A multi-evidence approach for flagging taxonomic misidentifications in marine eDNA metabarcoding datasets

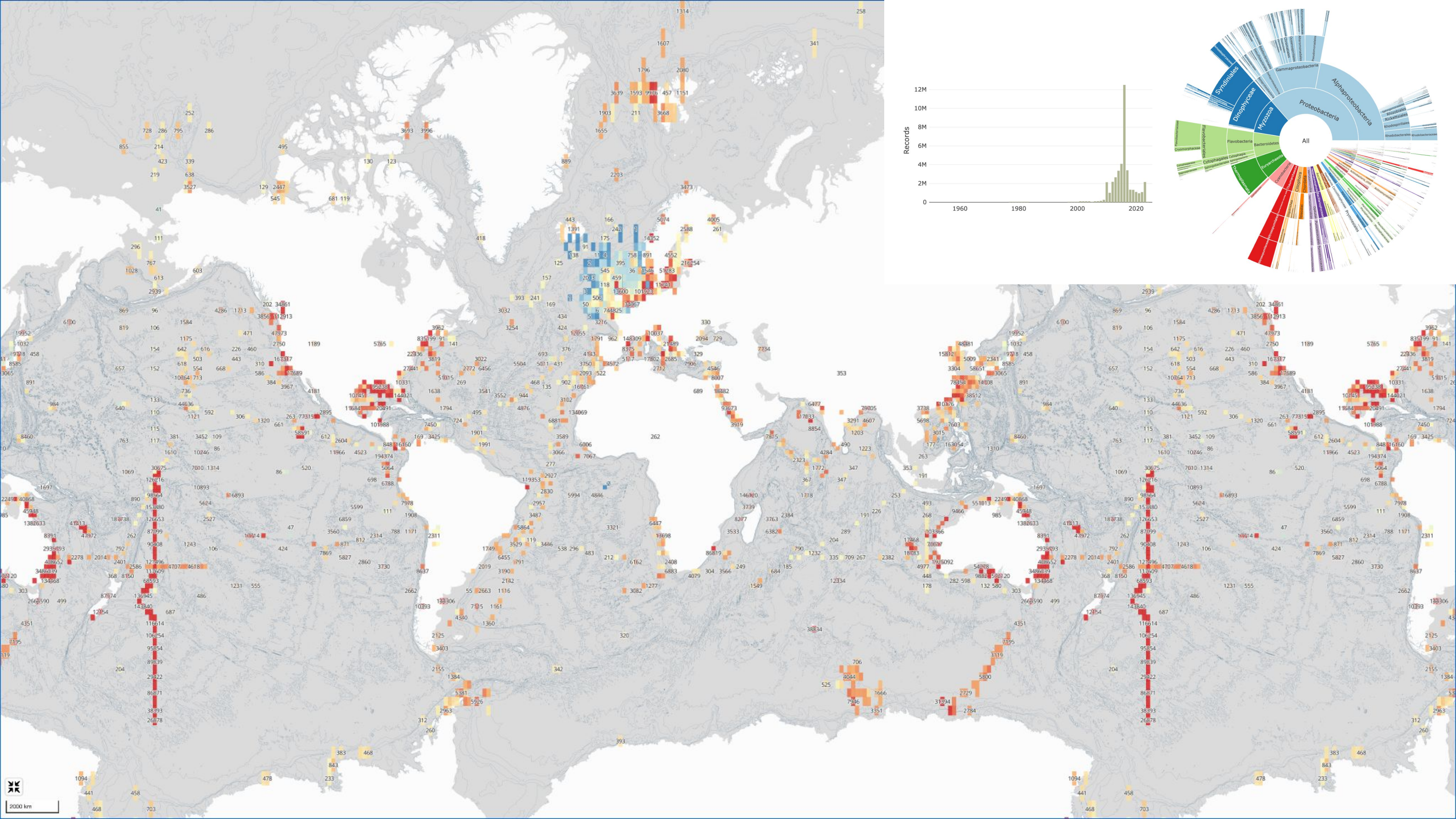
Pieter Provoost, Saara Suominen, Silas Principe, Ward Appeltans

Ocean Biodiversity Information System (OBIS)
Intergovernmental Oceanographic Commission (IOC-UNESCO)
helpdesk@obis.org

LIVING DATA 2025

unesco
Intergovernmental Oceanographic Commission

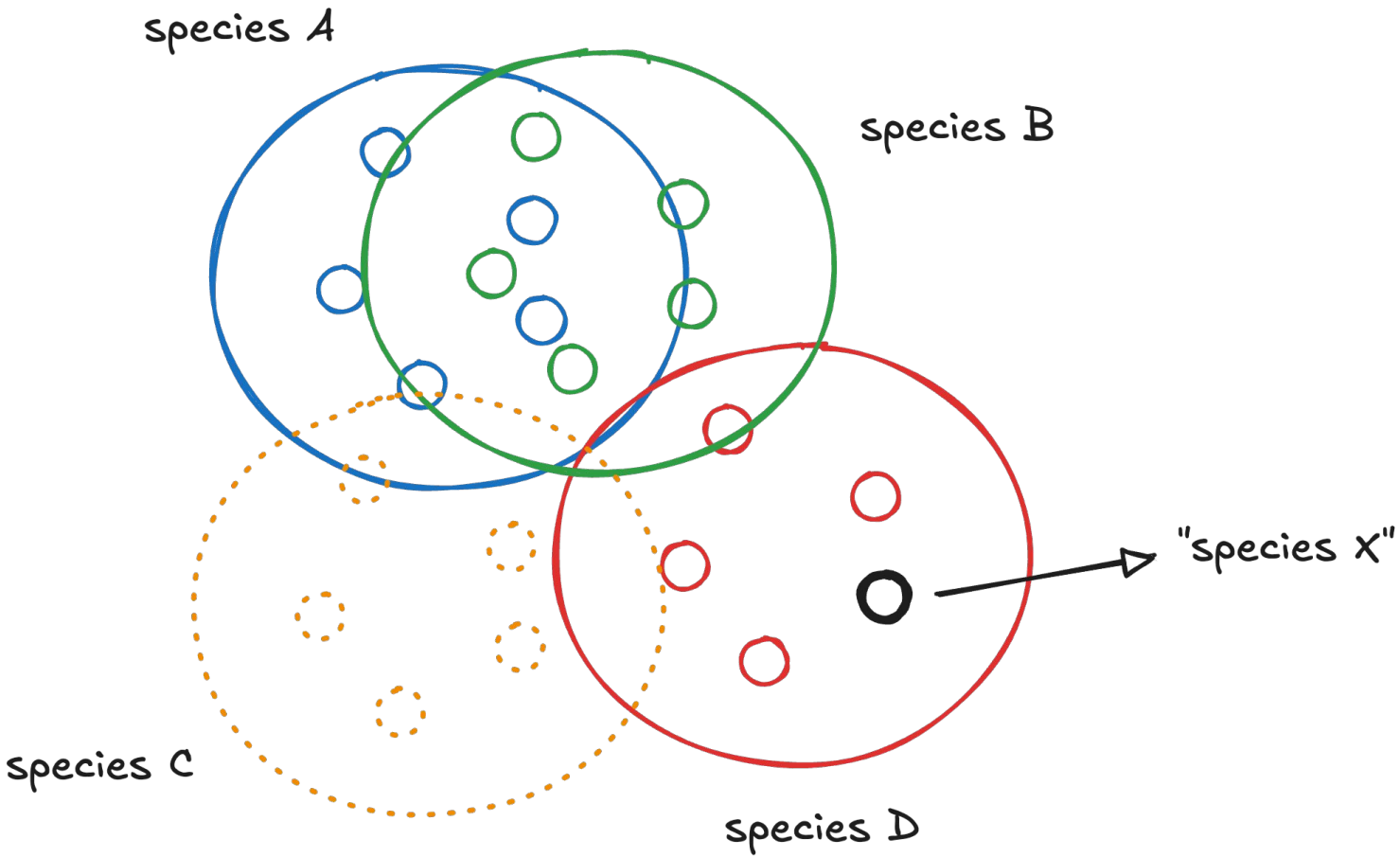
OBIS
OCEAN BIODIVERSITY INFORMATION SYSTEM

