Diatom metabarcoding for biomonitoring : 4th part

F. Rimet, A. Bouchez

Application of diatom biomonitoring, Source of bias, Intercalibration exercise

Schedule

1- Application of diatom biomonitoring to rivers ←

- Imitation: use of existing diatom indices
- Index based on phylogenetic inference of ecological profiles
- Taxonomy free indices
- Which bootstrap % to use for taxonomic assignation?

2- Source of bias

3- Intercalibration exercise

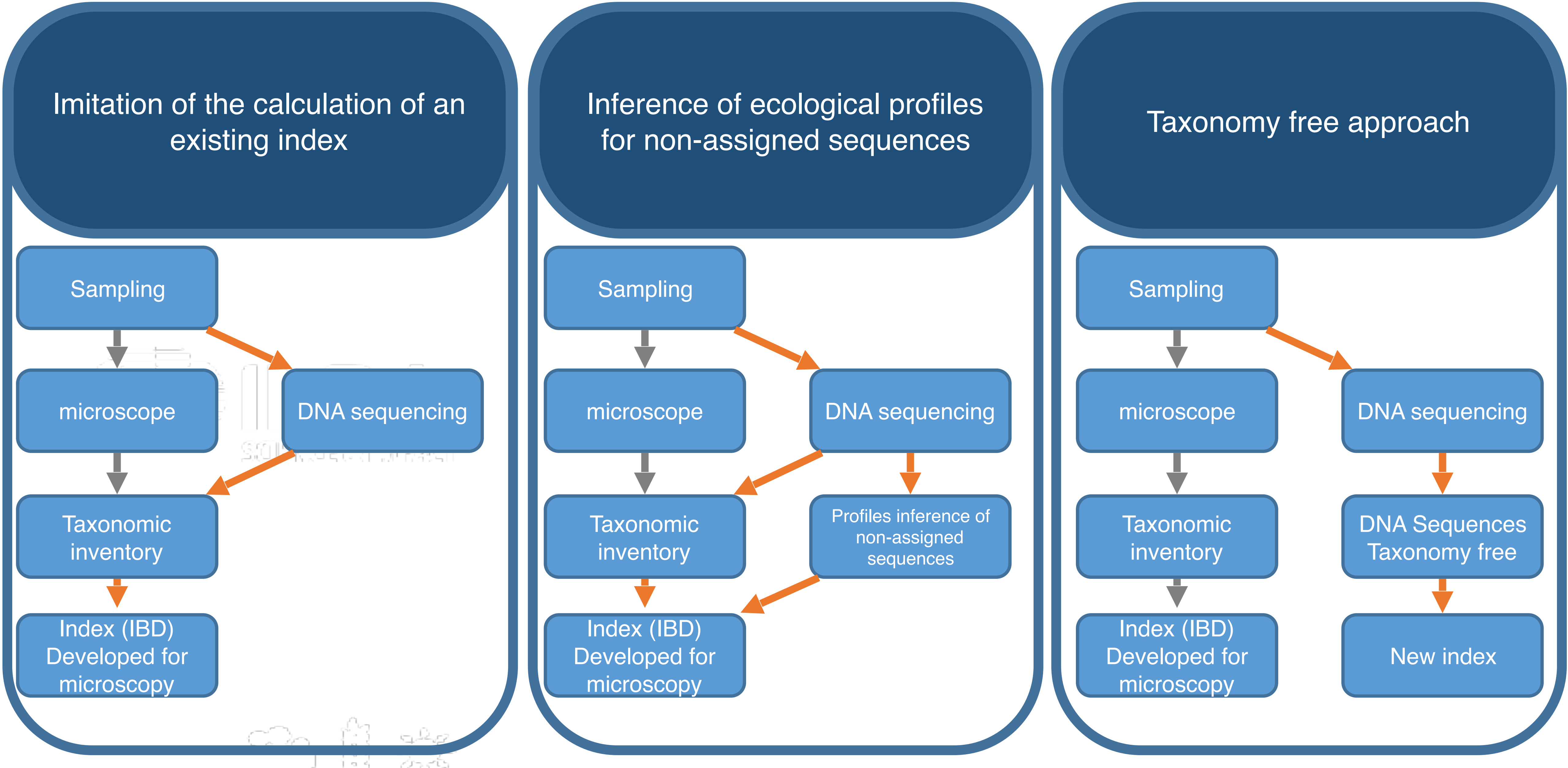
Schedule

1- Application of diatom biomonitoring to rivers

- Imitation: use of existing diatom indices ←
- Index based on phylogenetic inference of ecological profiles ←
- Taxonomy free indices ←
- Which bootstrap % to use for taxonomic assignation?

2- Source of bias

3- Intercalibration exercise



Imitation of the calculation of an existing index

Sampling

microscope

DNA sequencing

Taxonomic inventory

Index (IBD) Developed for microscopy

Inference of ecological profiles for non-assigned sequences

Sampling

microscope

DNA sequencing

Taxonomic inventory

Profiles inference of non-assigned sequences

Index (IBD) Developed for microscopy

Taxonomy free approach

Sampling

microscope

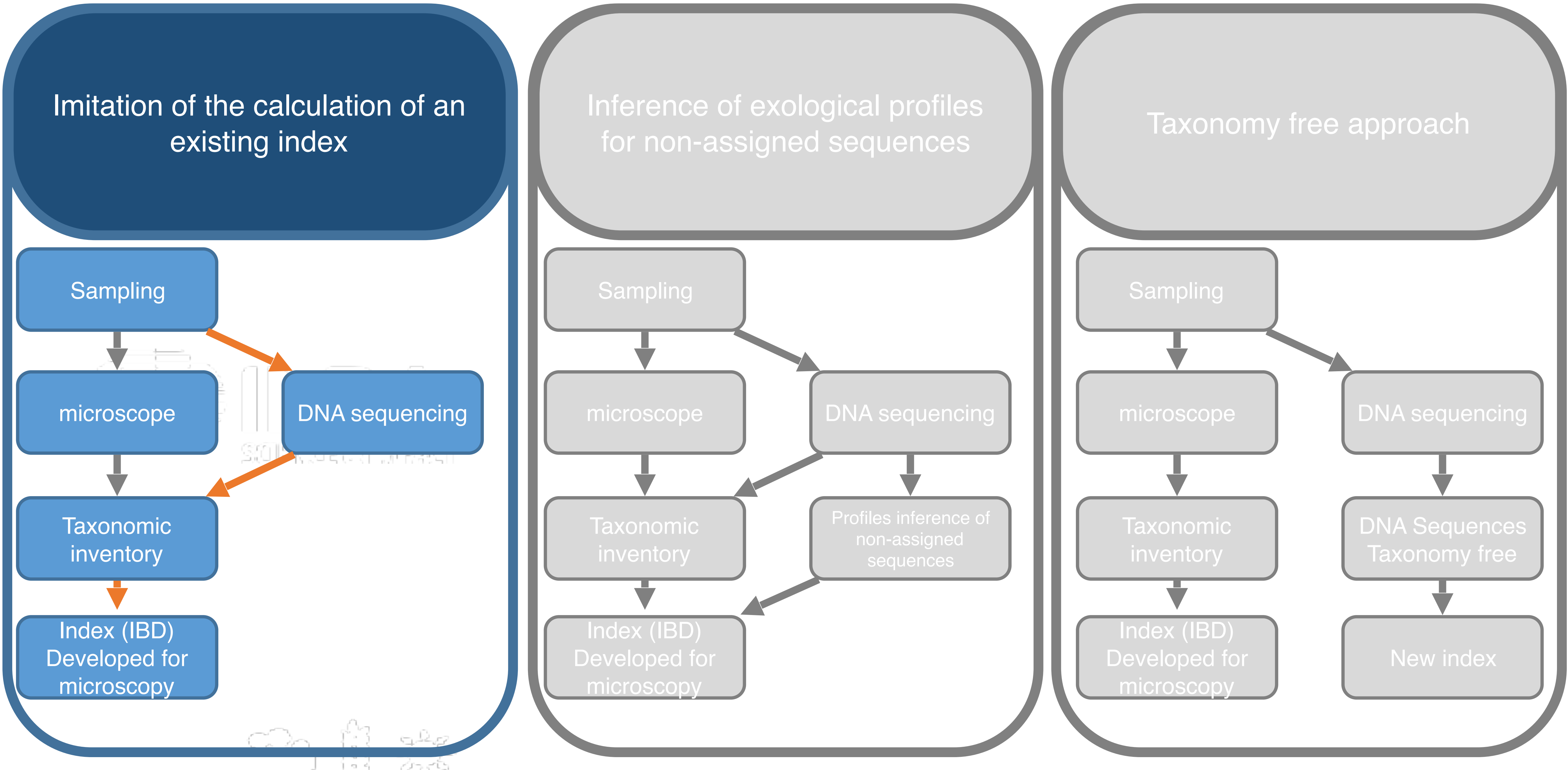
DNA sequencing

Taxonomic inventory

DNA Sequences Taxonomy free

Index (IBD) Developed for microscopy

New index



Imitation of the calculation of an existing index

Sampling

microscope

DNA sequencing

Taxonomic inventory

Index (IBD) Developed for microscopy

Inference of exological profiles for non-assigned sequences

Sampling

microscope

DNA sequencing

Taxonomic inventory

Profiles inference of non-assigned sequences

Index (IBD) Developed for microscopy

Taxonomy free approach

Sampling

microscope

DNA sequencing

Taxonomic inventory

DNA Sequences Taxonomy free

Index (IBD) Developed for microscopy

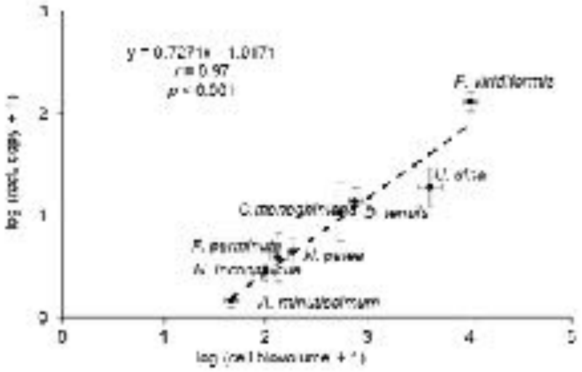
New index

Floristic list

Floristic list modified

Taxon	Séquences %	biovol (µm3)	CF v2	Sequence % / CF v2	expressed in %	frustule %
Achnanthidium minutissimum	14,0	76,0	2,2	6,4	68,3	75,0
Amphora pediculus	2,0	72,0	2,1	0,9	10,0	13,0
Navicula cryptotenella	10,0	386,0	5,6	1,8	18,9	10,0
Melosira varians	73,0	14515,0	282,9	0,3	2,7	2,0

