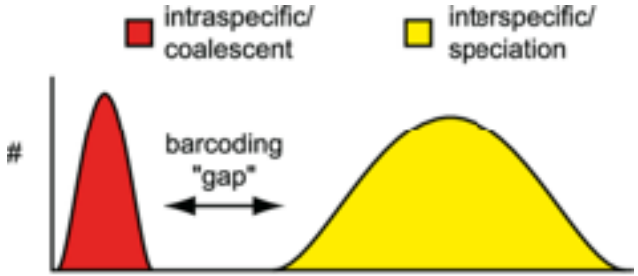


Depends on the taxonomic group



Tresoldi-Gonsaves et al., 2021

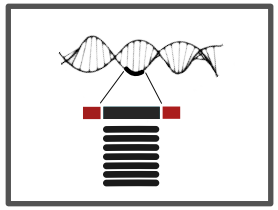
COI of various genera of worms (Nematoda)

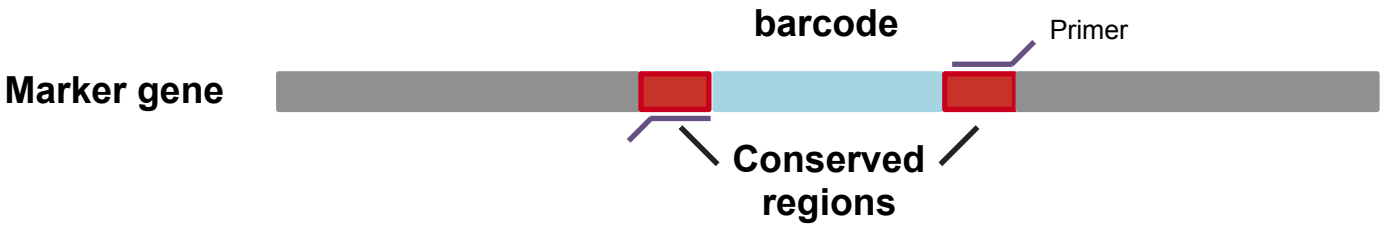


There is often a trade-off between capacity to target a large diversity and the barcode variability

Funded by European Union

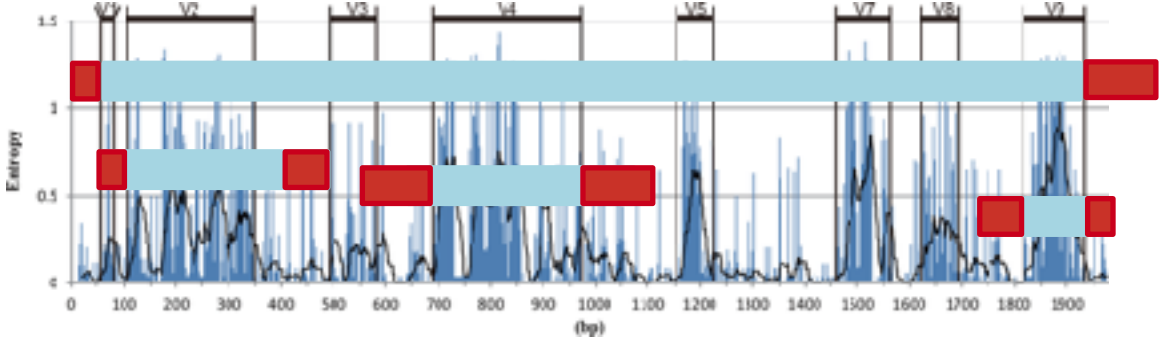
Barcode selection - a fundamental criteria





1. Universal
2. Conserved regions
3. Variable enough
4. Match sequencing technology size
5. Represented in reference libraries