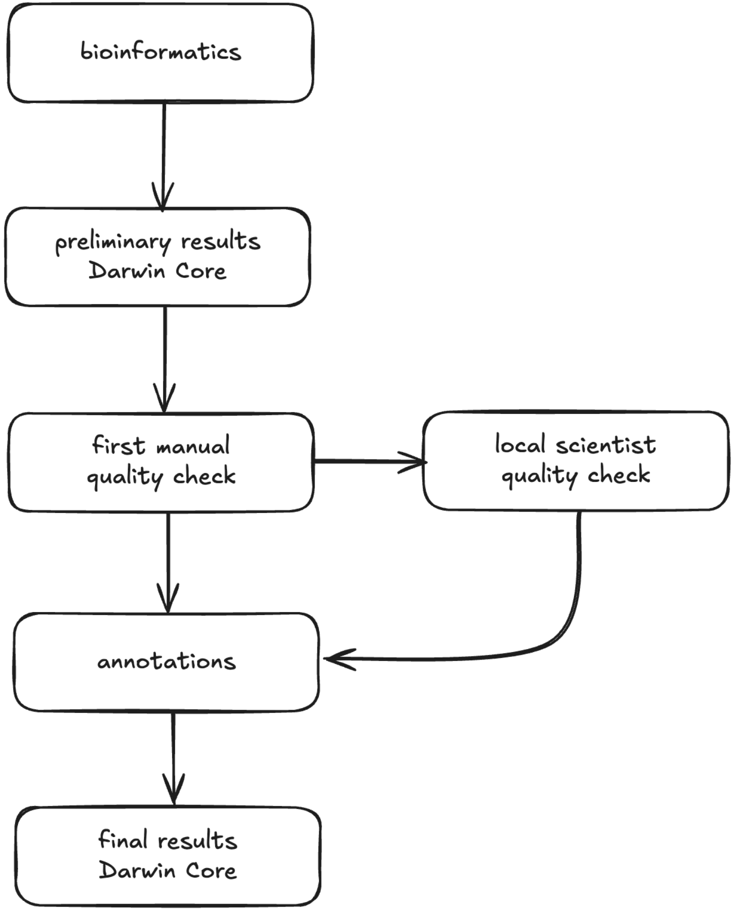

Records
12M
10M
8M
6M
4M
2M
0
1960
1980
2000
2020
All
Proteobacteria
Alphaproteobacteria
Gammaproteobacteria
Rhodobacterales
Rickettsiales
Myzozoa
Dinophyceae
Syndiniales
Bacteroidetes
Flavobacteria
Flavobacteriales
Cytophagia
Euryarchaeota
2000 km


unesco
Intergovernmental Oceanographic Commission
OBIS
OCEAN BIODIVERSITY INFORMATION SYSTEM
Suva, Fiji