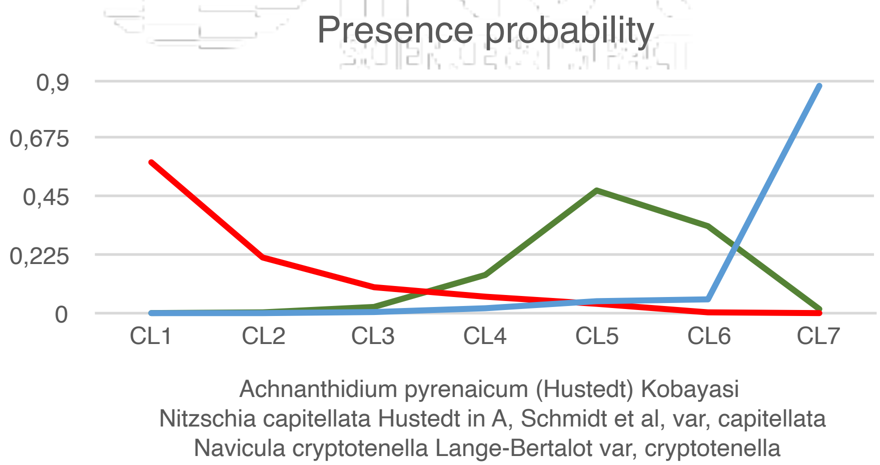
CFV2 =
10^(0,0703*((LOG(biovolume in µm3))^2,4908))



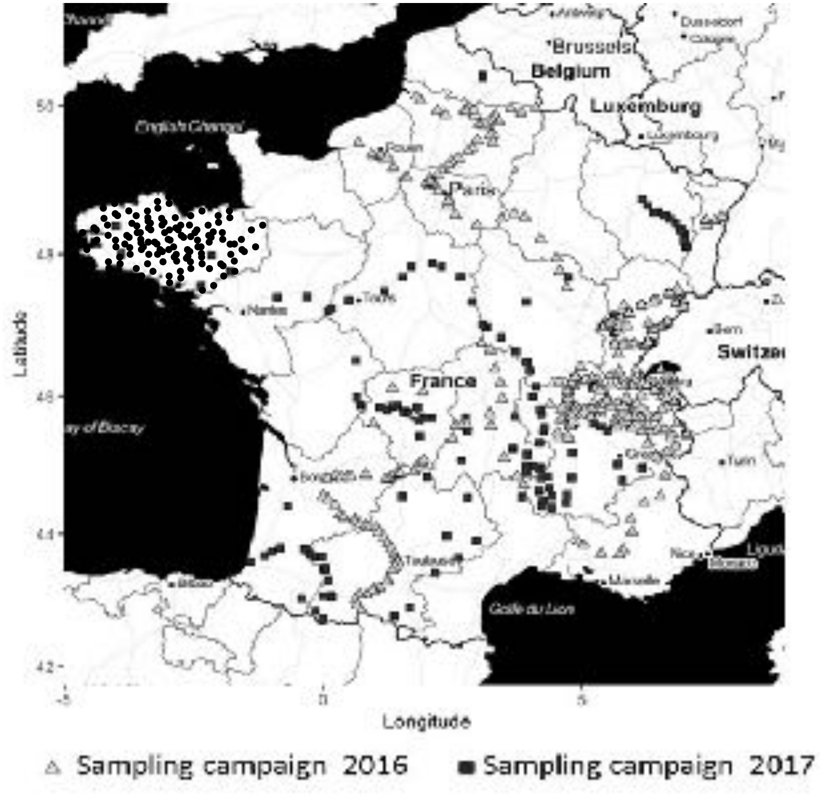
Comparison of water quality assessment with microscopy and DNA based inventories
> Application to France and the IBD

IBD : « Indice Biologique Diatomées »
French diatom index used in routine to assess river quality for the WFD

Ecological profiles are defined for 828 taxa

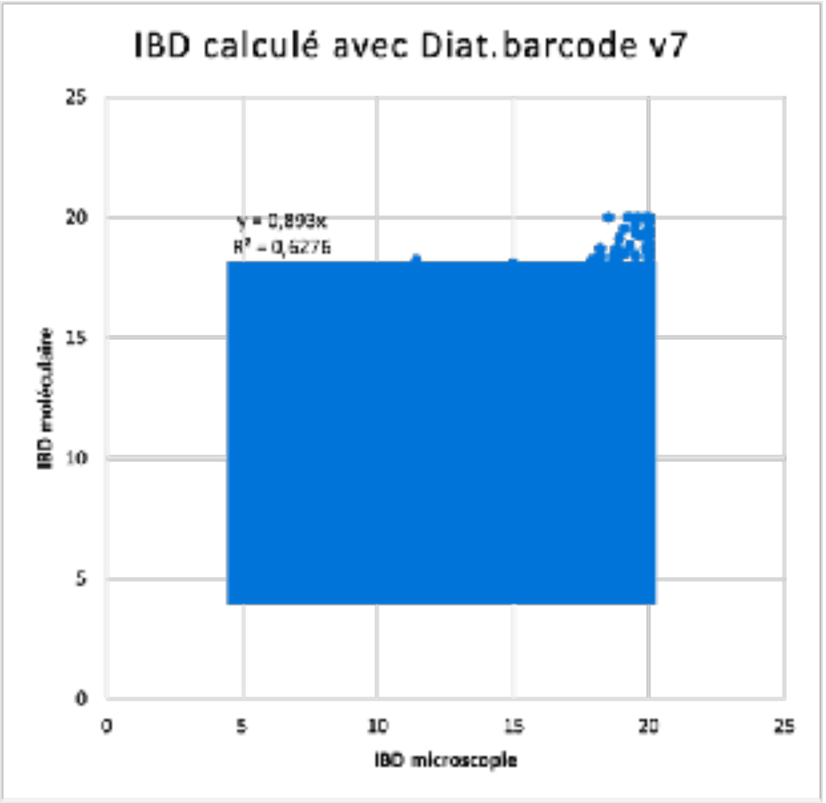


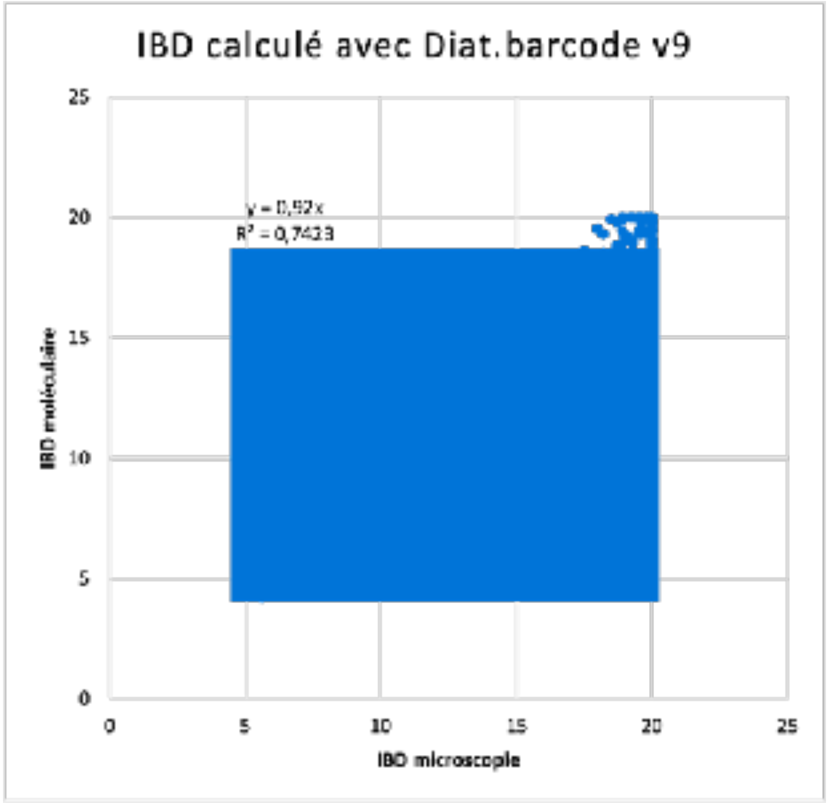
Test carried out on sampling campains 2016, 2017, 2019, in 658 sites

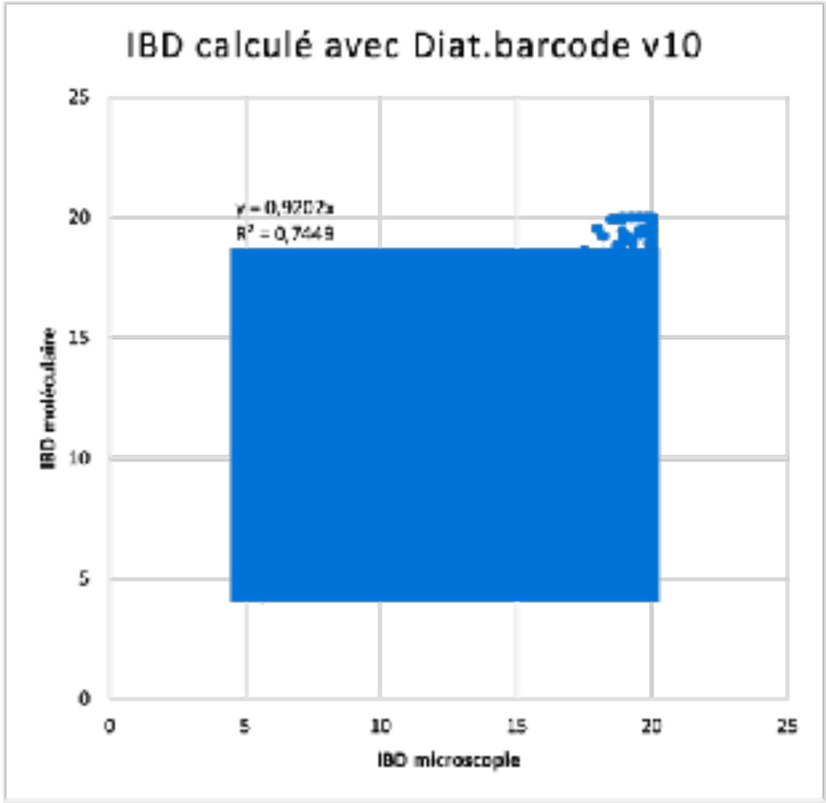


Comparison of water quality assessment with microscopy and DNA based inventories

> Application to France and the IBD


IBD calculé avec Diat.barcode v7


IBD calculé avec Diat.barcode v9


IBD calculé avec Diat.barcode v10

Improvement of the correlation thanks to Diat.barcode completion
Very good correlation in mainland France, slope is almost 1:1

What is the proportion of reads assigned to species level depending on Diat.barcode versions (v7 and v10) for 6 different regions?

