iBOL EUROPE

eDNAqua-Plan
NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

Barcode reference library curation

Brent C. Emerson and Filipe O. Costa
Oct 28, 2024

Funded by the European Union

UK Research and Innovation

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Biodiversity Genomics Europe

Barcode reference library curation

The opportunity Barcode sequences offer tremendous potential to democratize taxonomy and meaningfully address the “taxonomic impediment”.

The challenge Existing repositories for barcode sequences contain inaccurate sequence records, from simple misspellings through to taxonomic misassignment and sample contamination.

The consequences

*For **barcoding***: conflicting or incorrect taxonomic assignment of newly sequenced material.

*For **metabarcoding***: problematic on two fronts
- incorrect assignment of taxonomy to OTUs
- overestimation of species number

Biodiversity Genomics Europe

Standard processing of metabarcode data

no strict order

01 Demultiplexing

02 Adapter removal and quality filtering

03 Mate pairing and merging, length trimming and filtering, dereplication

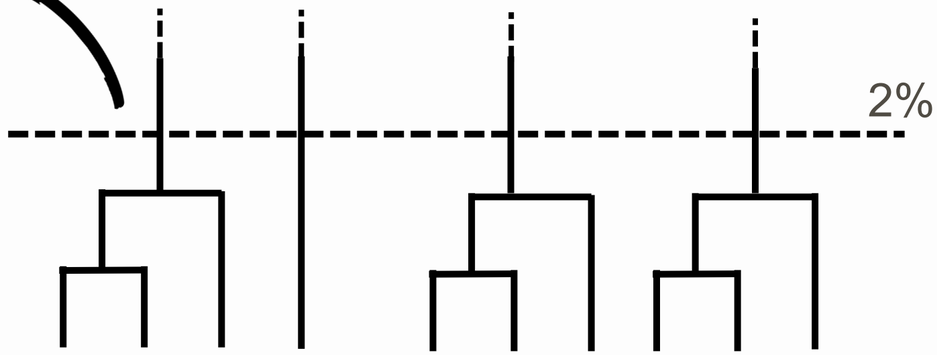
04 Chimera removal

05 Denoising

06 Optionally clustering or post-denoising filtering

07 Diversity estimations, taxonomic classification etc.

- PCR errors
- Sequencing errors
- **Nuclear-mitochondrial DNA segments (NUMTS)**