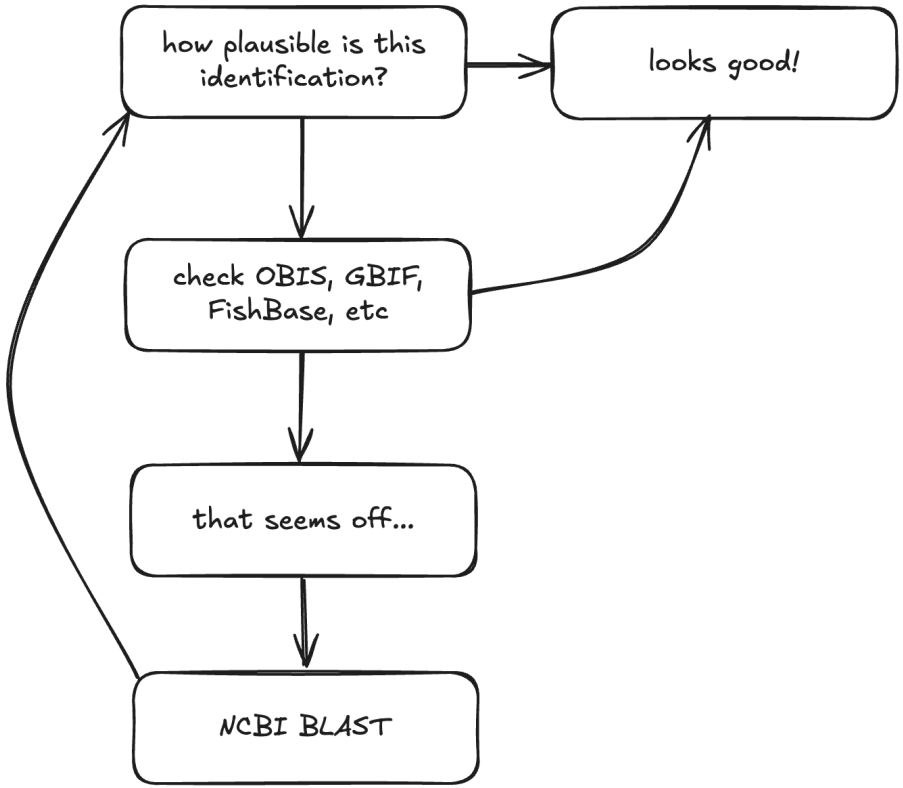
Banc d’Arguin National Park, Mauritania

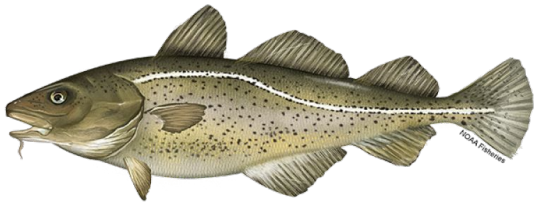
iSimangaliso, South Africa

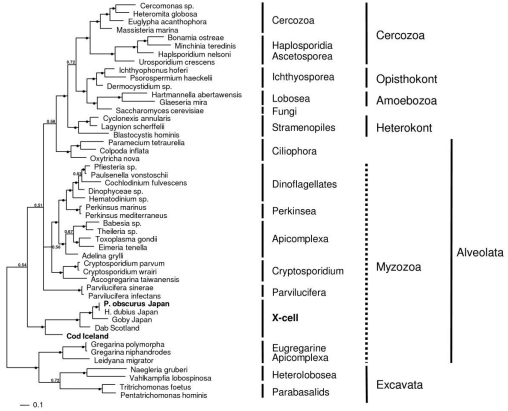

species A
species B
"species X"
species C
species D

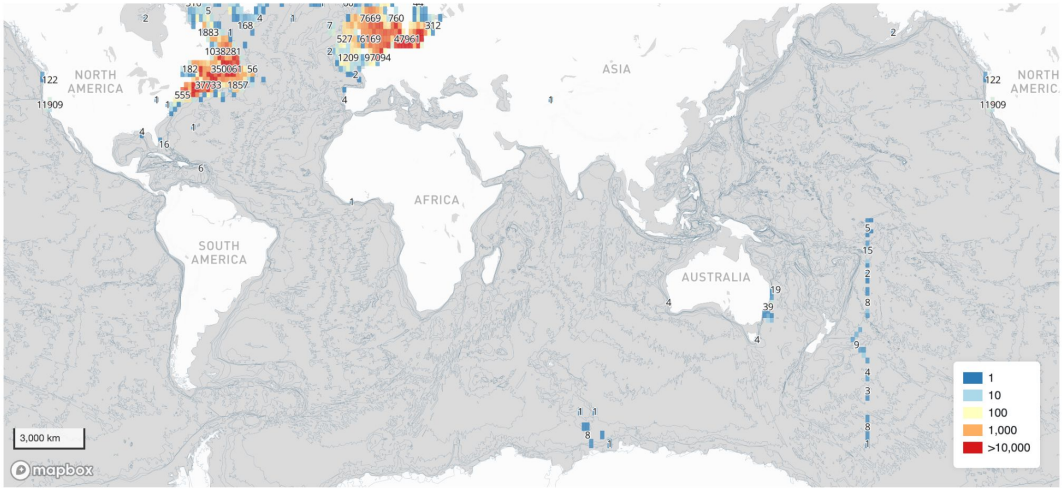

bioinformatics
preliminary results
Darwin Core
first manual
quality check
local scientist
quality check
annotations
final results
Darwin Core


UNESCO environmental DNA (eDNA) Expeditions in marine World Heritage sites
Biodiversity survey for Lord Howe Island Group
Interim Results
Not for public distribution
Credit: Justin Gilligan
1
Figure 2: Number of DNA reads (bubbles) and species detected (bars) by family.
4


how plausible is this identification?
looks good!
check OBIS, GBIF, FishBase, etc
that seems off...
NCBI BLAST







Freeman et al. Parasites & Vectors 2011, 4:15
http://www.parasitesandvectors.com/content/4/1/15

Parasites & Vectors

RESEARCH Open Access

Molecular identification and transmission studies of X-cell parasites from Atlantic cod *Gadus morhua* (Gadiformes: Gadidae) and the northern black flounder *Pseudopleuronectes obscurus* (Pleuronectiformes: Pleuronectidae)

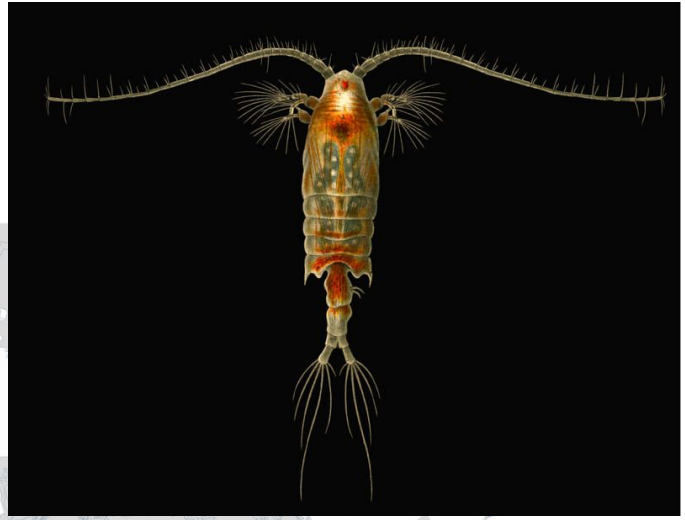
MA Freeman[1,2*†], M Eydal[3†], M Yoshimizu[5], K Watanabe[6], AP Shinn[2], K Miura[7], K Ogawa[4†]