Diat.barcode: a curated barcoding database

DIAT.BARCODE

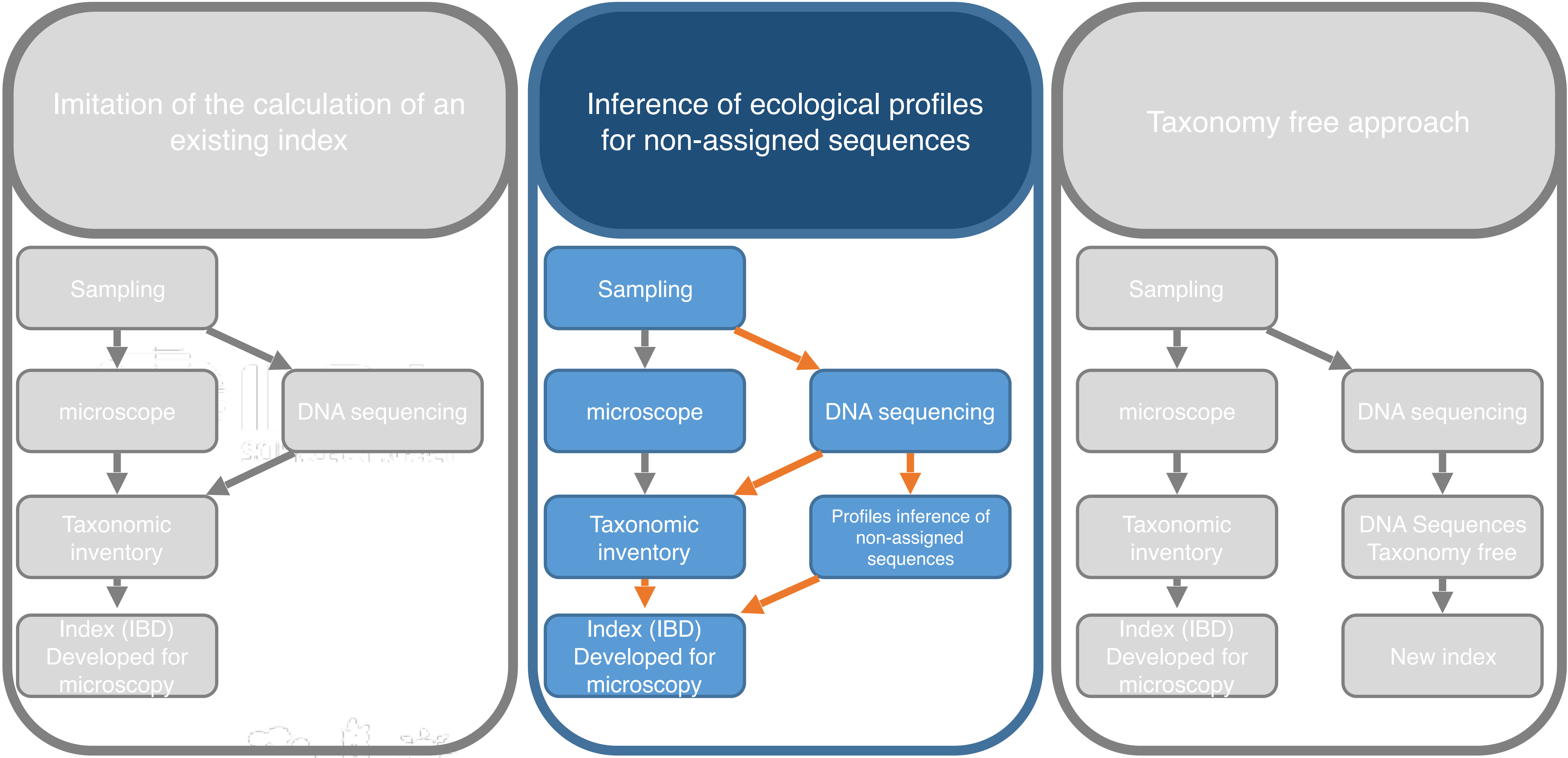
Assigned reads to species (%)	v7	v10
France mainland	83%	91%
Guyane	44%	45%
Guadeloupe	43%	84%
Martinique	44%	80%
Mayotte	47%	58%
Réunion	71%	90%

Direct calculation of IBD is adapted for mainland France

What can we do in these situations where Diat.barcode is uncomplete?

To sum up:

Method adapted when the reference library is almost complete: Microscopy and DNA are very well correlated. However, a few differences (eg. Fragile frustules, dead frustules, cryptic and species difficult to identify).



Imitation of the calculation of an existing index

Sampling

microscope

DNA sequencing

Taxonomic inventory

Index (IBD) Developed for microscopy

Inference of ecological profiles for non-assigned sequences

Sampling

microscope

DNA sequencing

Taxonomic inventory

Profiles inference of non-assigned sequences

Index (IBD) Developed for microscopy

Taxonomy free approach

Sampling

microscope

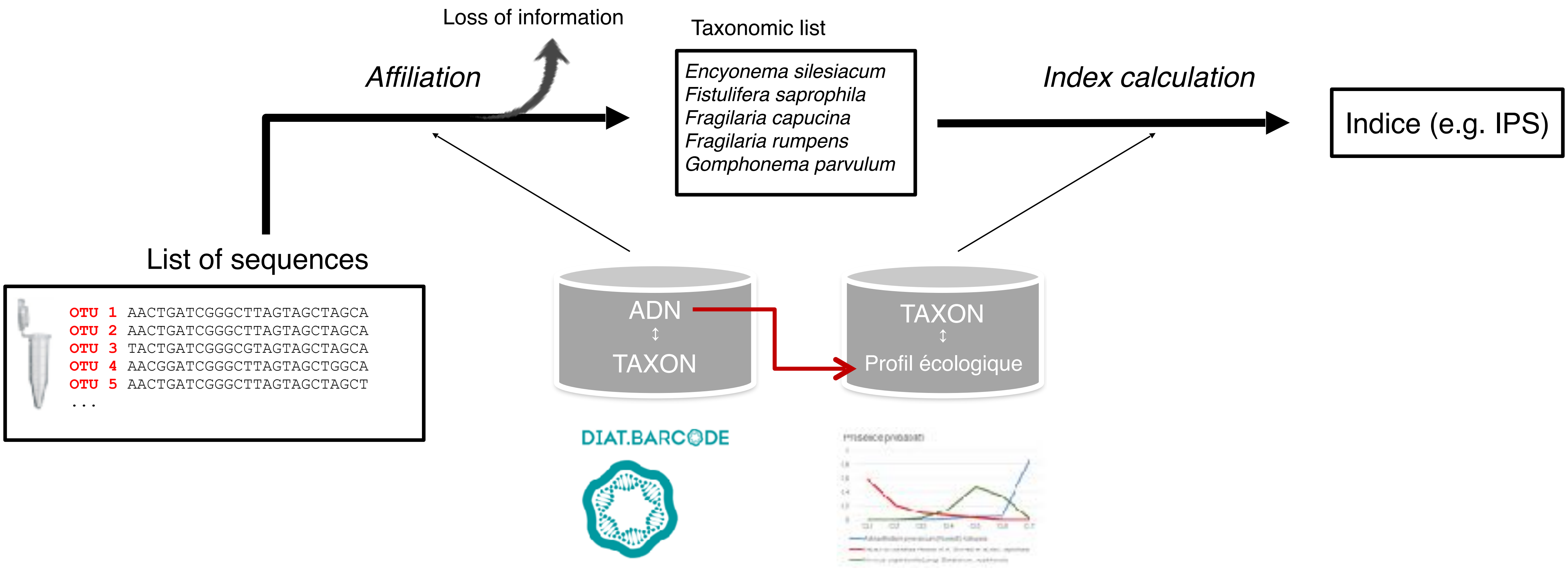
DNA sequencing

Taxonomic inventory

DNA Sequences Taxonomy free

Index (IBD) Developed for microscopy

New index



Loss of information

Affiliation

Taxonomic list

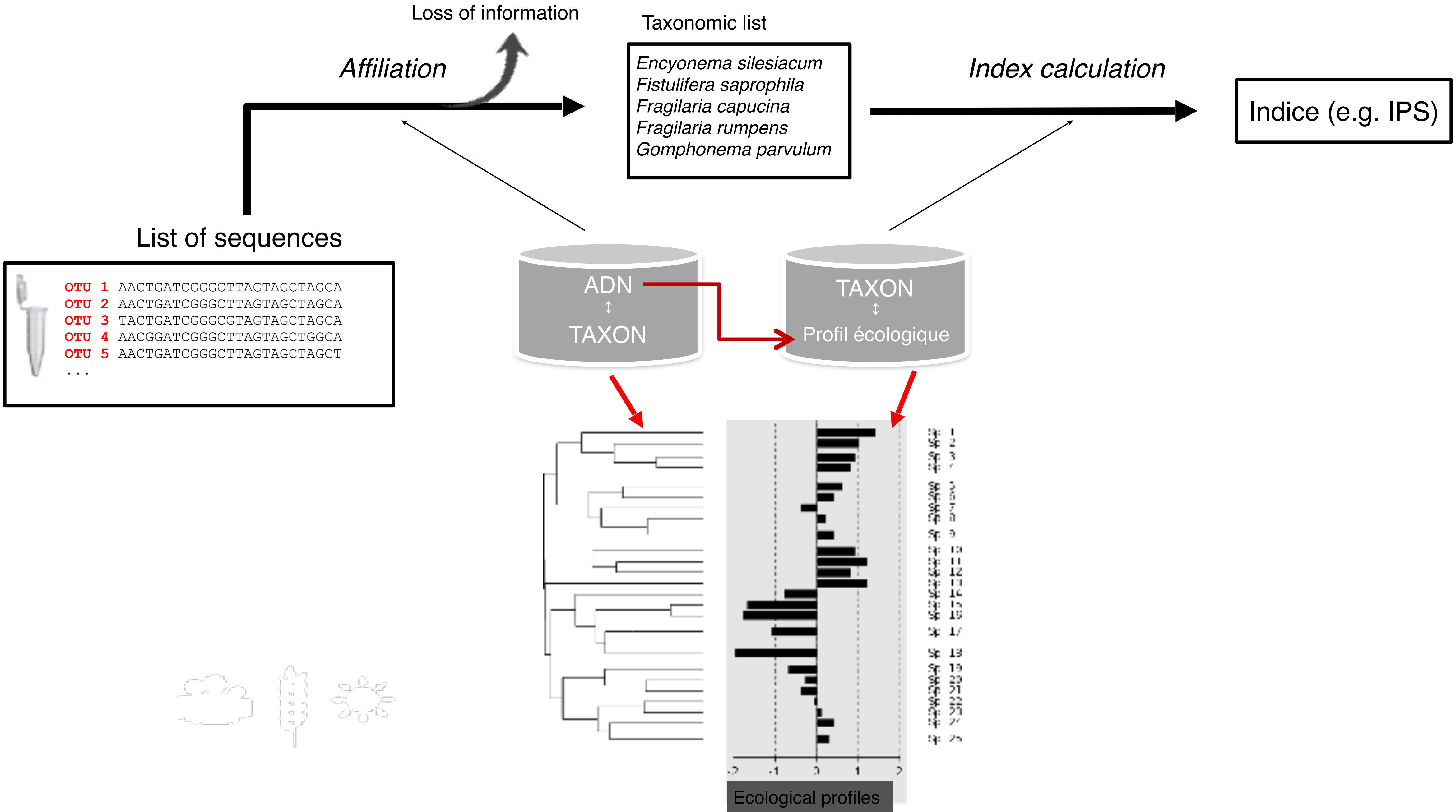
Encyonema silesiacum
Fistulifera saprophila
Fragilaria capucina
Fragilaria rumpens
Gomphonema parvulum