Biodiversity Genomics Europe

iBOL EUROPE

Standard processing of metabarcode data

01 Demultiplexing

02 Adapter removal and quality filtering

03 Mate pairing and merging, length trimming and filtering, dereplication

04 Chimera removal

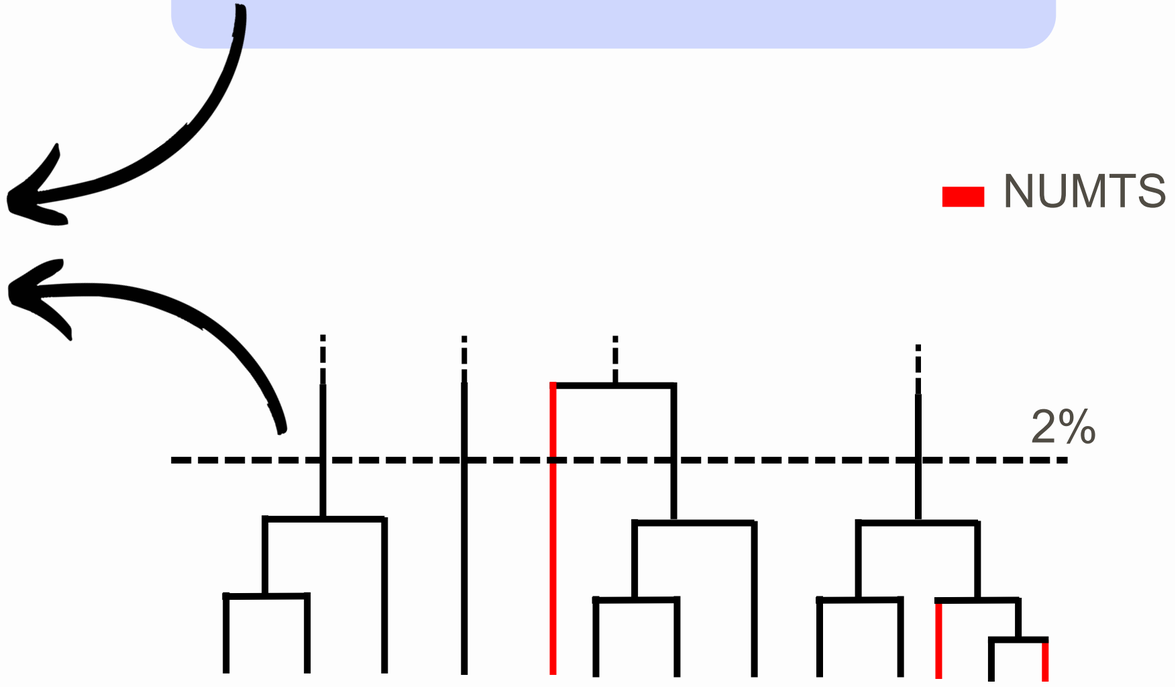
no strict order

05 **Denoising**

06 Optionally clustering or post-denoising filtering

07 Diversity estimations, taxonomic classification etc.

- PCR errors
- Sequencing errors
- **Nuclear-mitochondrial DNA segments (NUMTS)**



Secondary denoising of metabarcode data

01 Demultiplexing

02 Adapter removal and quality filtering

03 Mate pairing and merging, length trimming and filtering, dereplication

no strict order

04 Chimera removal

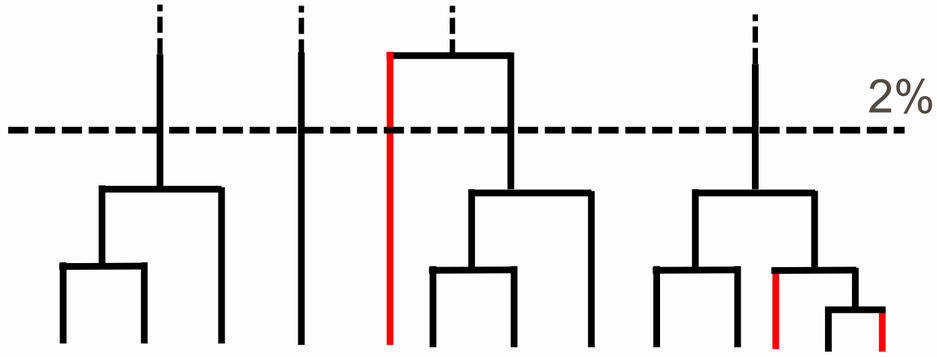
05 Denoising

06 Optionally clustering or post-denoising filtering

07 Diversity estimations, taxonomic classification etc.

OTUs from NUMTs can **exceed** the number of OTUs from real species

Barcode reference sequences can remove NUMT OTUs using secondary denoising with validation



BGE pathway to curated reference libraries

Biodiversity Genomics Europe

Work Package 10: Barcoding Applications

Task 10.1: Reference Library Curation

AIMS

- Curate about 200,000 already available sequences from ~25,000 species (on BOLD) to address current inconsistencies
- ~45,000 new barcode sequences from ~15,000 species will be generated during the project period to fill existing gaps in the database
- Main focus groups are pollinators (lepidopterans, hoverflies and bees - and freshwater and marine invertebrate species used in environmental monitoring)
- publicly available reference libraries of European barcode sequences

Biodiversity Genomics Europe

Biodiversity Genomics Europe

Work Package 10: Barcoding Applications

Task 10.1: Reference Library Curation

APPROACH

(1) Identify inconsistencies

- BIN sharing (2 or more species in the same BIN)
  - *outdated taxonomy, misspelling, misidentification, inadequate resolution, introgression*
- Multiple BINS (1 species has more than 1 BIN)
  - *Monophyly (cryptic diversity) vs paraphyly (misidentification)*