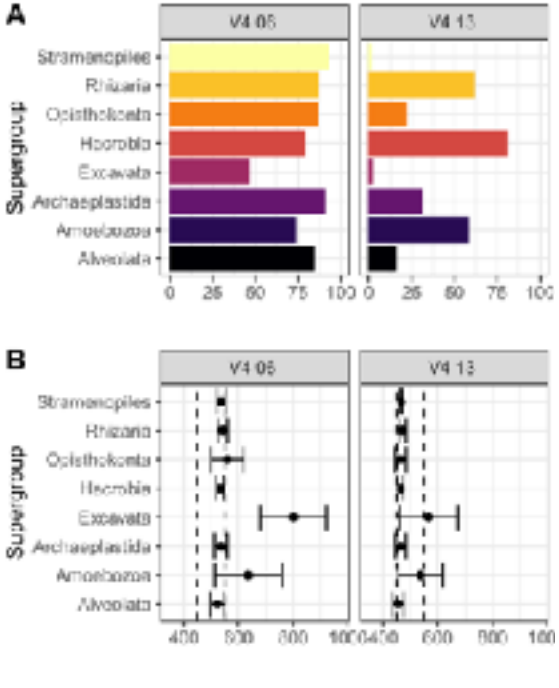
Match sequencing technology size



ADNr 18S : 1800bp
ADNr 16S : 1500bp
Rbcl : 1400bp
MatK : 1500bp
COI : 650bp

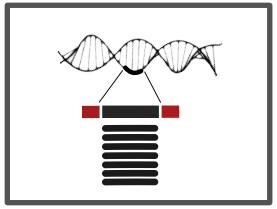
Funded by European Union

Match sequencing technology size



870

Geisen et al., 2019

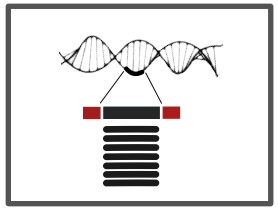


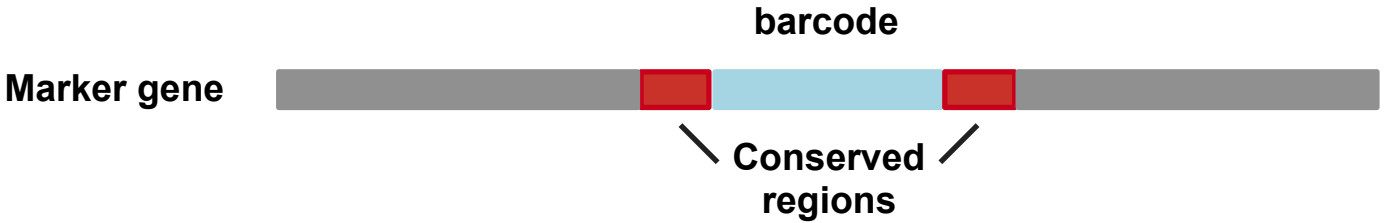
For the same primers, the amplicon size will vary between major supergroup of protists

A broader diversity of protist will be recovered with long-read sequencing

Funded by European Union

Barcode selection



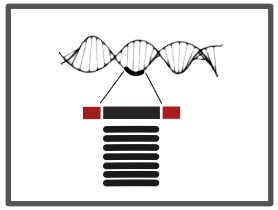


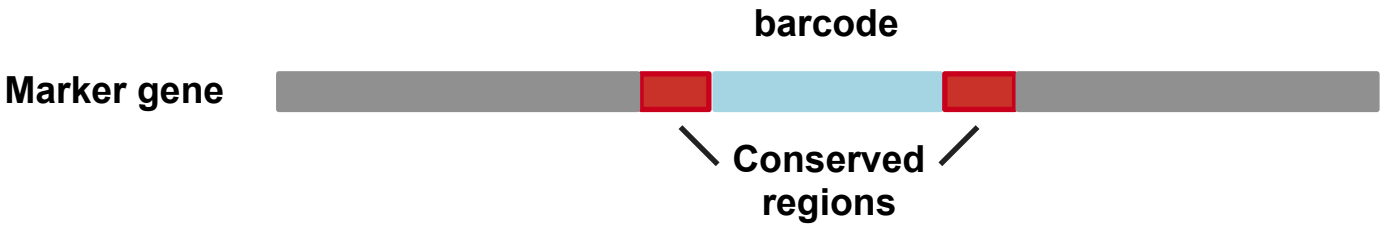
1. Universal
2. Conserved regions
3. Variable enough
4. Match sequencing technology size
5. Represented in reference libraries