Index calculation

Indice (e.g. IPS)

List of sequences

```
OTU 1  AACTGATCGGGCTTAGTAGCTAGCA
OTU 2  AACTGATCGGGCTTAGTAGCTAGCA
OTU 3  TACTGATCGGGCGTAGTAGCTAGCA
OTU 4  AACGGATCGGGCTTAGTAGCTGGCA
OTU 5  AACTGATCGGGCTTAGTAGCTAGCT
...
```



Loss of information

Affiliation

Taxonomic list

Encyonema silesiacum
Fistulifera saprophila
Fragilaria capucina
Fragilaria rumpens
Gomphonema parvulum

Index calculation

Indice (e.g. IPS)

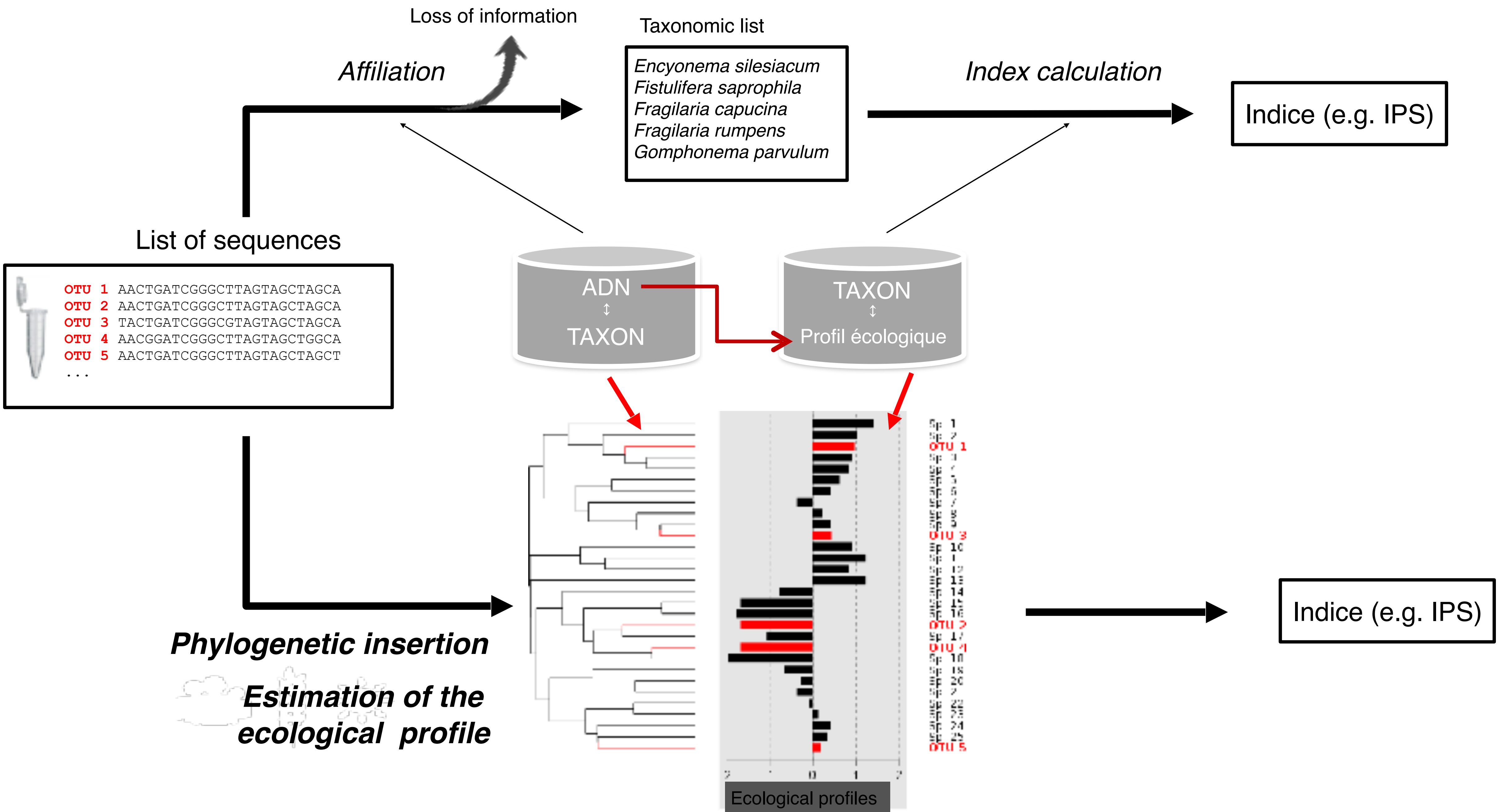
List of sequences

OTU 1 AACTGATCGGGCTTAGTAGCTAGCA
OTU 2 AACTGATCGGGCTTAGTAGCTAGCA
OTU 3 TACTGATCGGGCGTAGTAGCTAGCA
OTU 4 AACGGATCGGGCTTAGTAGCTGGCA
OTU 5 AACTGATCGGGCTTAGTAGCTAGCT
...

ADN
↕
TAXON

TAXON
↕
Profil écologique

Ecological profiles



Loss of information

Affiliation

Taxonomic list

Encyonema silesiacum
Fistulifera saprophila
Fragilaria capucina
Fragilaria rumpens
Gomphonema parvulum

Index calculation

Indice (e.g. IPS)

List of sequences

```
OTU 1 AACTGATCGGGCTTAGTAGCTAGCA
OTU 2 AACTGATCGGGCTTAGTAGCTAGCA
OTU 3 TACTGATCGGGCGTAGTAGCTAGCA
OTU 4 AACGGATCGGGCTTAGTAGCTGGCA
OTU 5 AACTGATCGGGCTTAGTAGCTAGCT
...
```

ADN ↕ TAXON

TAXON ↕ Profil écologique

Phylogenetic insertion

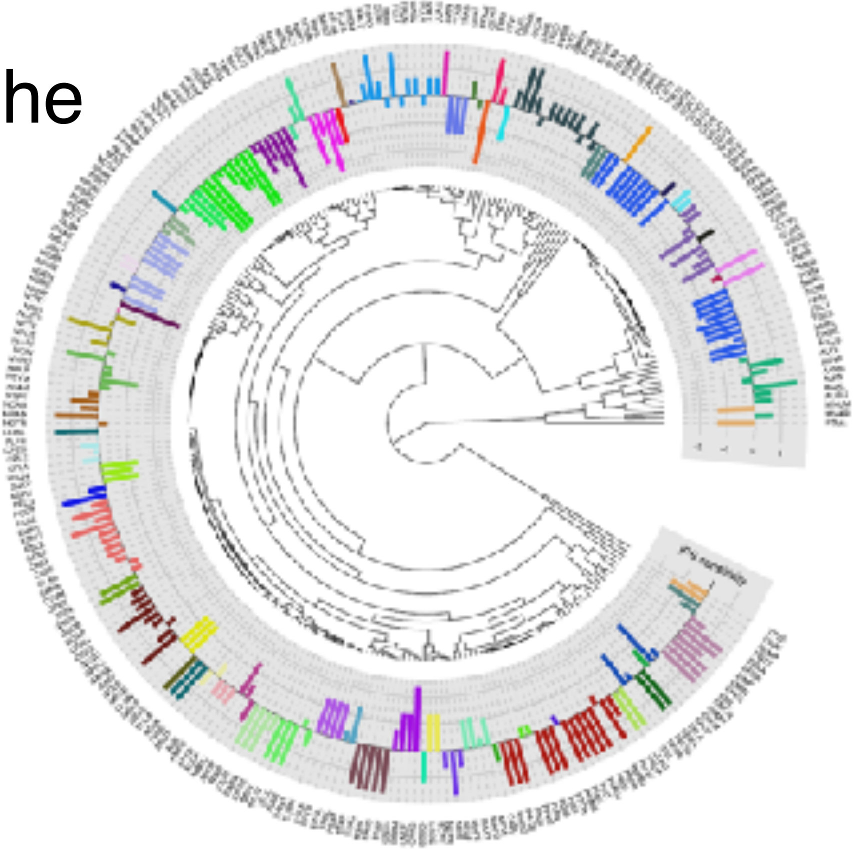
Estimation of the ecological profile

Ecological profiles

Indice (e.g. IPS)

There is a phylogenetic signal of the ecological profile of diatoms

(Keck et al. 2016 J. Appl Ecol)

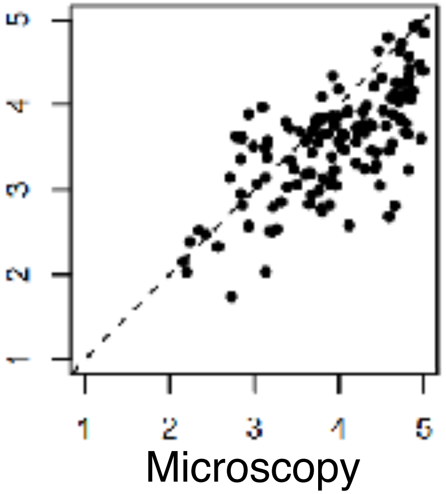


In situ test (139 sites, mainland France)

(Keck et al. 2018, Mol Ecol Res)

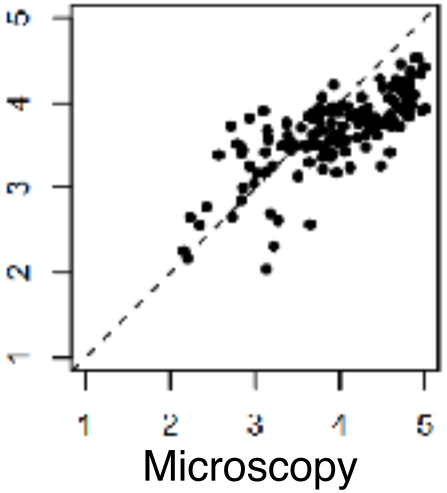
Comparison of performances (IPS)

DNA “classical”
with 30% of OTUs
classified



r: +7%
MSE: -29%

✓



DNA “classical”
30% of classified OTUs
+
70% OTUs non-classified
but with a profile inferred
from the phylogeny