(2) Correction of inconsistencies and metadata acquisition

(3) Sequence records scored from 1-6 according to sequence and metadata quality

(4) Species are graded according to BIN-sharing and -splitting

BGE pathway to curated reference libraries

	specimen rank					
Criteria	**1**	**2**	**3**	**4**	**5**	**6**
Species level ID	✓	✓	✓	✓	✓	✓
Type specimen	✓					
Good quality sequence		✓	✓	✓	✓	✓
Public voucher (has museum ID)		✓(or)	✓(or)			
Public voucher (agreed institution)		✓(or)	✓(or)			
Public voucher (agreed voucher_type)		✓(or)	✓(or)			
image(s)		✓	✓	✓	✓	
Identifier named		✓(or)	✓(or)			
ID method (incl. morphology)		✓(or)	✓(or)			
Collection (Site)		✓				
Collection (Date)		✓				
Collection (Country)		✓	✓	✓		
GPS coordinates		✓				
Collector named		✓				

Figure 1: Ranking system to pick representatives for each haplotype / species

Blue = **minimum requirements** to go forward into BAGS analysis.
Green = good records.
Orange = bad records.

GRADE A = >10 specimens in 1 BIN
GRADE B = 3-10 specimens in 1 BIN
GRADE C = >1 BIN
GRADE D = < 3 specimens in 1 BIN
GRADE E = BIN sharing (>1 species in single BIN)

BGE pathway to curated reference libraries

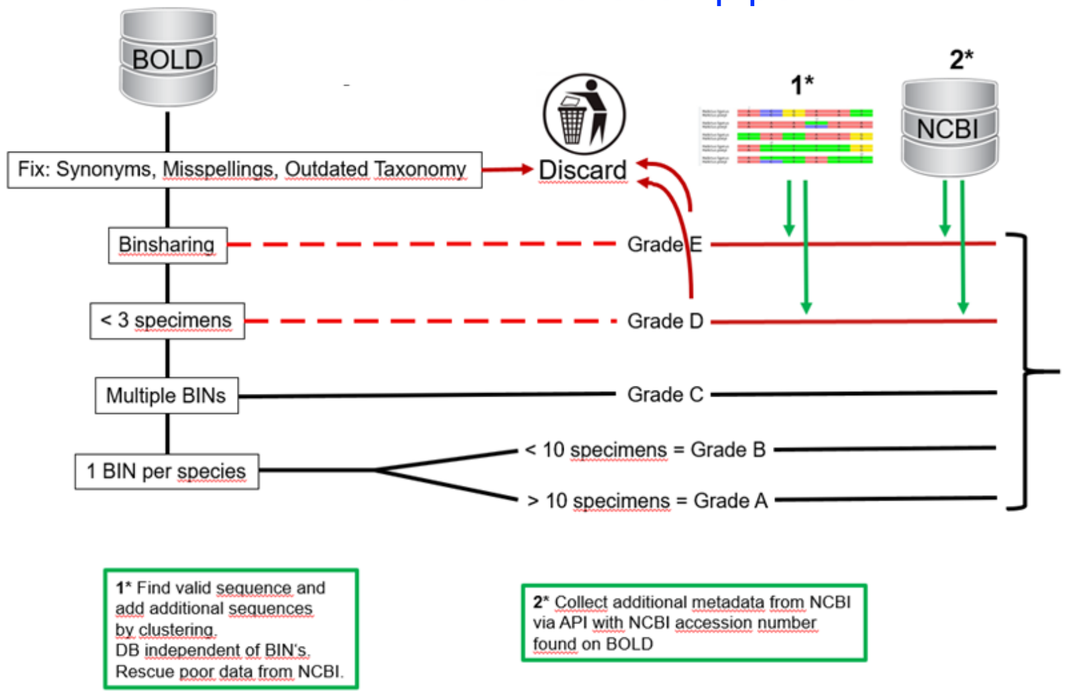
Biodiversity Genomics Europe

Work Package 10: Barcoding Applications

Task 10.1: Reference Library Curation

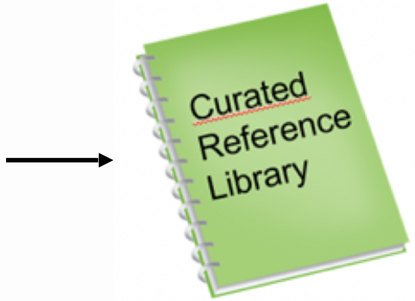
APPROACH

Automated curation pipeline



Manual curation and validation by taxonomic expert(s)

Typically family-level



BGE pathway to curated reference libraries

Biodiversity Genomics Europe

Work Package 10: Barcoding Applications

Task 10.1: Reference Library Curation

PROGRESS

- Curated barcode reference library for the Geometridae of Europe completed and being prepared for publication
- Automated curation for other Lepidoptera completed, manual curation in progress
- Other pollinators + select freshwater and marine invertebrate species pending

Thank you!