<u>The perfect barcode does not exist</u>

Particularly if you want to target very diverse groups of organisms (e.g. all the protists, all the macroinvertebrates, all the phytoplankton)

Barcode selection





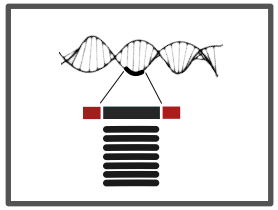
1. Universal
2. Conserved regions
3. Variable enough
4. Match sequencing technology size
5. Represented in reference libraries

<u>The perfect barcode does not exist</u>

Find the one that best suits to your research question

Funded by European Union

How to select a barcode ?



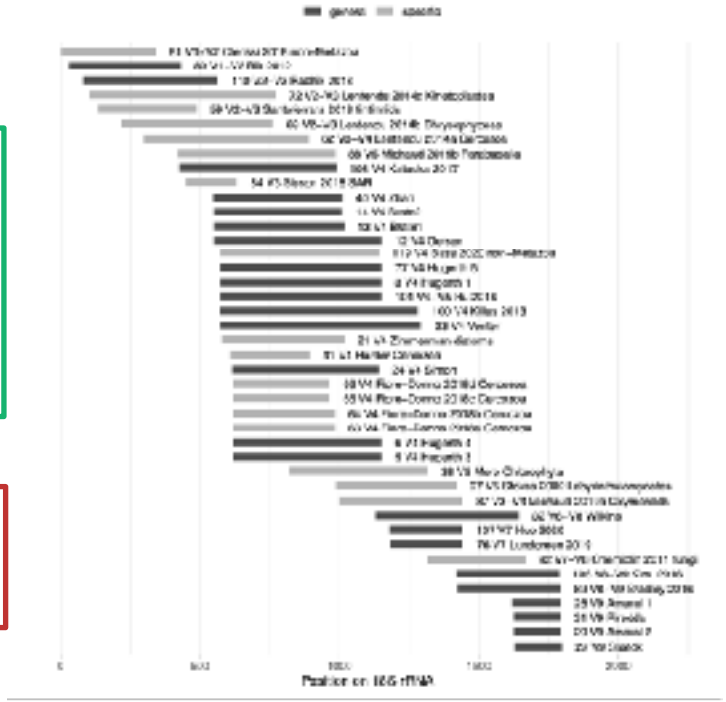
In literature

- Comparison between dataset

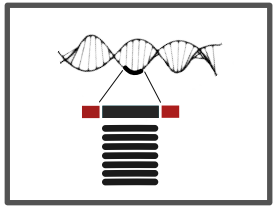
- Known bias

- Tested

- Possibly not adapted to your study/environment



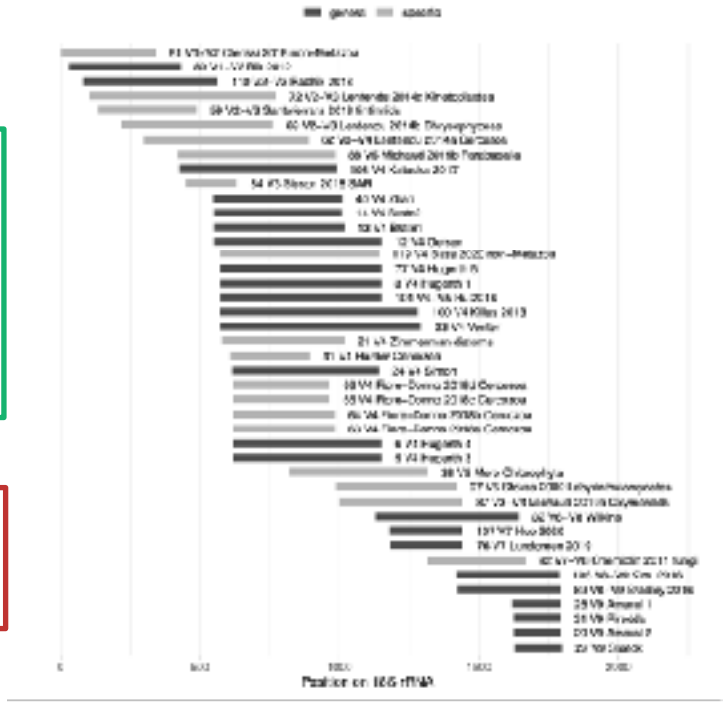
How to select a barcode ?