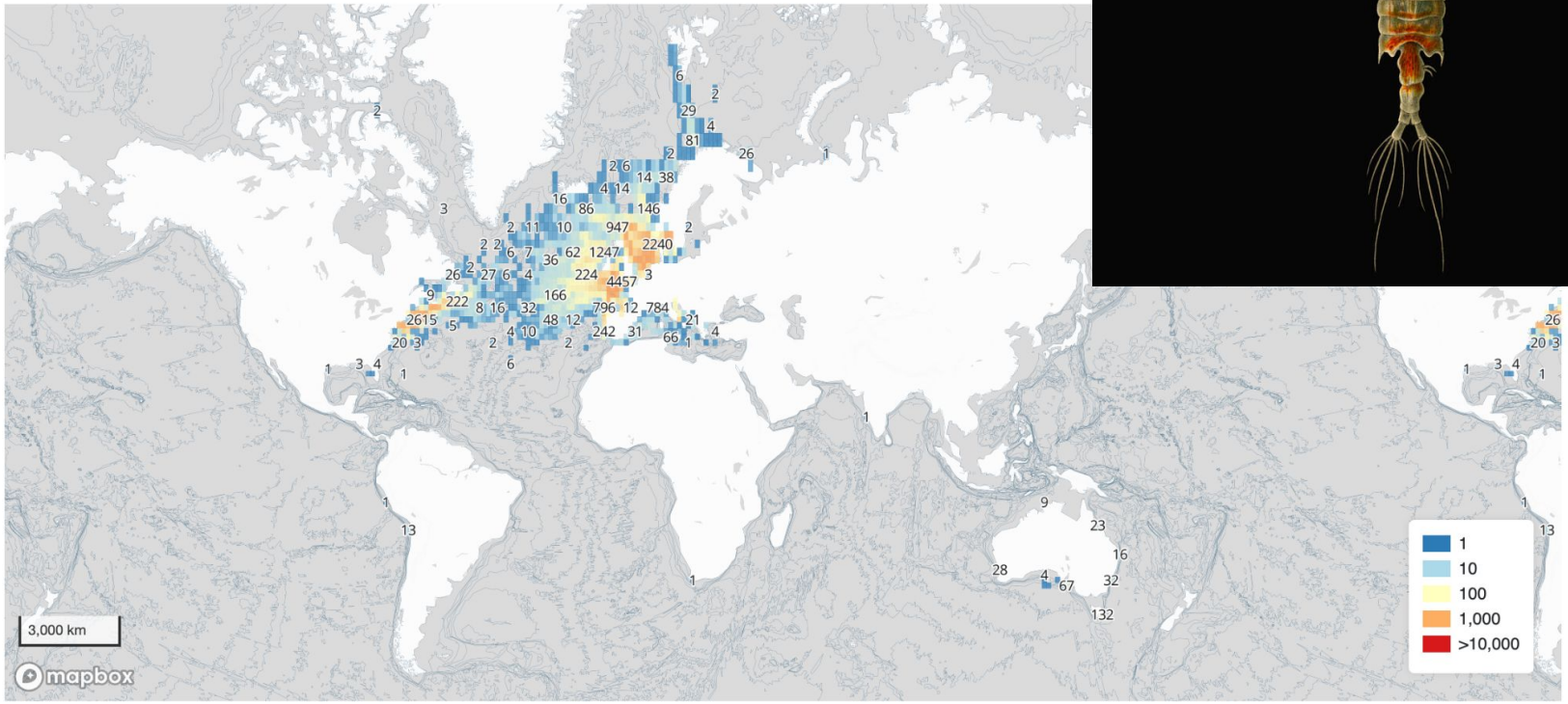
Figure 2 Epidermal X-cell pseudotumours on formalin-fixed *Pseudopleuronectes obscurus* from Hokkaido, Japan. a) A large central dorsal pseudotumour, seen in cross section (b). c) Pseudotumours can extend from the dorsal surface to the ventral surface and remain pigmented. d) Juvenile fish are also infected and ventral pseudotumours can also be unpigmented. Scale bars a, c & d = 3 cm, b = 10 mm.

Centropages typicus Krøyer, 1849 Species

Animalia > Arthropoda > Copepoda > Calanoida > Centropagidae > Centropages

WoRMS: 104499 NCBI: 463189 Open in mapper





Porites astreoides Lamarck, 1816 Species

Animalia > Cnidaria > Hexacorallia > Scleractinia > Poritidae > Porites

Vernacular names: knobby porous coral, mustard hill coral, yellow porites, yellow porous coral

WoRMS: 288889 NCBI: 104758 Open in mapper

Image by Dan Schofield

Anguilla dieffenbachii Gray, 1842 Species EN

Animalia > Chordata > Teleostei > Anguilliformes > Anguillidae > Anguilla

Vernacular names: New Zealand longfin eel

WoRMS: 271703 NCBI: 61127 Open in mapper

SPECIES | ACCEPTED

Alligator sinensis Fauvel, 1879

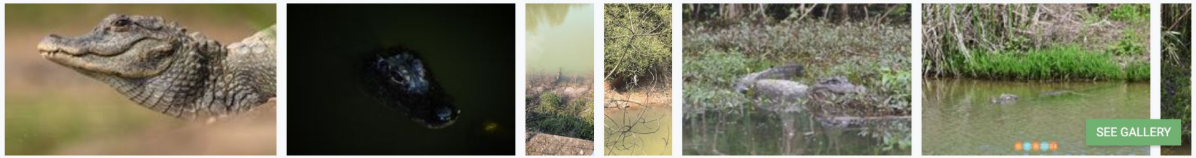
Published in: Fauvel, A. A. Alligators in China: Their History, Description and Identification.