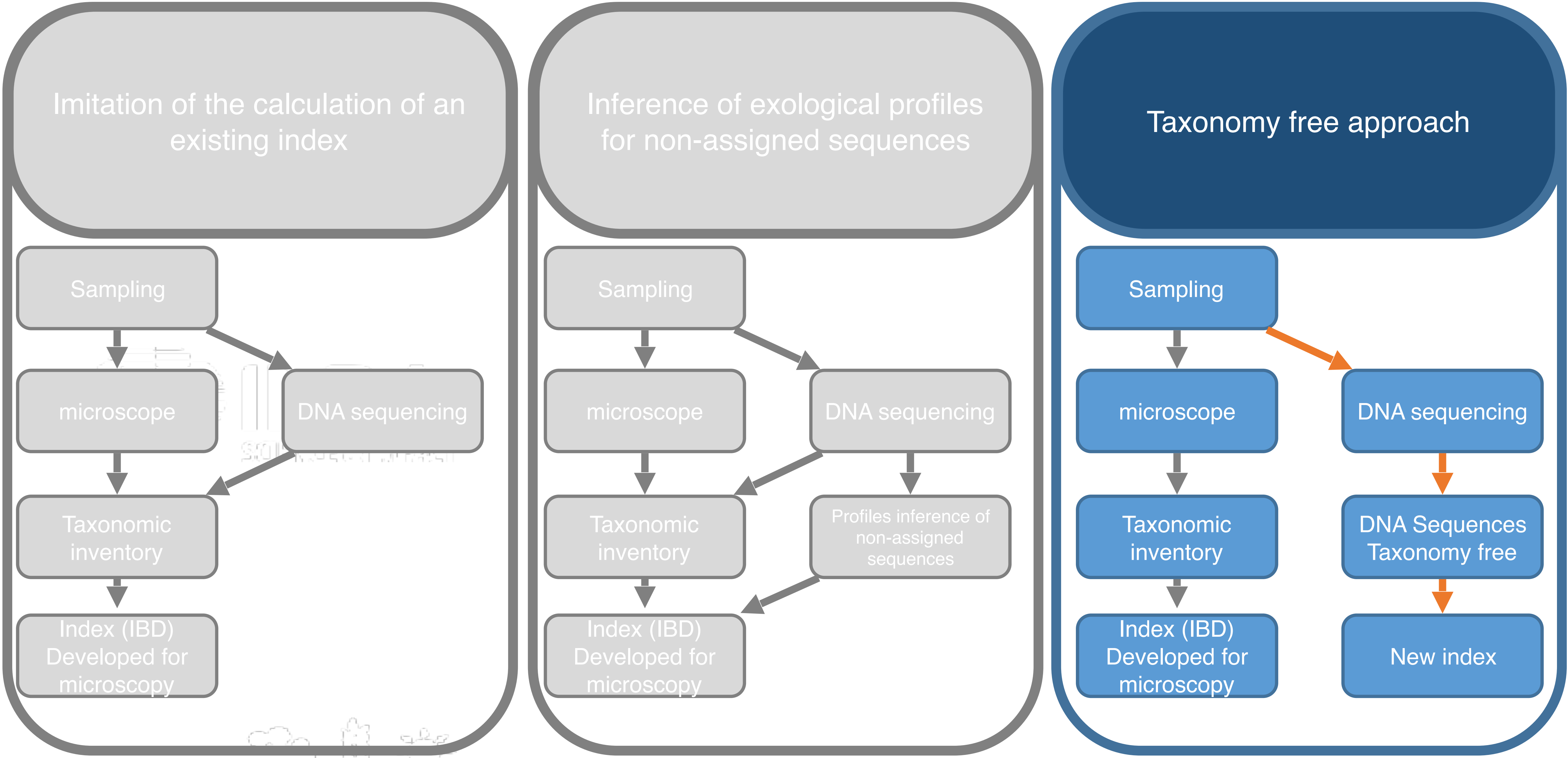
To sum up

The method:

- takes advantage of 100% of the DNA data
- improves the estimation of diatom indices compared to the "classical" approach
- is based on a solid theoretical foundation
(The phylogenetic signal is a direct consequence of the Darwinian principle of descent with modification)
- is built on existing databases/indices

We can consider this method as a transition to de-novo molecular indices such as:

- Taxonomy-free indices
- Supervised machine learning indices



Imitation of the calculation of an existing index

Sampling

microscope

DNA sequencing

Taxonomic inventory

Index (IBD) Developed for microscopy

Inference of exological profiles for non-assigned sequences

Sampling

microscope

DNA sequencing

Taxonomic inventory

Profiles inference of non-assigned sequences

Index (IBD) Developed for microscopy

Taxonomy free approach

Sampling

microscope

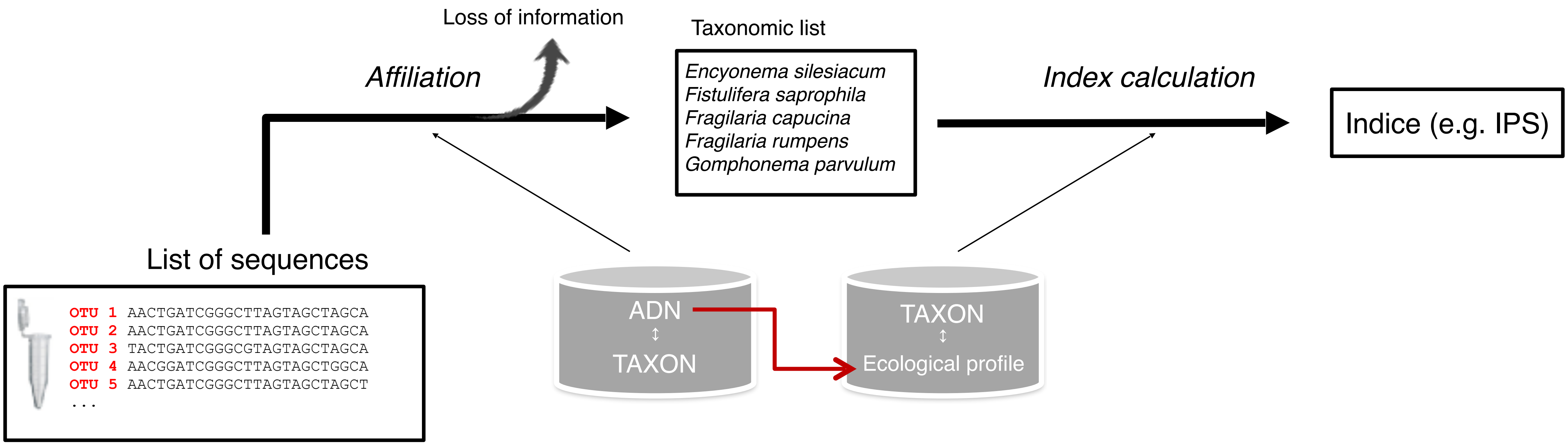
DNA sequencing

Taxonomic inventory

DNA Sequences Taxonomy free

Index (IBD) Developed for microscopy

New index



Loss of information

Affiliation

Taxonomic list

Encyonema silesiacum
Fistulifera saprophila
Fragilaria capucina
Fragilaria rumpens
Gomphonema parvulum

Index calculation

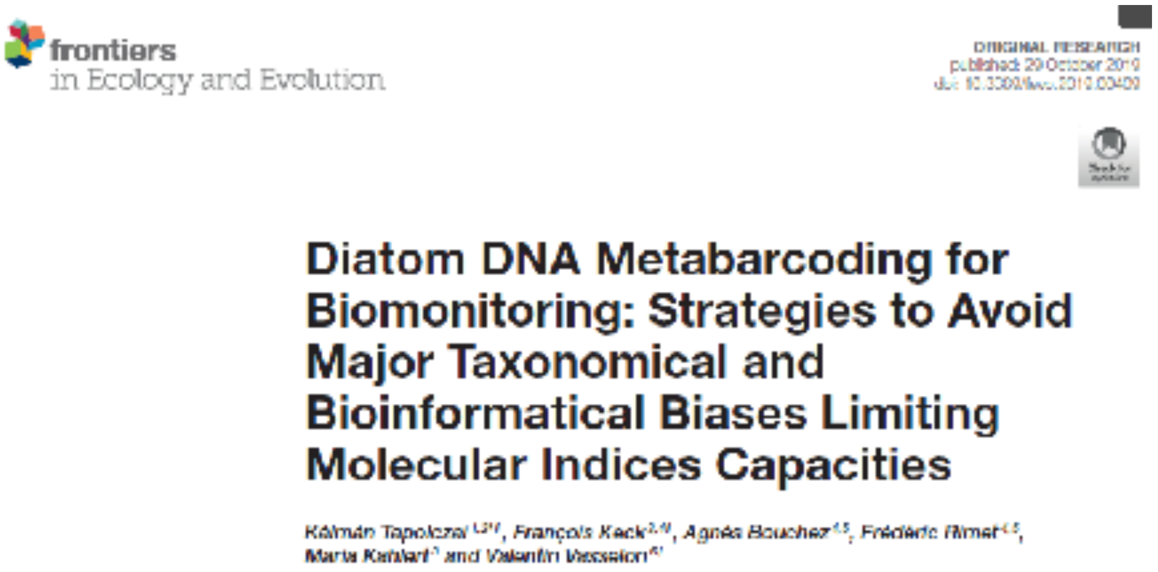
Indice (e.g. IPS)

List of sequences

```
OTU 1 AACTGATCGGGCTTAGTAGCTAGCA
OTU 2 AACTGATCGGGCTTAGTAGCTAGCA
OTU 3 TACTGATCGGGCGTAGTAGCTAGCA
OTU 4 AACGGATCGGGCTTAGTAGCTGGCA
OTU 5 AACTGATCGGGCTTAGTAGCTAGCT
...
```