In literature

- Comparison between dataset
- Known bias
- Tested

- Possibly not adapted to your study/environment



Design your own

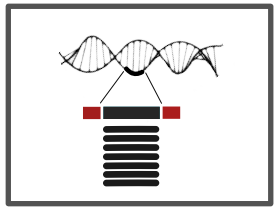
- Most adapted to your study
- Up to date reference libraries

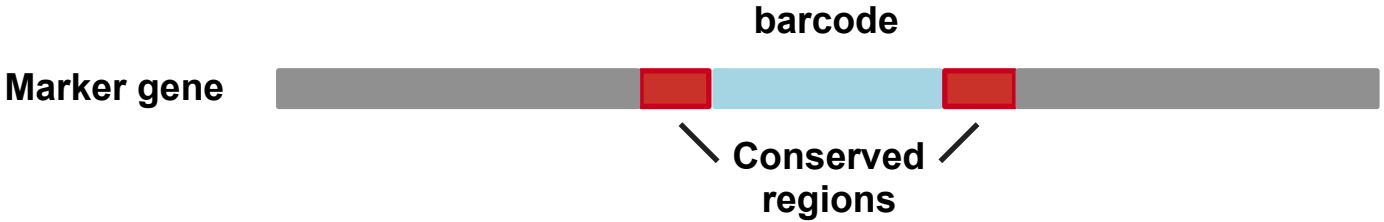
Time-consuming !

- A lot of optimization steps

Funded by European Union

Barcode selection





1. Universal
2. Conserved regions
3. Variable enough
4. Match sequencing technology size
5. Represented in reference libraries

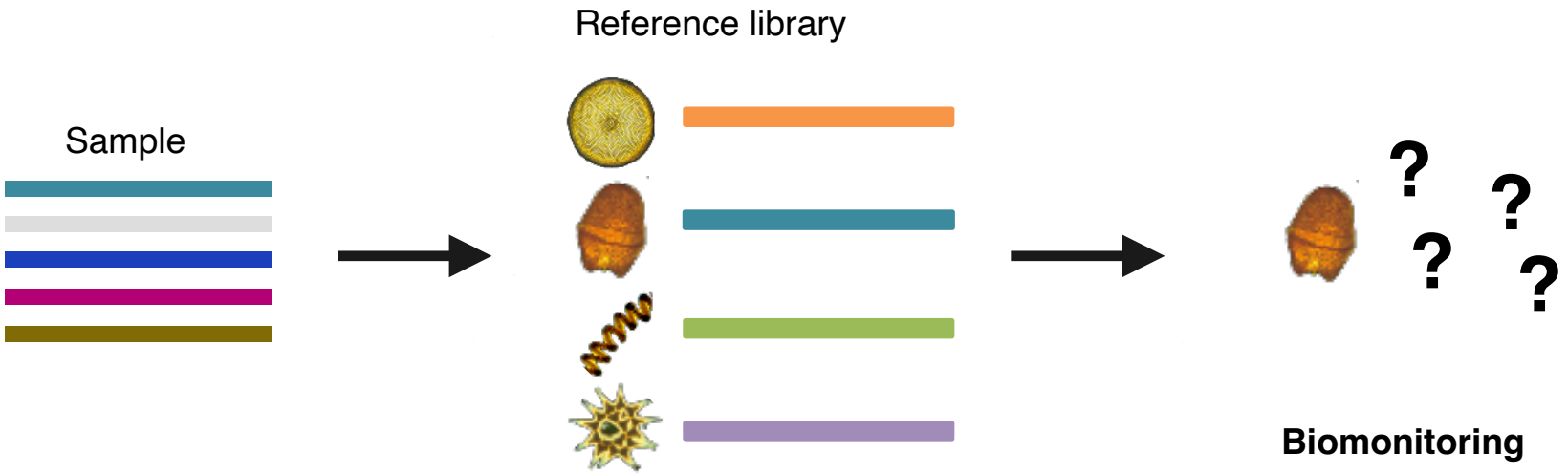
« The power of DNA metabarcoding is directly proportional to data available in reference databases »