Funded by
the European Union

UK Research
and Innovation

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

eDNAqua-Plan

NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

A **Plan** towards an **eDNA reference library** and **data repository** for **Aquatic** Organisms, navigating **Europe** towards the next generation **biodiversity monitoring**

iBOL EUROPE

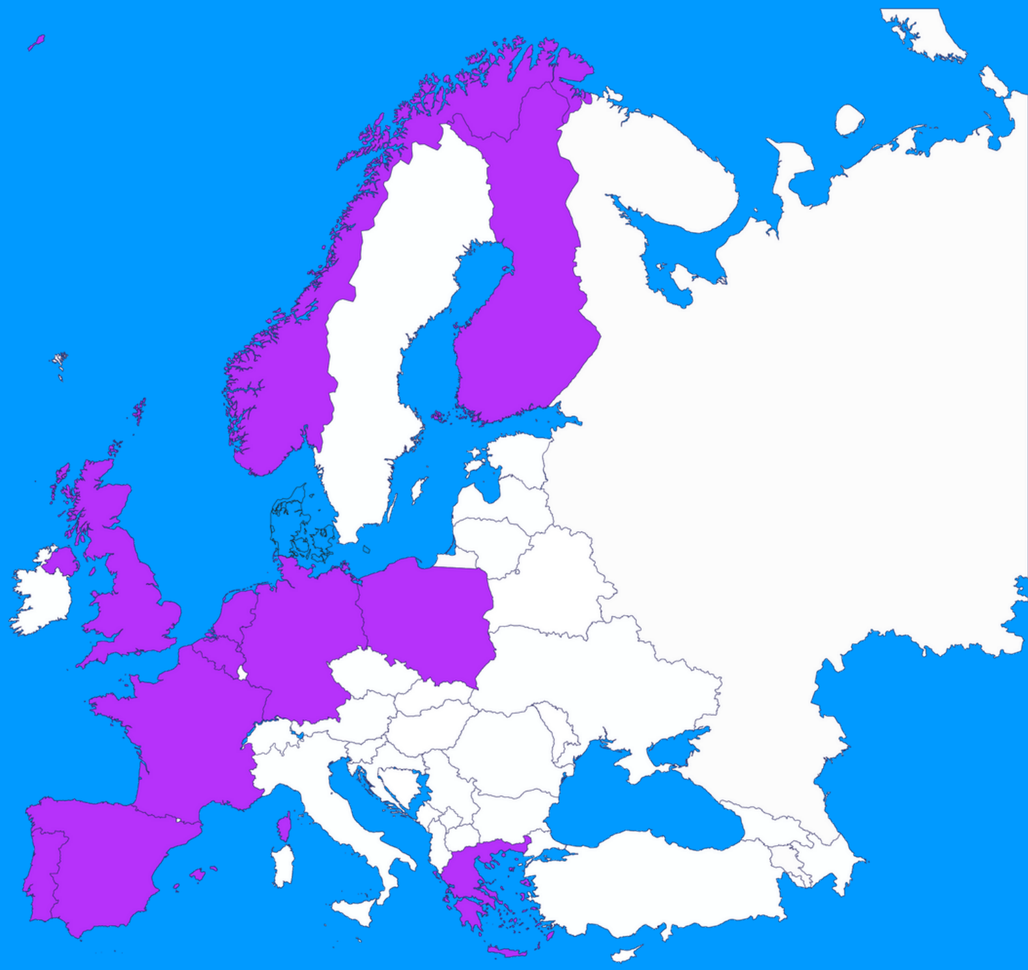
Funded by the European Union

Funded by the European Union under the Horizon Europe Programme, Grant Agreement No. 101112800 (eDNAqua-Plan). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