In: GBIF Backbone Taxonomy

China Alligator In English **Trade restrictions I** CITES 2020

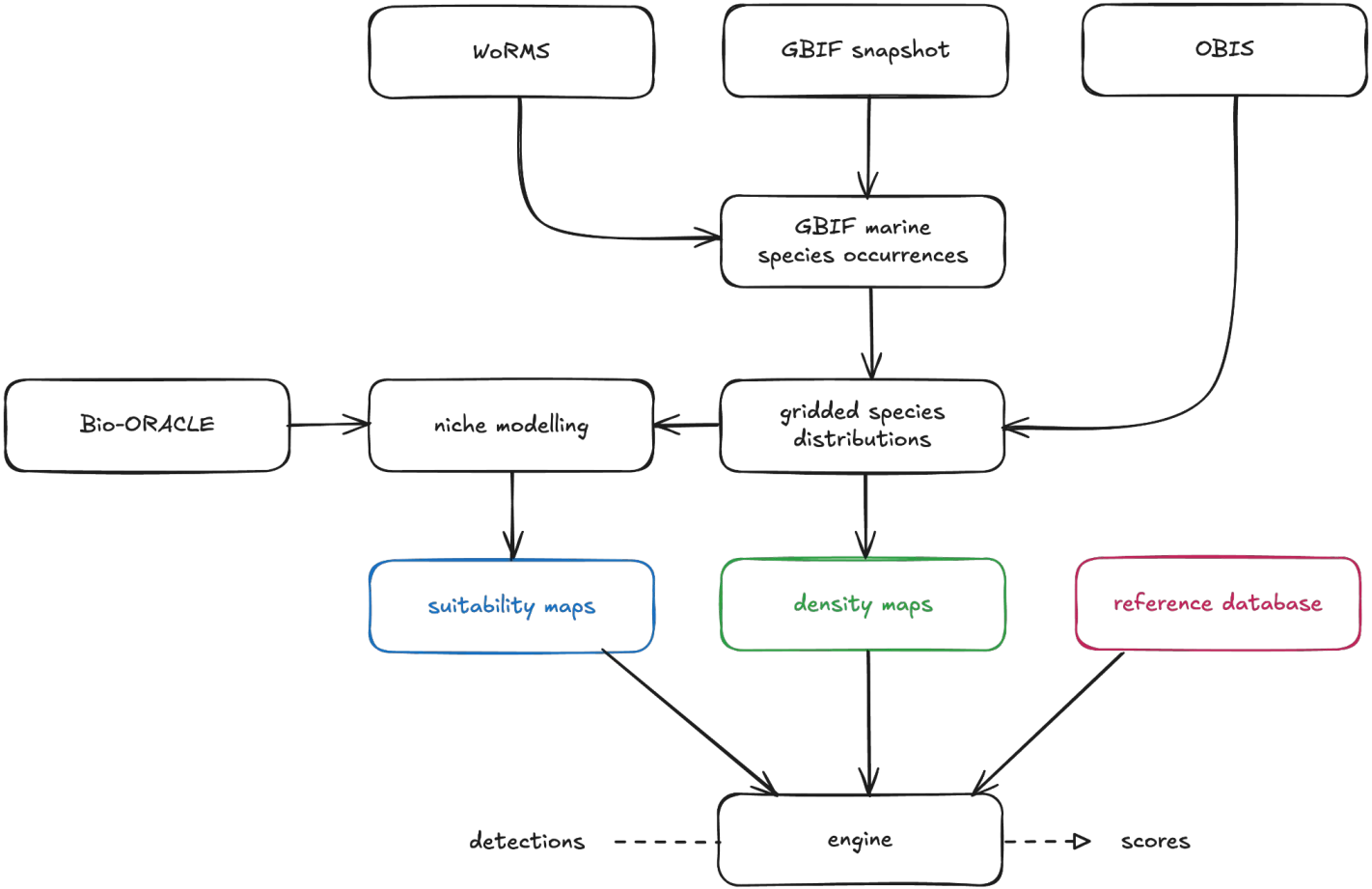
OVERVIEW METRICS 286 OCCURRENCES 1 INFRASPECIES

34 OCCURRENCES WITH IMAGES

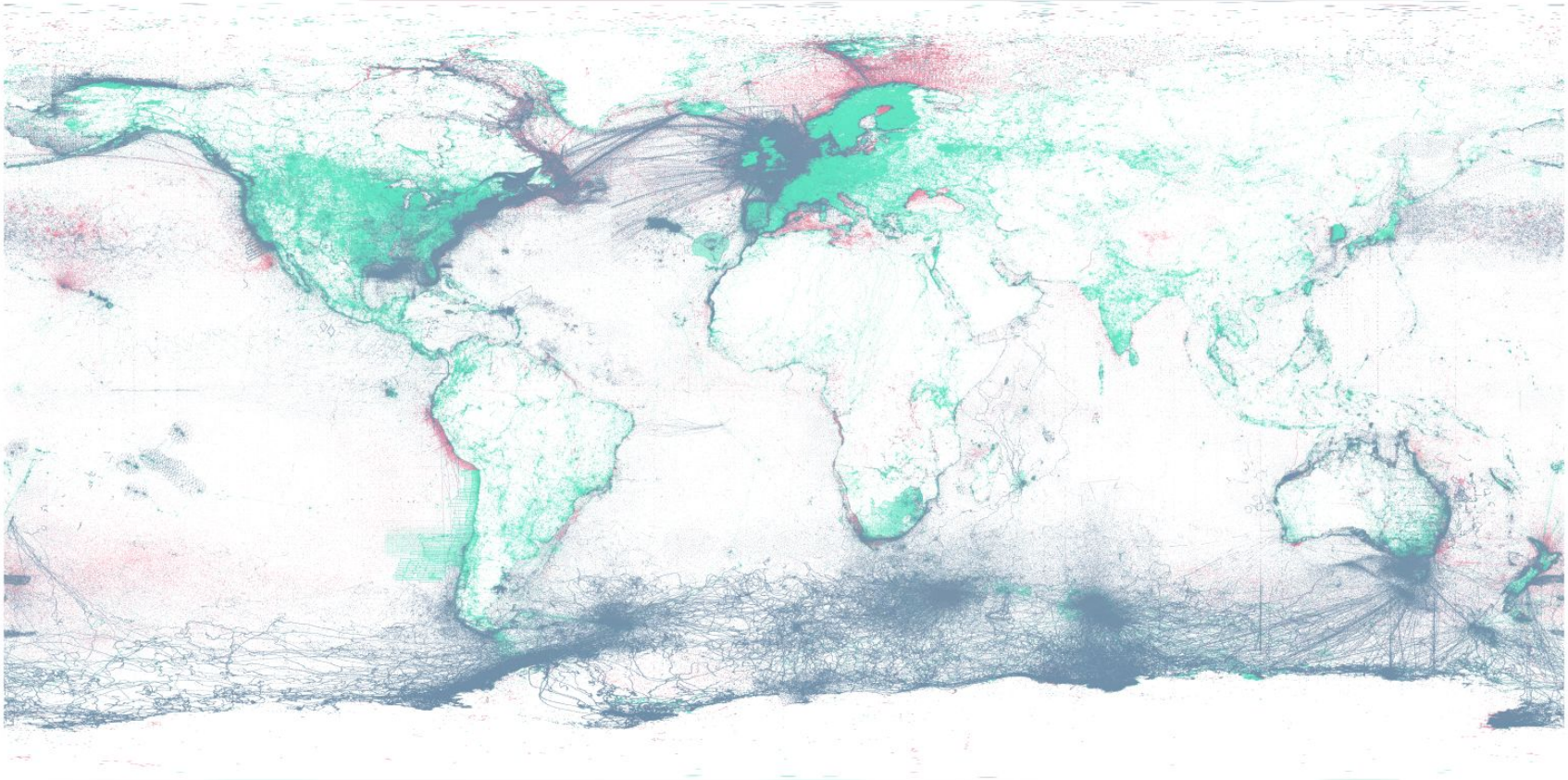


39 GEOREFERENCED RECORDS

Any year 1922 - 2025 EXPLORE AREA


WoRMS
GBIF snapshot
OBIS
GBIF marine species occurrences
Bio-ORACLE
niche modelling
gridded species distributions
suitability maps
density maps
reference database
detections
engine
scores