Funded by European Union

Reference libraries

ATCGCTTTGGACCT
ATCGCTTTGGACCT
ATCGCTTTGGACCT

They are a prerequisite to identify your sequence and give it a taxonomy

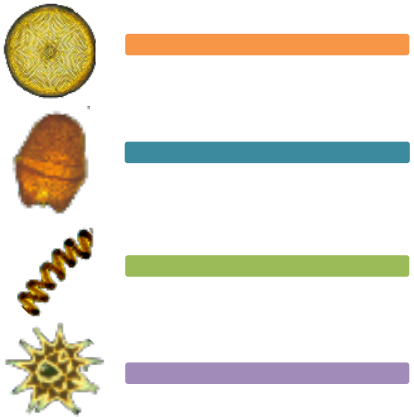


Funded by European Union

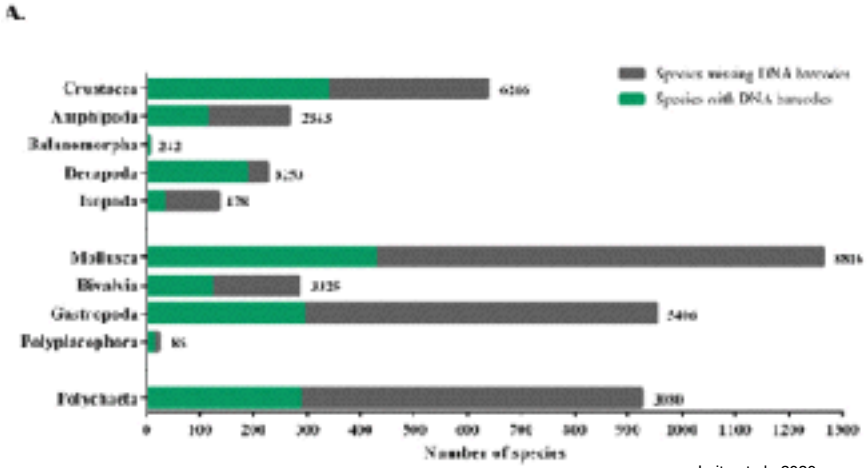
Reference libraries

They are a prerequisite to identify your sequence and give it a taxonomy

Reference library



→ The most complete as possible



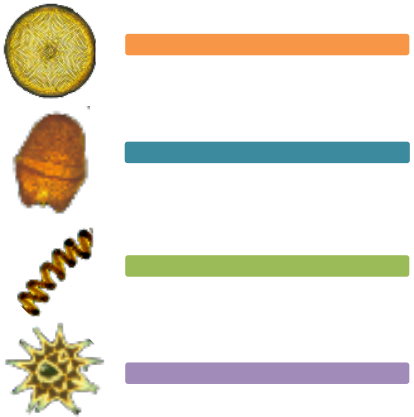
Leite et al., 2020

Reference libraries

ATCGCTTTGGACCT
ATCGCTTTGGACCT
ATCGCTTTGGACCT

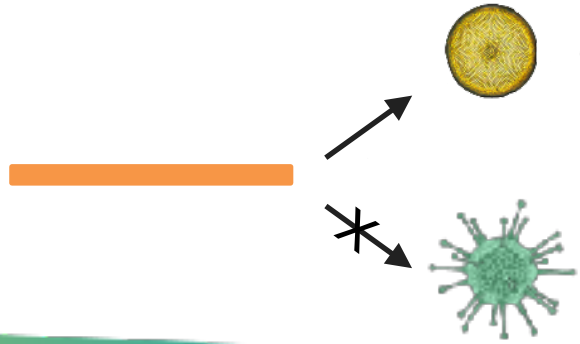
They are a prerequisite to identify your sequence and give it a taxonomy