eDNAqua-Plan

NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

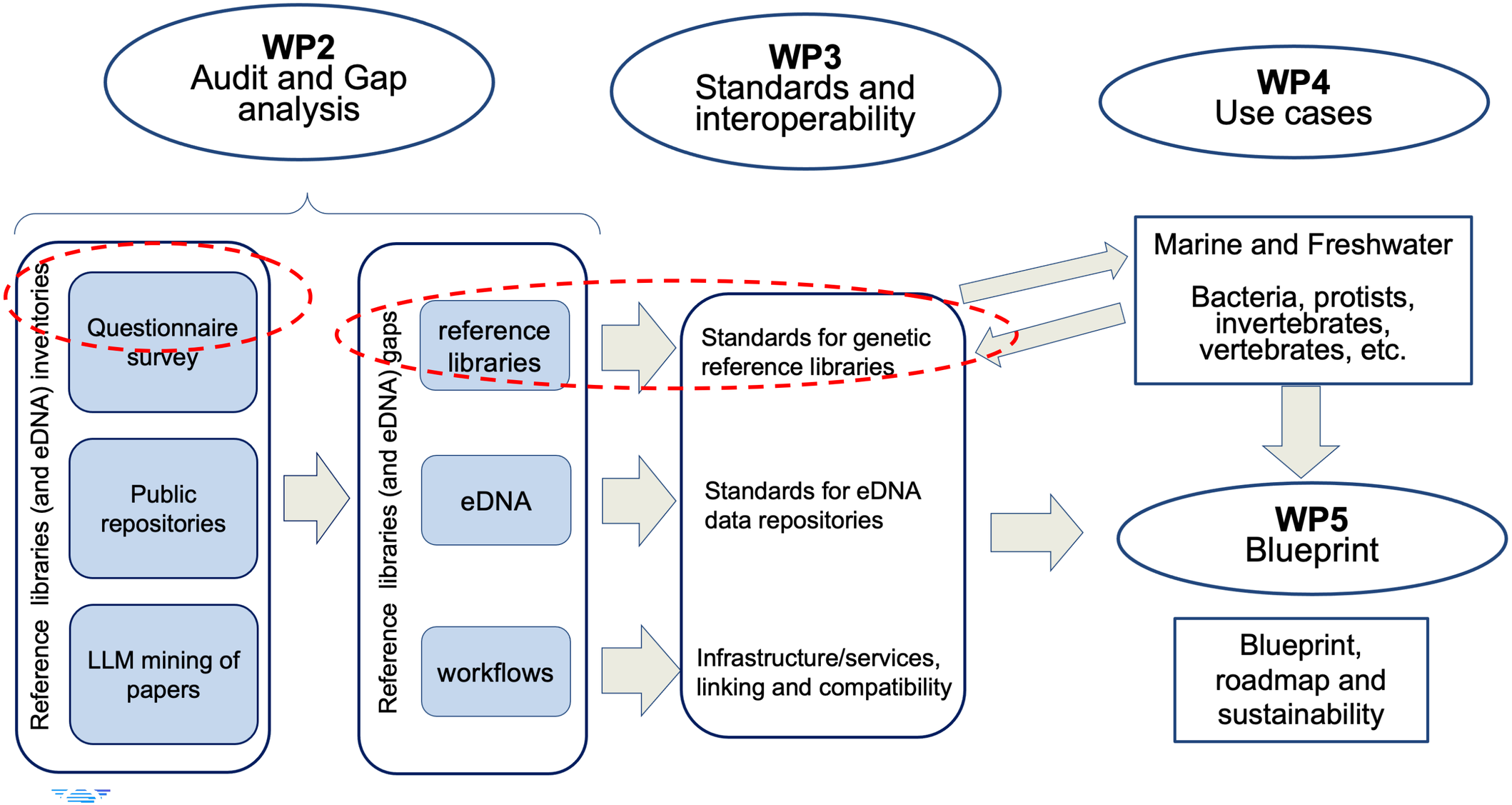
- 18 partners from 11 countries
- 36 months (started Sept. 2023)