ADN
↕
TAXON

TAXON
↕
Ecological profile

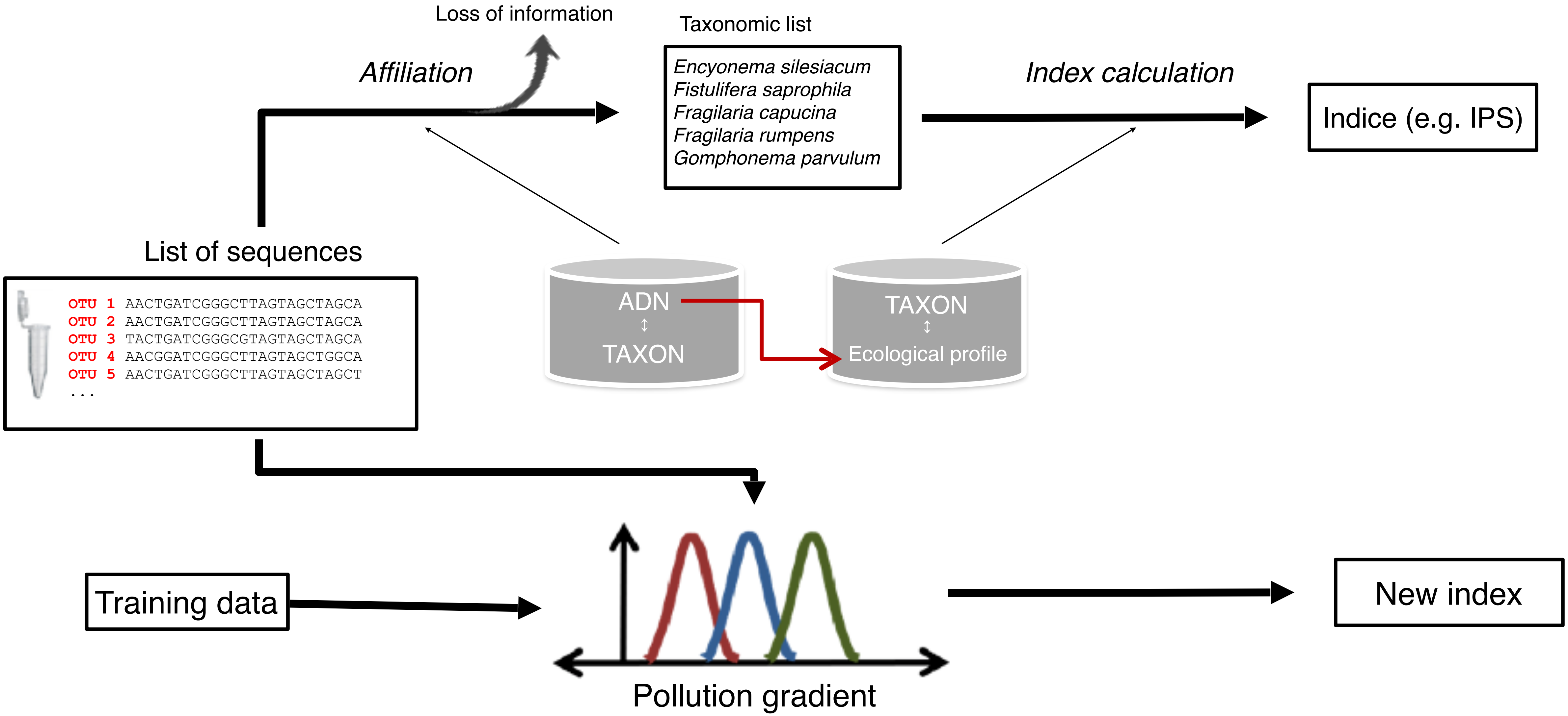


frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 29 October 2019
doi: 10.3389/fevo.2019.00409

Diatom DNA Metabarcoding for Biomonitoring: Strategies to Avoid Major Taxonomical and Bioinformatical Biases Limiting Molecular Indices Capacities

Kálmán Tapolczai[1,2,3*], François Keck[2,4], Agnès Bouchez[4,5], Frédéric Rimet[4,5],
Maria Kahlert[3] and Valentin Vasselon[6*]



Loss of information

Affiliation

Taxonomic list

Encyonema silesiacum
Fistulifera saprophila
Fragilaria capucina
Fragilaria rumpens
Gomphonema parvulum

Index calculation

Indice (e.g. IPS)

List of sequences

```
OTU 1 AACTGATCGGGCTTAGTAGCTAGCA
OTU 2 AACTGATCGGGCTTAGTAGCTAGCA
OTU 3 TACTGATCGGGCGTAGTAGCTAGCA
OTU 4 AACGGATCGGGCTTAGTAGCTGGCA
OTU 5 AACTGATCGGGCTTAGTAGCTAGCT
...
```

ADN ↕ TAXON

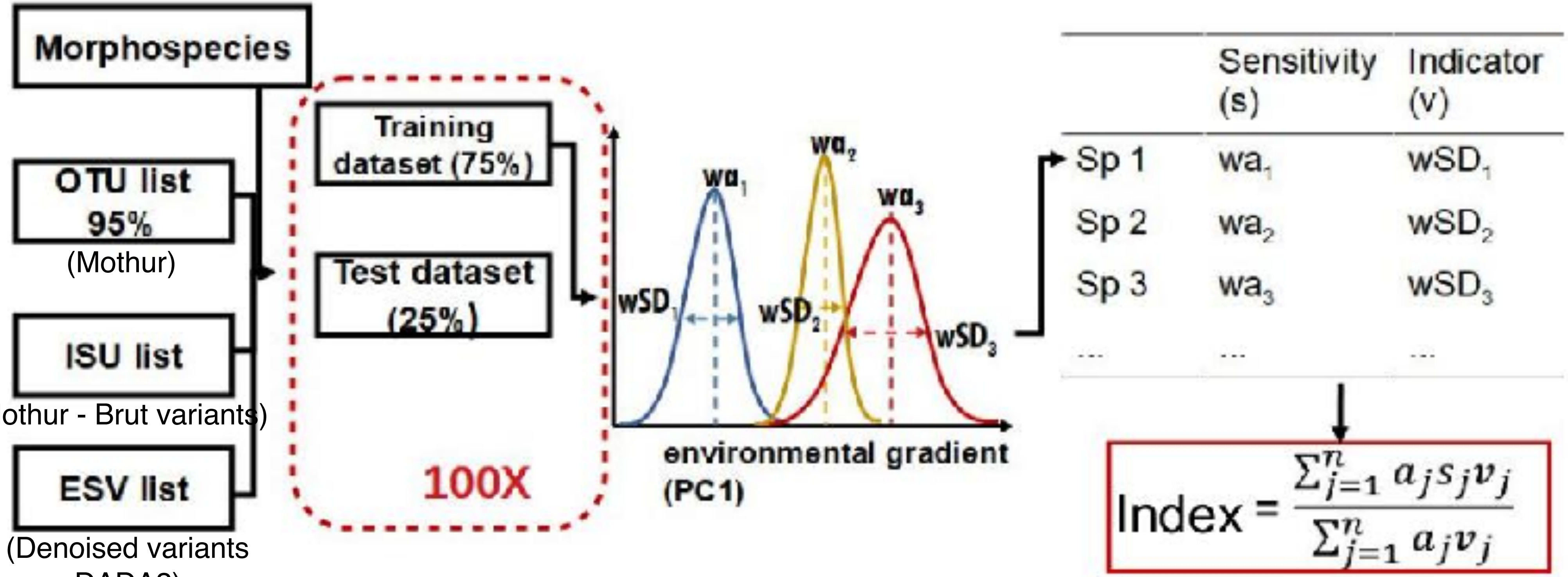
TAXON ↕ Ecological profile

Training data

Pollution gradient

New index

Strategy for the development and test of indices:



(Mothur)

(Mothur - Brut variants)

(Denoised variants DADA2)

	Sensitivity (s)	Indicator (v)
Sp 1	wa_1	wSD_1
Sp 2	wa_2	wSD_2
Sp 3	wa_3	wSD_3
...	...	...

$$\text{Index} = \frac{\sum_{j=1}^{n} a_j s_j v_j}{\sum_{j=1}^{n} a_j v_j}$$

ISU: to keep good quality DNA reads using the trim.seqs() command and the following parameters: a sequence length of 263 ± 10 bp (rbcL barcode length without primers), a Phred quality score ≥23 over a moving window of 25 bp, 0 ambiguities (“N”), a maximum homopolymer length of 8 bp

OTU: (i) ISUs were aligned using the align.seqs() command and poorly aligned reads were removed using the command screen.seqs(start=28, optimize=end, criteria=90); (ii) we used the pre.cluster() command to denoise sequencing errors by preclustering rare ISUs with related more abundant ones (1 bp threshold); (iii) detection of chimeras was performed using the chimera.vsearch() command;

Raw sequences

Length / quality filter

Dereplication into ISU

Chimera removal

Taxonomic assignation of ISU
(ex: 60, 70, 80%)

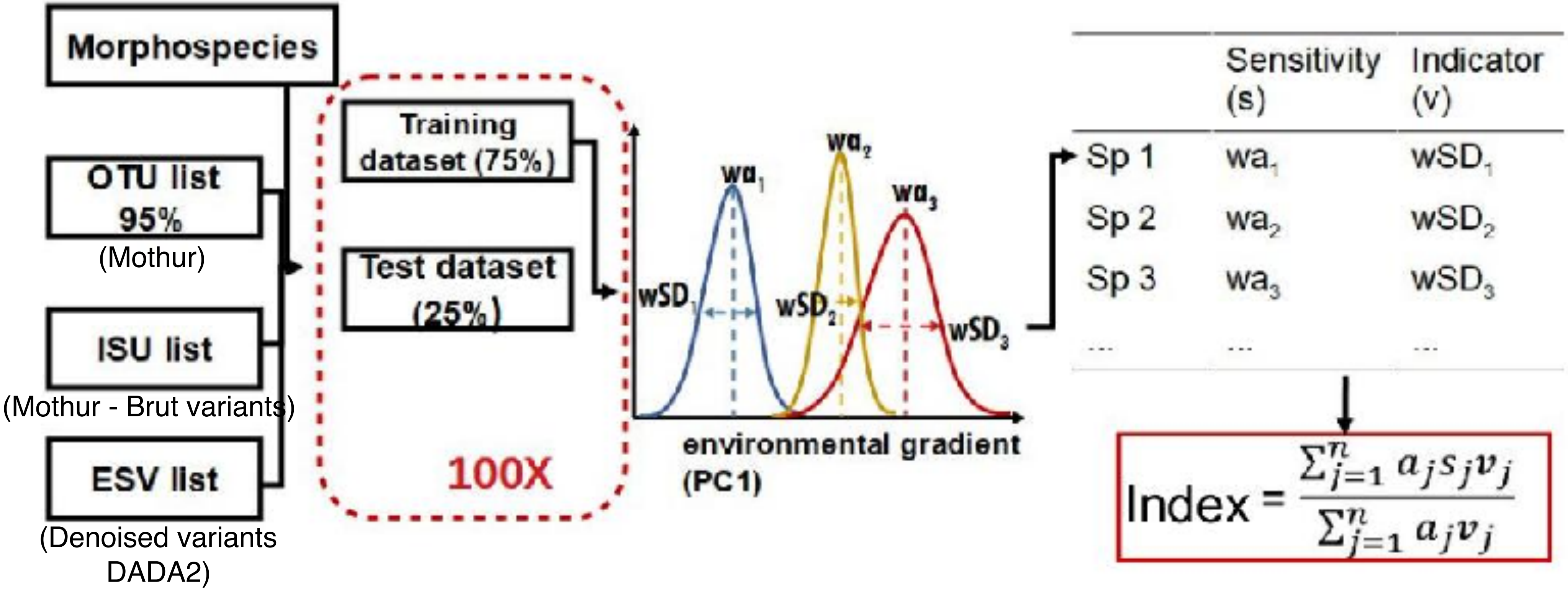
Clustering into OTU
(ex: Furthest neighbor, Opticlust)
(ex: 95%)

Taxonomic assignation of OTU
(ex: threshold 60, 70, 80%)

Molecular inventory

MOTHUR

Strategy for the development and test of indices:



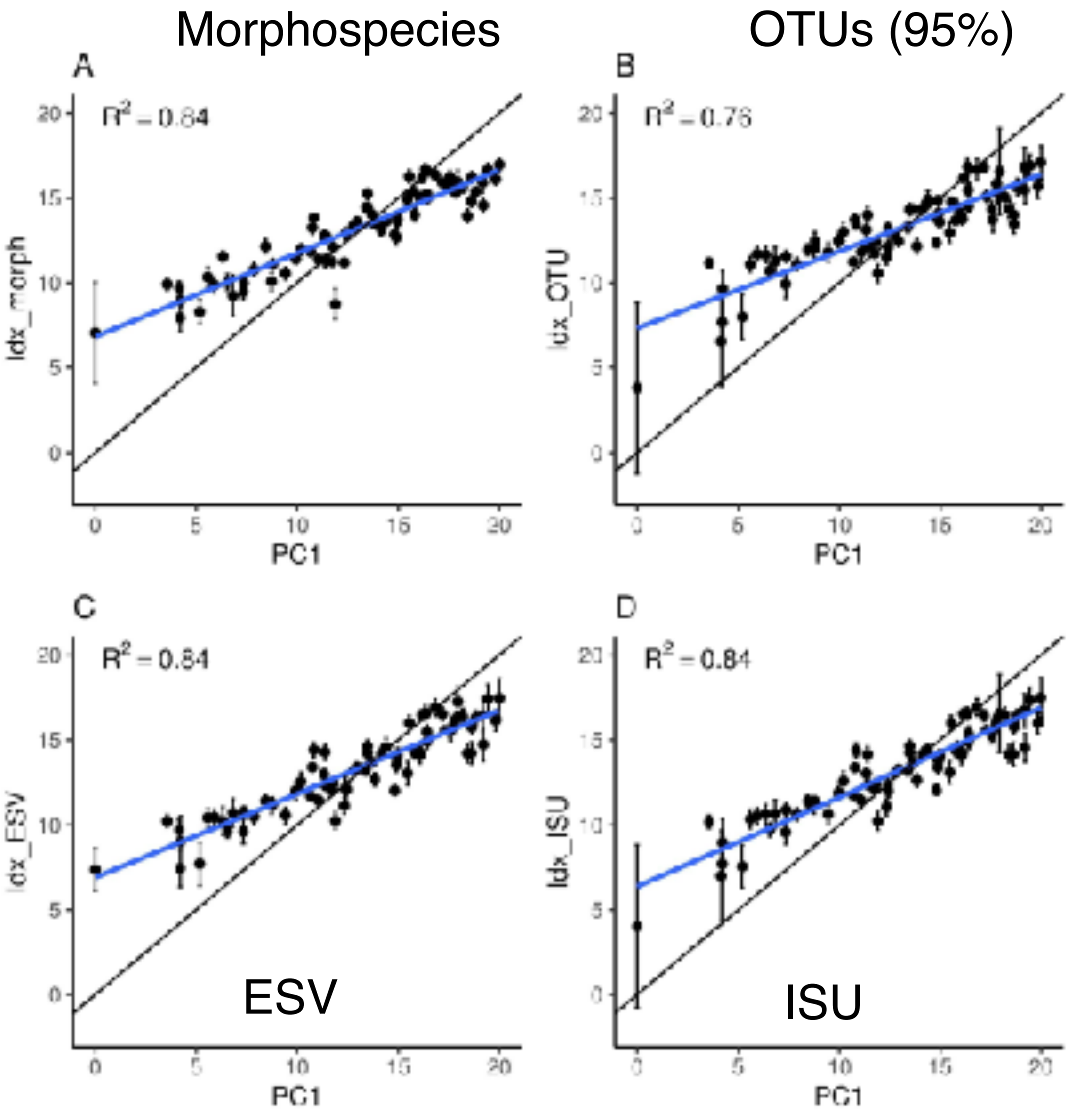
	Sensitivity (s)	Indicator (v)
Sp 1	wa_1	wSD_1
Sp 2	wa_2	wSD_2
Sp 3	wa_3	wSD_3
...	...	...

$$\text{Index} = \frac{\sum_{j=1}^{n} a_j s_j v_j}{\sum_{j=1}^{n} a_j v_j}$$

ISU: to keep good quality DNA reads using the trim.seqs() command and the following parameters: a sequence length of 263 ± 10 bp (rbcL barcode length without primers), a Phred quality score ≥23 over a moving window of 25 bp, 0 ambiguities (“N”), a maximum homopolymer length of 8 bp

OTU: (i) ISUs were aligned using the align.seqs() command and poorly aligned reads were removed using the command screen.seqs(start=28, optimize=end, criteria=90); (ii) we used the pre.cluster() command to denoise sequencing errors by preclustering rare ISUs with related more abundant ones (1 bp threshold); (iii) detection of chimeras was performed using the chimera.vsearch() command;

Results



- The 4 indices are working quite well
- Morpho, ISU, ESV indices are very similar
- The OTU index is the least good

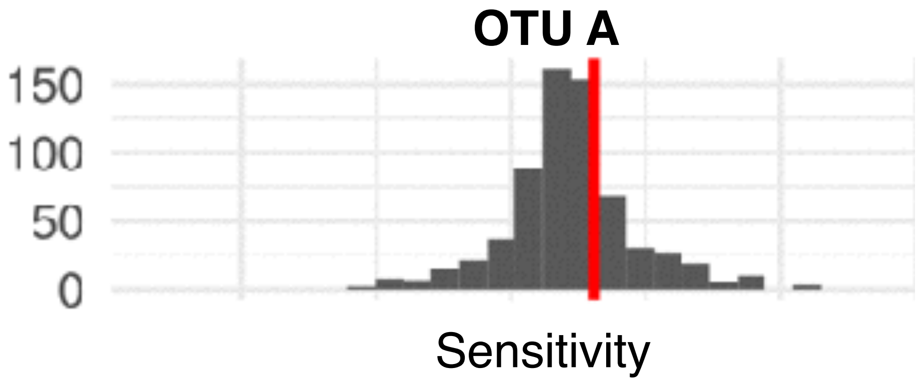
Why does the ISU (non-denoised) index work?

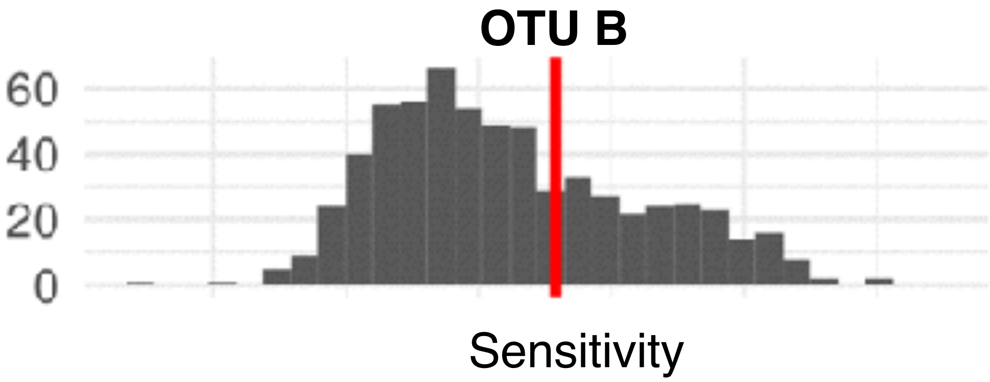
- Errors probably removed with the quality filtering
- Errors have probably an ecological profile similar to their original sequence

This questions the need for denoising for bioindication purpose

Why does the OTU index work less well?

OTUs group sequences based on genetic similarity and not on ecology





Optimum OTU = optimum of variants

Optimum OTU ≠ optimum of variants

To sum up

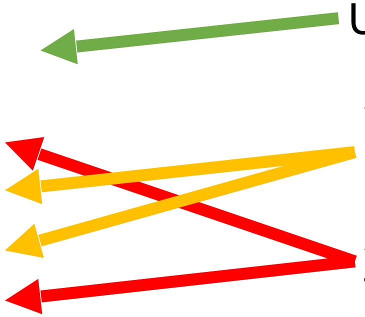
- Taxonomy-free approaches work well (provided you have a suitable dataset for training)
- OTU clustering can hide important ecological information
- Doing without clustering or even denoising (i.e. keeping bioinfo to a minimum) can facilitate the development and standardization of bioindication methods

To conclude

Diat.barcode: a curated barcoding database

DIAT.BARCODE

Assigned reads to species (%)	v7	v10
France métropolitaine	83%	91%
Guyane	44%	45%
Guadeloupe	43%	84%
Martinique	44%	80%
Mayotte	47%	58%
Réunion	71%	90%



Use of the classical IBD

Adapted to phylogenetic assignation of ecological profiles

Adapted to taxonomy free when taxonomy is poorly known

Schedule

1- Application of diatom biomonitoring to rivers
  - Imitation: use of existing diatom indices
  - Index based on phylogenetic inference of ecological profiles
  - Taxonomy free indices
  - Which bootstrap % to use for taxonomic assignation? ←

2- Source of bias

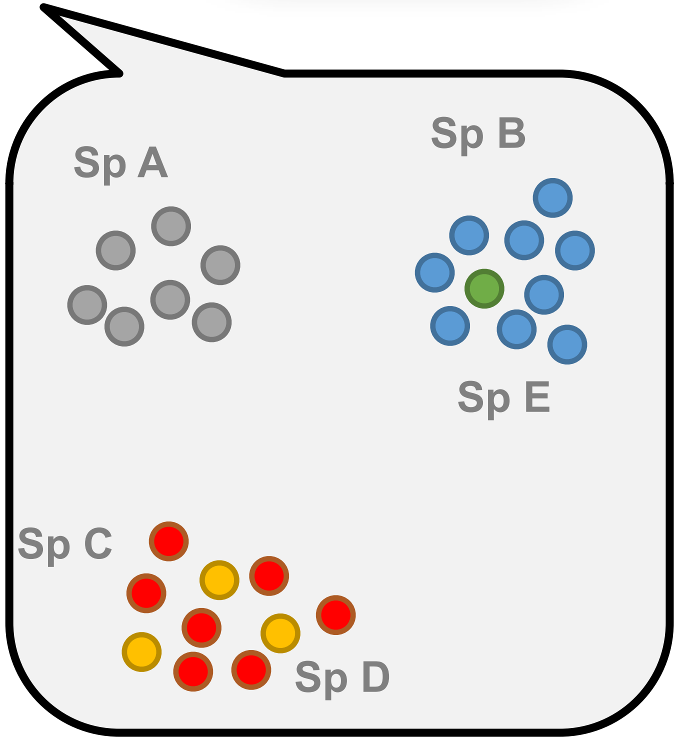
3- Intercalibration exercise

What is the bootstrap value for taxonomic assignation?

Bootstrap 95%

Seq1: SpA
Seq2: ?
Seq3: ?