Reference library



→ The most complete as possible

→ The most curated as possible



Up to date taxonomy

Funded by European Union

Reference libraries - ribosomal genes

Generalists

silva
high quality ribosomal RNA databases

SILVA SSU 138.1 update release

	SSU Parc	SSU Ref NR 99	LSU Parc	LSU Ref NR 99
Minimal length	300	1200/900	300	1900
Quality filtering	basic	strong	basic	strong
Guide Tree	no	yes	no	yes
Release date	27.08.20	27.08.20	27.08.20	27.08.20
Aligned rRNA sequences	9,469,124	510,508	1,312,534	95,286

For bacteria, archaea and eukaryotes
Both nuclear and plastid SSU and LSU

Specialists

PR²

Nuclear 18S

200,000 sequences

For protists

PhytoRef

Plastidal 16S

6486 sequences

Phytool_v2

Nuclear 18S
Plastidal 16S, 23S

18S : *16654*
16S : *8479*
23S : *1997*

For phytoplankton

Reference libraries - other genes

Generalists

BOLD SYSTEMS

BOLD: The Barcode of Life Data System (www.barcodinglife.org)

SUJEEVAN RATNASINGHAM and PAUL D. N. HEBERT

Around 1.3M of COI barcodes



Specialists

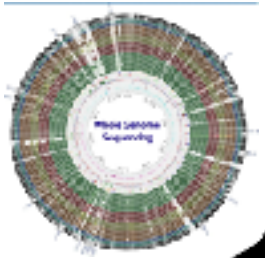
Diatbarcode	COIns
	
rbcl	COI
4000-5000 sequences	532 000 sequences
Diatoms	Insects

Reference libraries - non curated database

Your study

Big eDNA projects

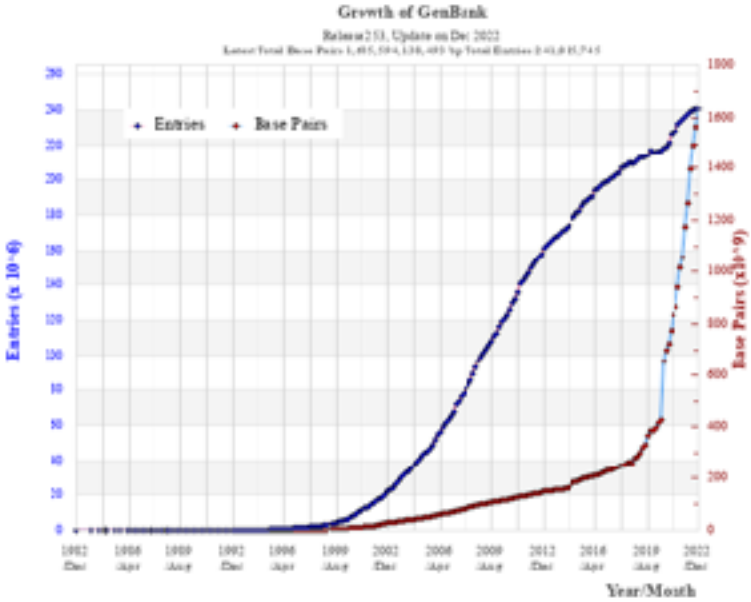
WGS



ENA
European Nucleotide Archive

DDBJ
DNA Data Bank of Japan

NCBI



For GenBank : 240 millions of sequences

High potential but **not curated** !