Funded by
the European Union

eDNAqua-Plan
NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

iBOL EUROPE



eDNAqua-Plan
NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

iBOL EUROPE

WP2: Landscape and gap analysis

- **40 data attributes** reviewed / all taxa and markers considered
- **21 repositories** searched
- **38 questionnaire** replies from European Countries (acknowledgments to contributors!)
- **Large Language Model (LMM) searched > 850 publications**
- **Long report**, yet to be published, but summary available on the website: **https://ednaquaplan.com**

eDNAqua-Plan
NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

iBOL EUROPE

WP2: Landscape and gap analysis

Summary of main gaps and general findings

- Status of reference libraries **suboptimal**, with high **heterogeneity in quality** and levels of taxonomic **completion**
- **Absence (or minimal) metadata standards**, and many key metadata components missing
- Basal and/or insufficient implementation of comprehensive data **QA/QC systems**
- In most cases absence of post-barcoding assessment of **taxonomic accuracy**, and respective **annotation** systems
- Numerous cases of limited (or absent) **accessibility and interoperability** (low levels of FAIRness)

eDNAqua-Plan
NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

iBOL EUROPE

BG24

WP3: Standards for genetic reference libraries

- **Building on the literature, and gaps and findings of WP2**

Develop recommendations for genetic reference libraries with a focus on:

- **Propose standards,** required information and validation criteria for reliable reference records
- **Evaluate systems for curating** reference libraries and essential metadata that allow curation
- **Propose solutions to improve the interoperability** between standards and databases towards FAIR reference libraries

eDNAqua-Plan
NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

WP3: Standards for genetic reference libraries

- **Building on community expertise**

Discussions with related projects and initiatives

Focus on essential metadata and curation procedures

Workshop with the eDNAqua-Plan consortium (last September)

3 main discussion topics:

- Storage of vouchers, data and metadata
- Taxonomic curation procedures
- Machine-readable workflows and interoperability

eDNAqua-Plan
NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

iBOL EUROPE

BG24

CONCLUSIONS

- **Strong investment needed**: further complete and improve aquatic European barcode reference libraries.
- Work in progress to produce recommendations for reference libraries that are:

  a) Accurate: integrates full-scale taxonomic curation procedures**;**

  b) Auditable and reliable: relevance of metadata quality and standards

  c) Harmonized and interoperable: structured metadata formats (e.g. JSON,)

  d) Emphasis on FAIR principles.

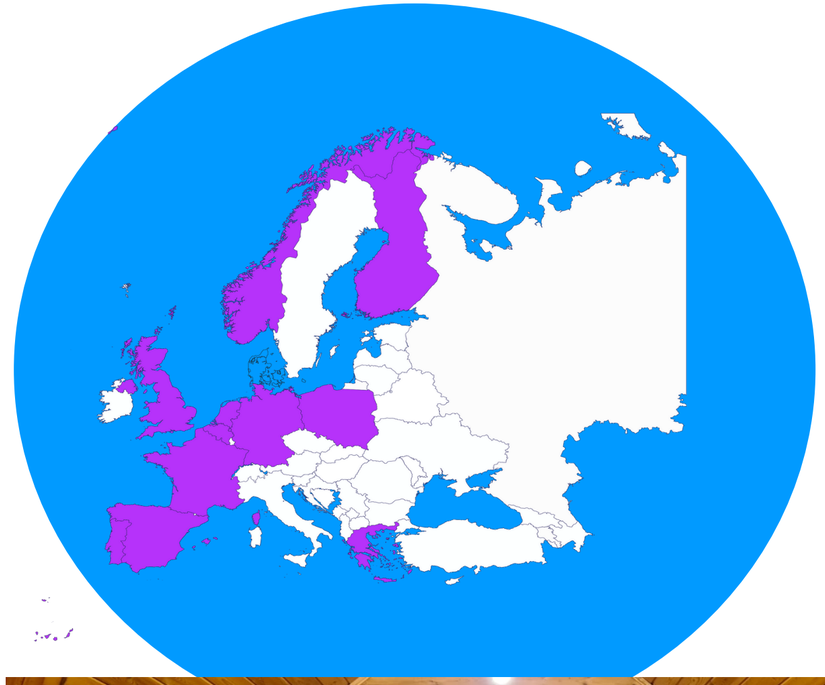
  e) Sustainable: sustainability plan to maintain, curate and update libraries

- A lot of work ahead (in partnership with iBOL Europe and other relevant initiatives)

eDNAqua-Plan

NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

iBOL EUROPE



External advisers

Jenny Giles – CSIRO, Kristy Deiner – UTH Zurich, Toshifumi Minamoto – Kobe University, Pier Luigi Buttigieg – Max Planck Institute for Marine Microbiology, Mehrdad Hajibabaei – iESTF

Funded by the European Union

https://ednaquaplan.com