Reference database

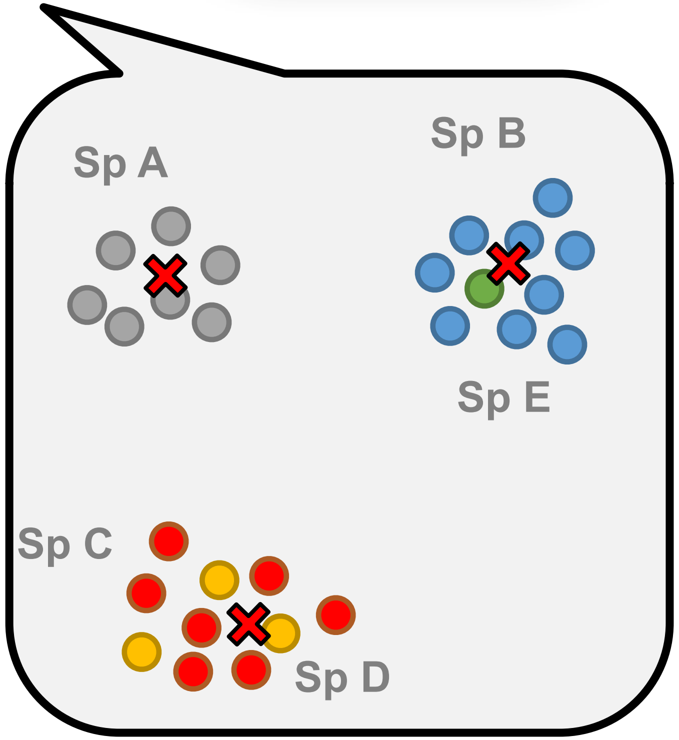


What is the bootstrap value for taxonomic assignation?

Bootstrap 95%

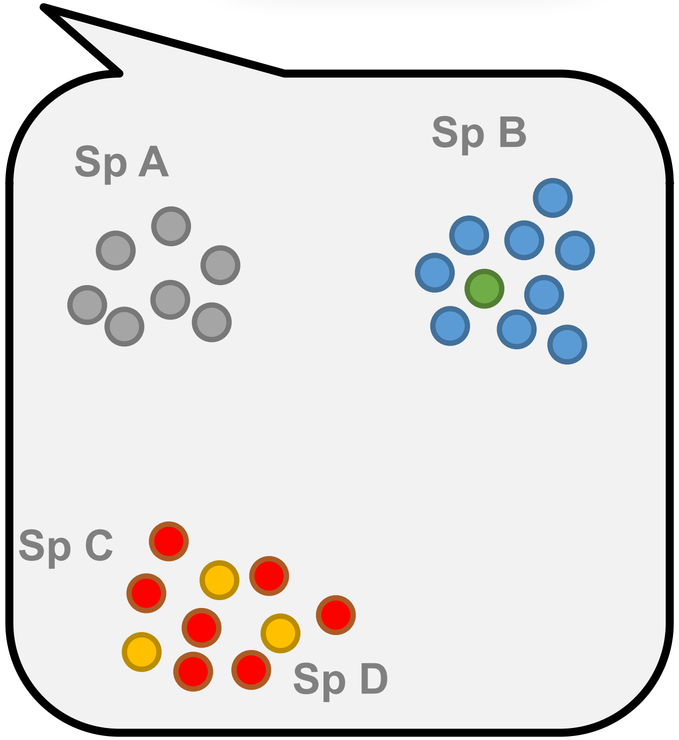
Seq1: SpA
Seq2: ?
Seq3: ?

Reference database



Bootstrap 60%

Seq1: SpA
Seq2: SpB
Seq3: SpC



Ecological Indicators 109 (2020) 105775

Contents lists available at ScienceDirect

Ecological Indicators

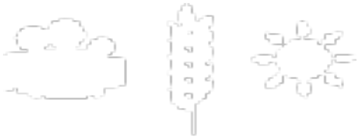
journal homepage: www.elsevier.com/locate/ecolind

Original Articles

Diatom metabarcoding applied to large scale monitoring networks: Optimization of bioinformatics strategies using Mothur software

Sinziana F. Rivera[a,*], Valentin Vasselon[b], Agnès Bouchez[a], Frédéric Rimet[a]

[a] Université Savoie Mont-Blanc, INRA, UMR CARRTEL, 75bis av. de Corzent – CS 50511, FR-74200 Thonon-les-Bains, France
[b] AFB, Pôle R&D «ECLA», INRA, UMR CARRTEL, 75bis av. de Corzent – CS 50511, FR-74200 Thonon-les-Bains, France



Raw sequences

Length / quality filter

Dereplication into ISU

Chimera removal

Taxonomic assignation of ISU
(ex: 60, 70, 80%)

Clustering into OTU
(ex: Furthest neighbor, Opticlust)
(ex: 95%)

Taxonomic assignation of OTU
(ex: threshold 60, 70, 80%)

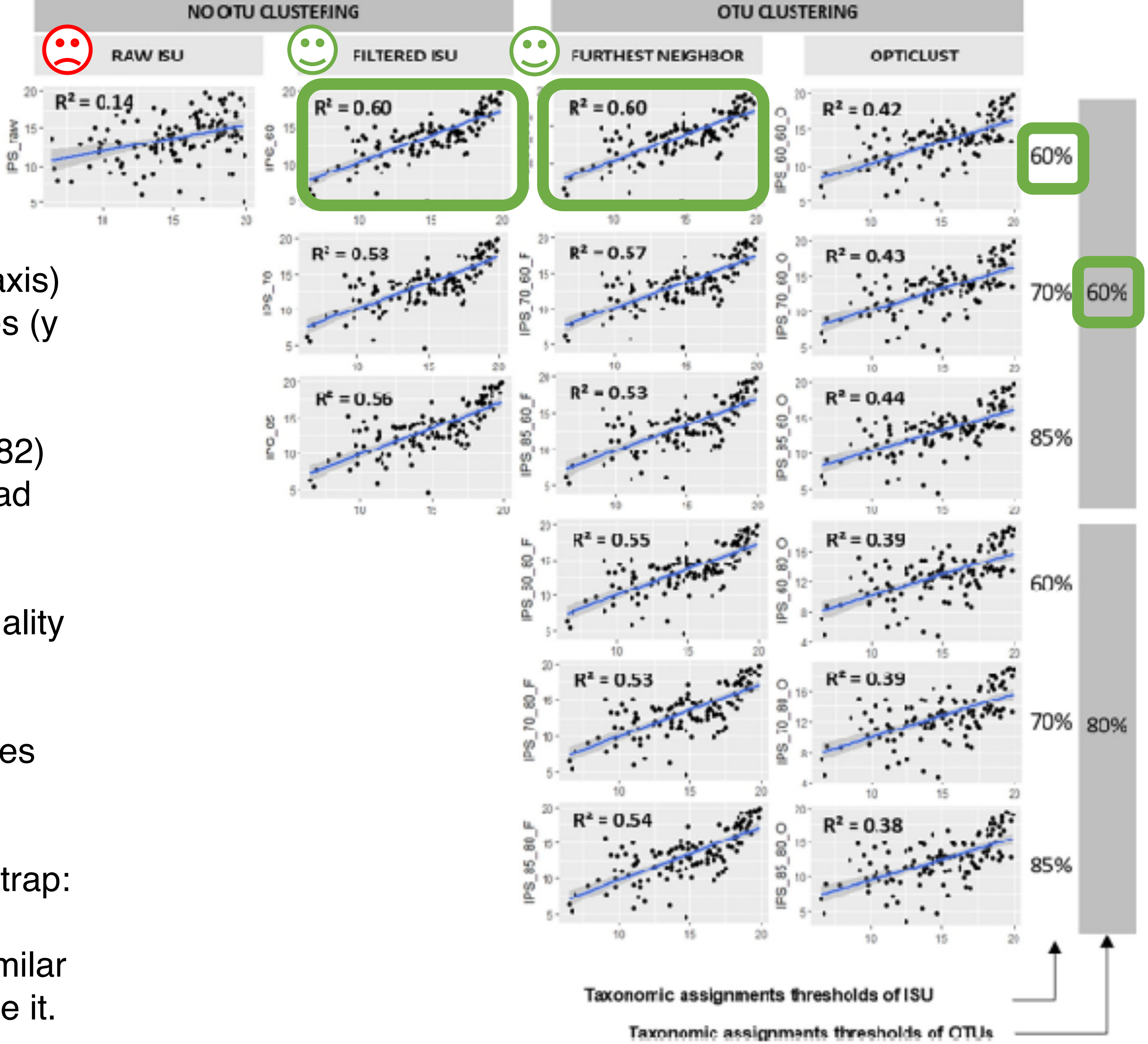
Molecular inventory

MOTHUR

Raw sequences	Raw sequences	Raw sequences
	Length / quality filter	Length / quality filter
Dereplication into ISU	Dereplication into ISU	Dereplication into ISU
	Chimera removal	Chimera removal
Taxonomic assignation of ISU (ex: 60, 70, 80%)	Taxonomic assignation of ISU (ex: 60, 70, 80%)	Taxonomic assignation of ISU (ex: 60, 70, 80%)
		Clustering into OTU (ex: Furthest neighbor, Opticlust) (ex: 95%)
		Taxonomic assignation of OTU (ex: threshold 60, 70, 80%)
Molecular inventory	Molecular inventory	Molecular inventory

Correlation between the water quality assessment obtained with microscopy (x axis) and the 16 different bioinformatic strategies (y axis).
The biotic diatom index IPS (indice de Polluosensibilité Spécifique, Cemagref 1982) was calculated. IPS scores vary from 1 (bad quality status) to 20 (good quality status).

IPS scores calculated from ISU with no quality filters (raw data) were poorly correlated to microscopy.
Furthest neighbor and filtered ISU strategies provided similar results and were better correlated to microscopy than OptiClust.
It is important not to have a too high bootstrap: for biomonitoring purpose, it is better to determine neigbor species (which have similar ecologies) than not being able to determine it.



Schedule

1- Application of diatom biomonitoring to rivers

- Imitation: use of existing diatom indices
- Index based on phylogenetic inference of ecological profiles
- Taxonomy free indices
- Which bootstrap % to use for taxonomic assignation?

2- Source of bias ←

3- Intercalibration exercise

Bias

- **Physiological status of the cell / dead frustules**

The amount of DNA depends on the status of the cells

Taxon	frustule %	status	Transfo / biovolume
Achnanthidium minutissimum	75	~	40
Amphora pediculus	13	X	0
Navicula cryptotenella	10	V	50
Melosira varians	2	V	10

Bias

Completness of the reference database

For some species: no barcodes in the library

Diat.barcode

Taxon	frustule %	Barcode library	Transfo / biovolume
Achnanthidium minutissimum	75	No	0
Amphora pediculus	13	Yes	52
Navicula cryptotenella	10	Yes	40
Melosira varians	2	Yes	8

Bias

Think about neighbor/sister species

Taxon	frustule %
Encyonema minutum	75
Amphora pediculus	13
Navicula cryptotenella	10
Melosira varians	2

Taxon	Transfo / biovolume
Encyonema silesiacum	75
Amphora nd	13
Navicula cryptotenelloides	10
Melosira varians	2

Schedule

1- Application of diatom biomonitoring to rivers
  - Imitation: use of existing diatom indices
  - Index based on phylogenetic inference of ecological profiles
  - Taxonomy free indices
  - Which bootstrap % to use for taxonomic assignation?

2- Source of bias

3- Intercalibration exercise ←