https://github.com/iobis/speciesgrids



Lycodes lavalaei

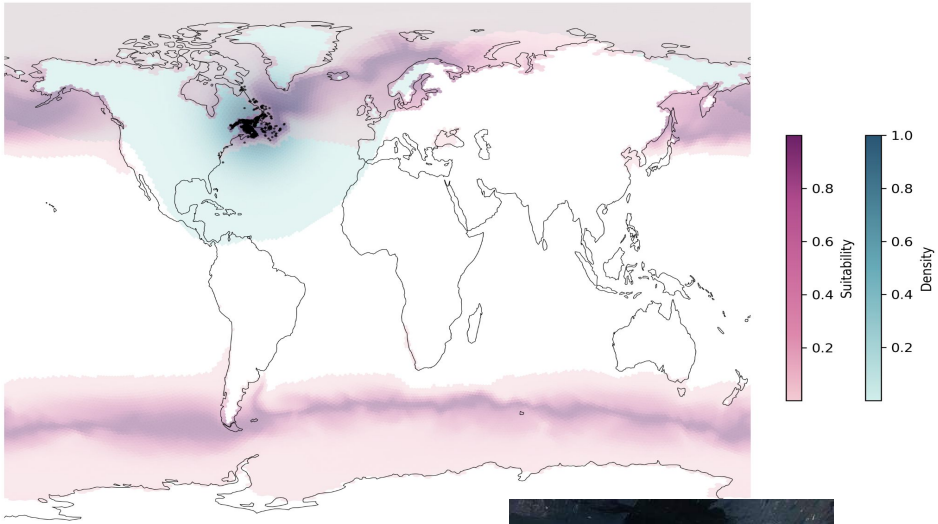
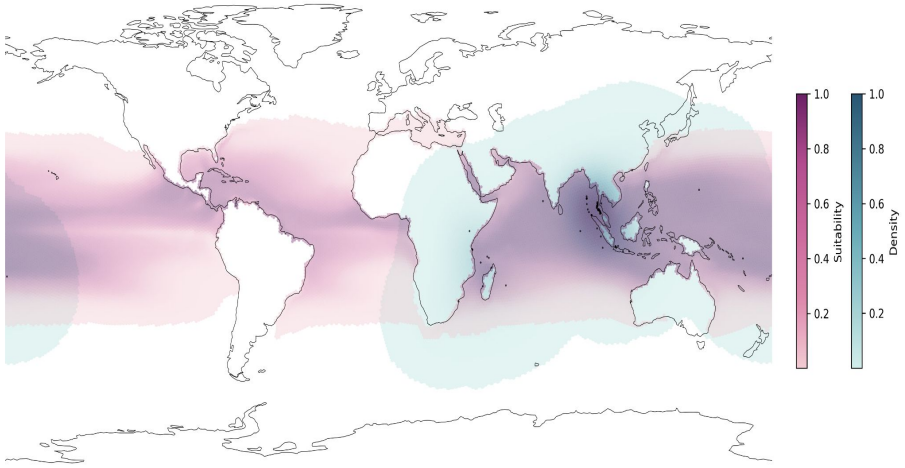
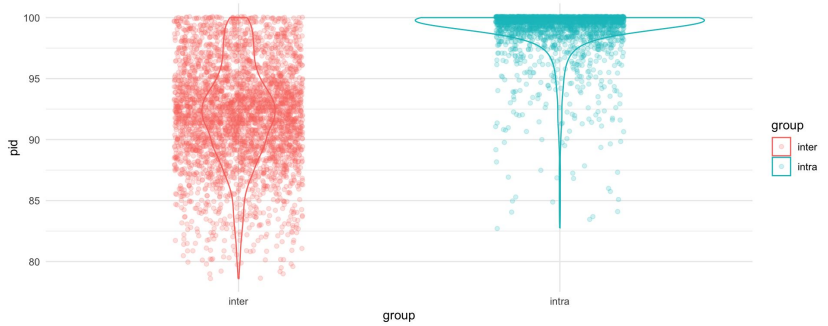
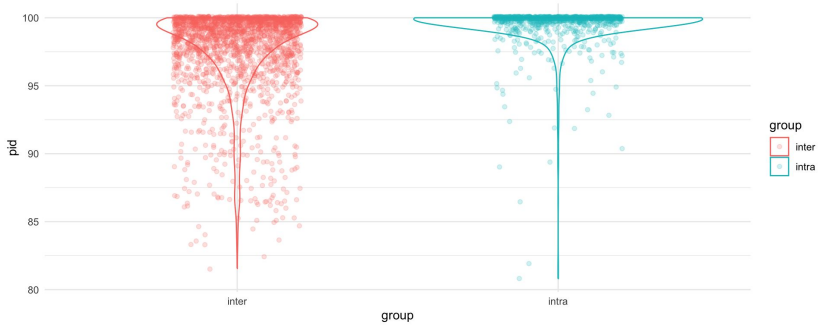