Reference libraries

All sequences

ENA European Nucleotide Archive

DDBJ DNA Data Bank of Japan

NCBI

Quality-checked

silva high quality ribosomal RNA databases

Curated taxonomy

PR²

PhytoRef

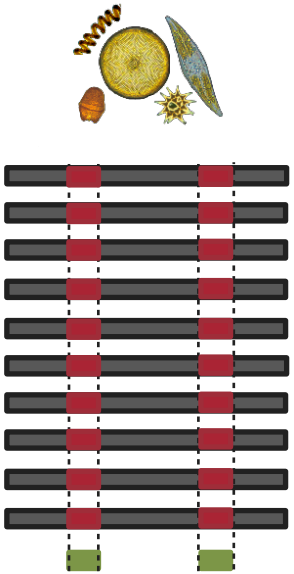
Funded by European Union

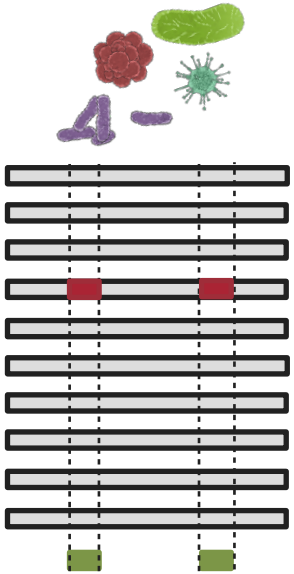
Reference libraries

ATCGCTTTGGACCT
ATCGCTTTGGACCT
ATCGCTTTGGACCT

But also they are essential to find the best design for the primers

Optimal design of the primers





Test for primer specificity

Funded by European Union

Reference libraries

ATCGCTTTGGACCT
ATCGCTTTGGACCT
ATCGCTTTGGACCT

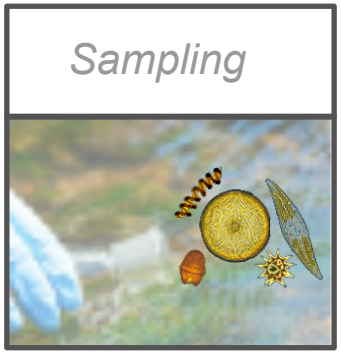
But also they are essential to find the best design for the primers

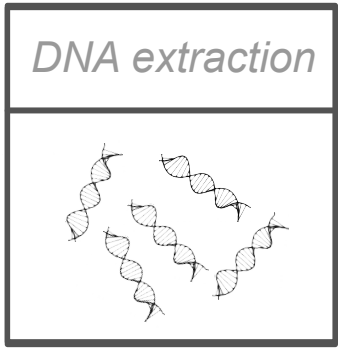
Optimal design of primers

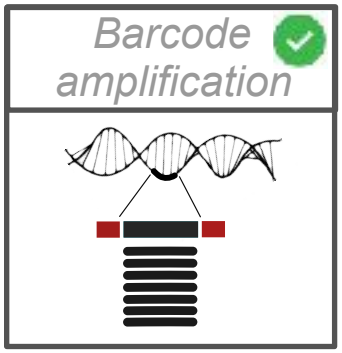
st for primer specificity

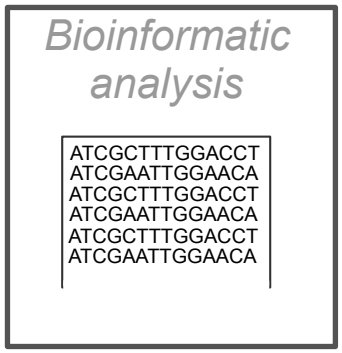
Funded by European Union

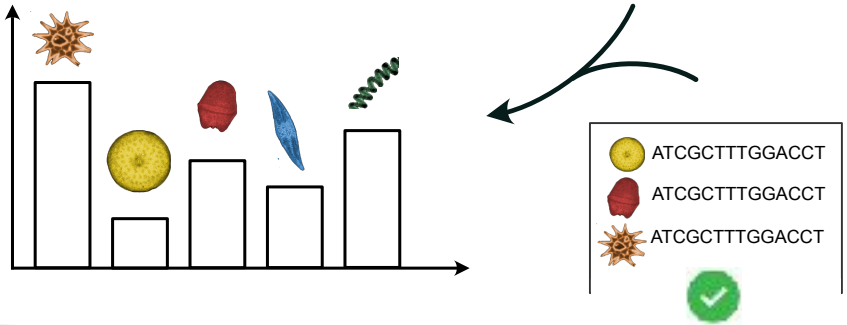
Metabarcoding steps











Funded by European Union

BIOLAWEB