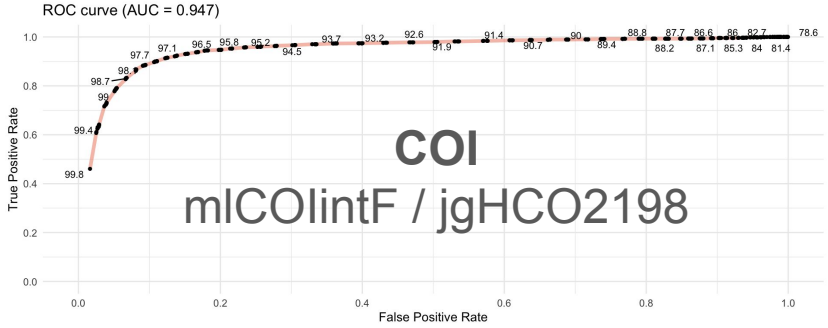
Image by
Vicki Johnson

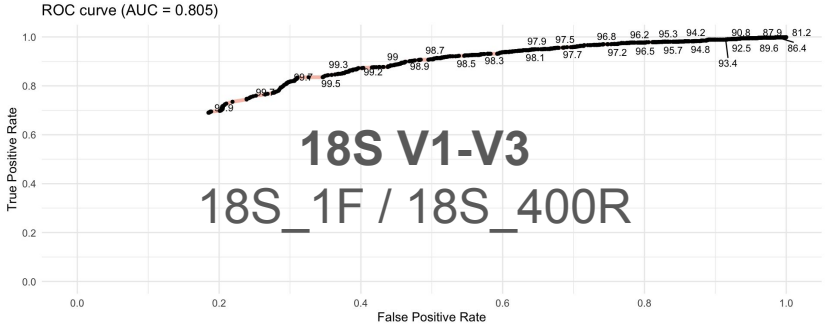
Amphiprion ephippium

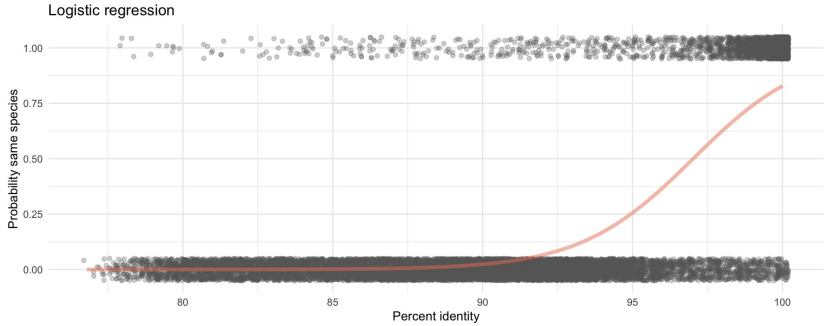


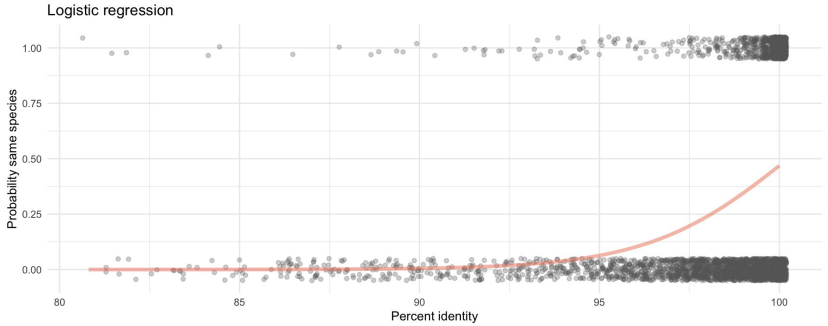

pid
group
inter
intra
group

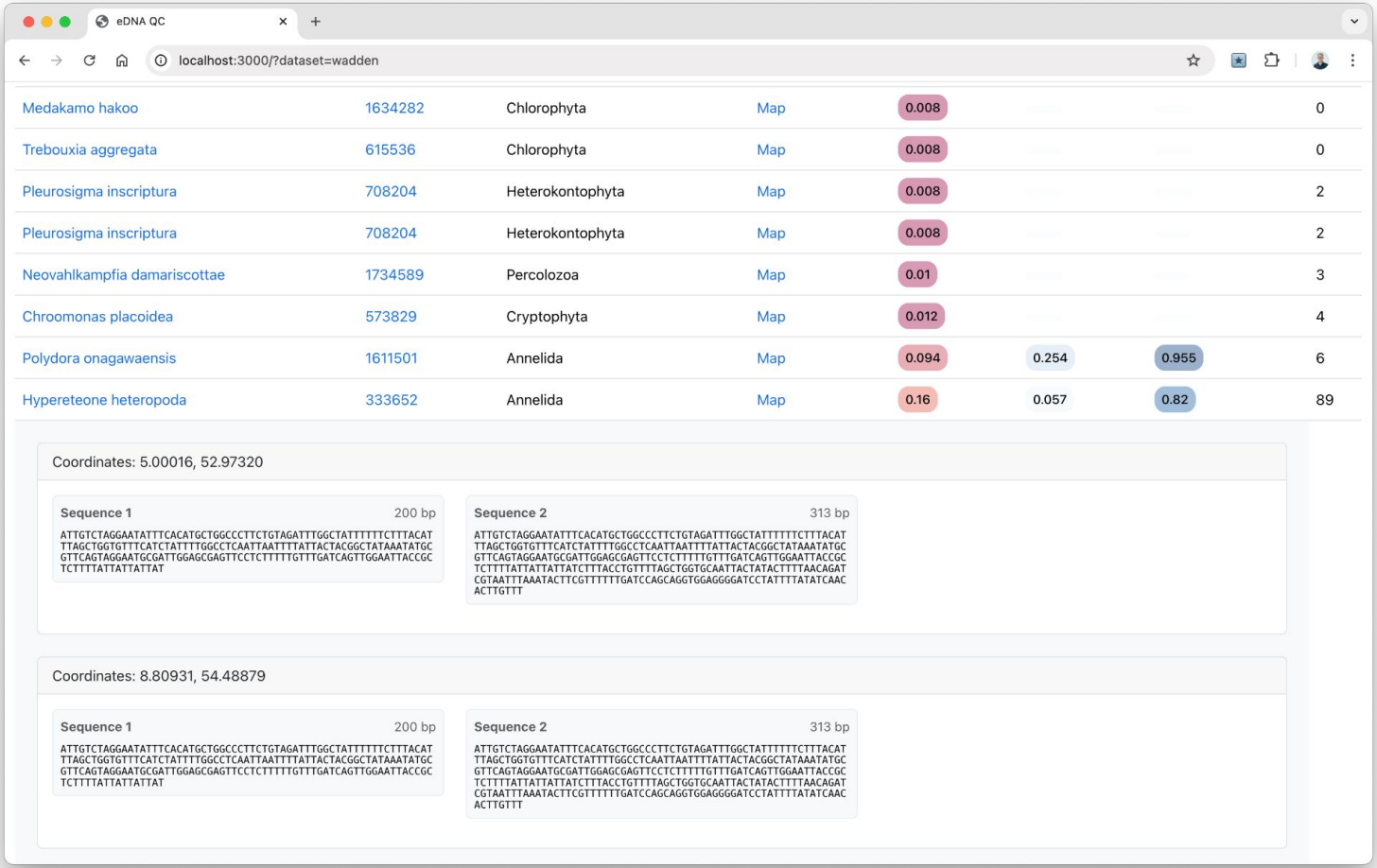

pid
group
inter
intra
group

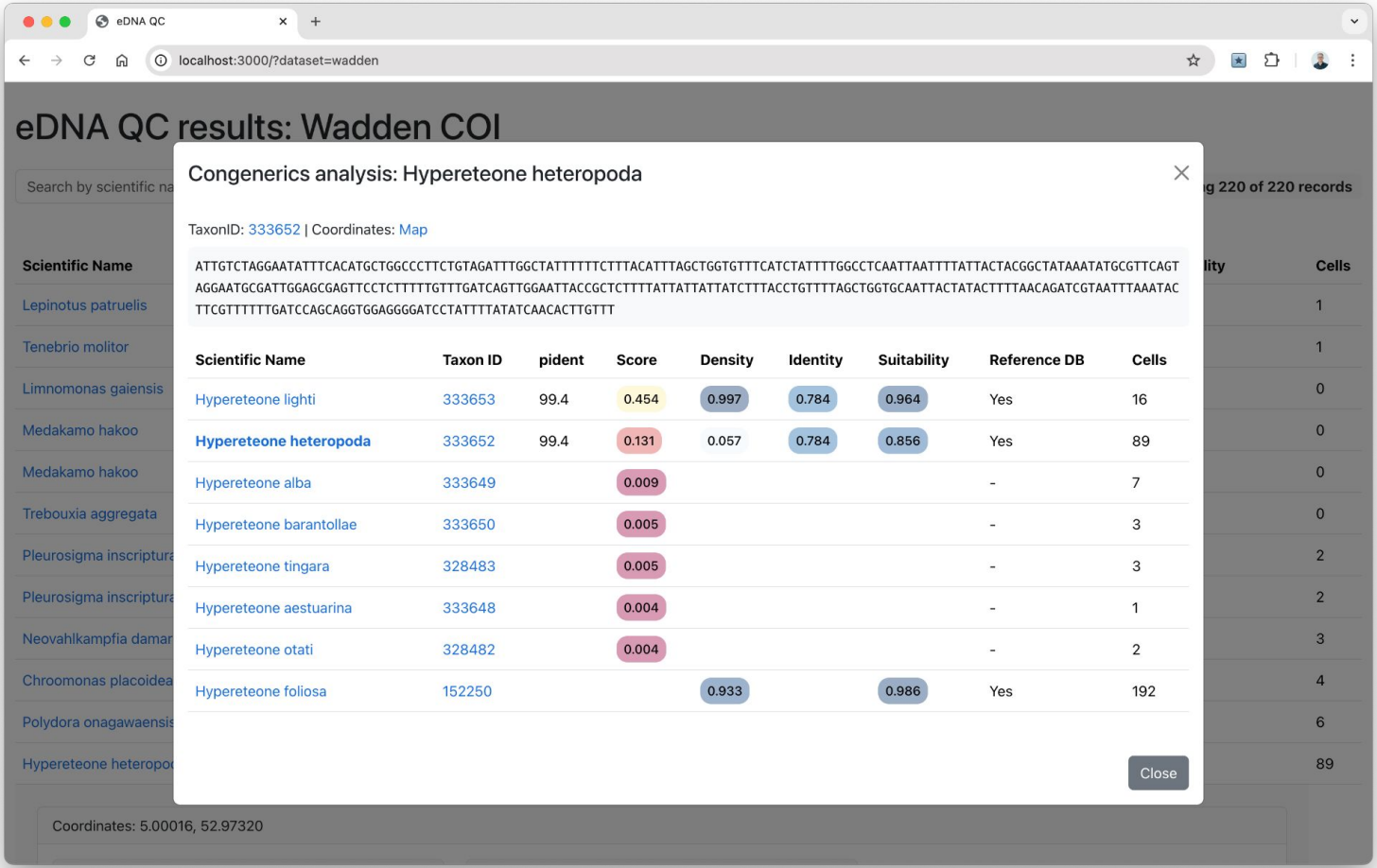

ROC curve (AUC = 0.947)
True Positive Rate
False Positive Rate
COI
mlCOIintF / jgHCO2198

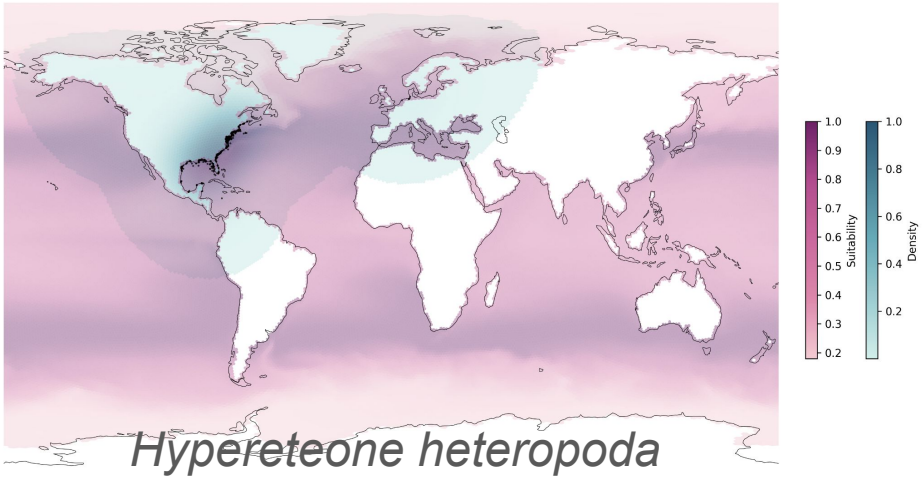

ROC curve (AUC = 0.805)
True Positive Rate
False Positive Rate
18S V1-V3
18S_1F / 18S_400R


Logistic regression
Probability same species
Percent identity

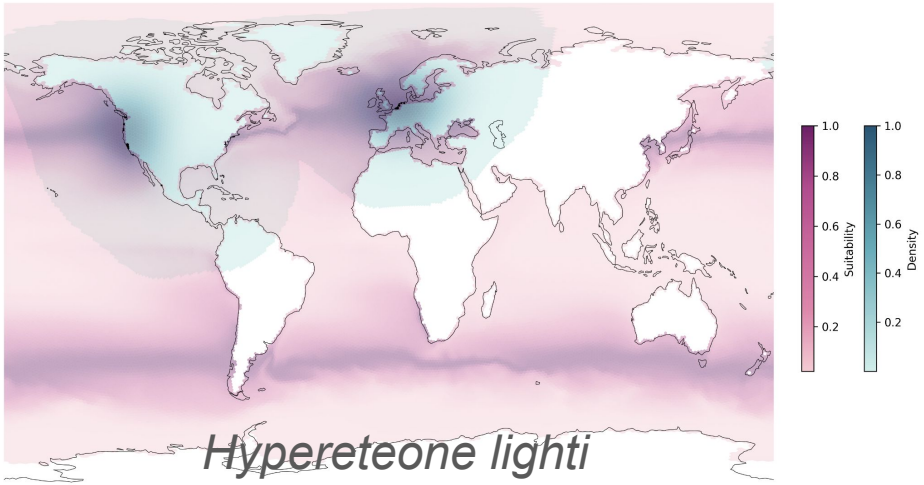

Logistic regression
Probability same species
Percent identity


eDNA QC
localhost:3000/?dataset=wadden
Medakamo hakoo 1634282 Chlorophyta Map 0.008 0
Trebouxia aggregata 615536 Chlorophyta Map 0.008 0
Pleurosigma inscriptura 708204 Heterokontophyta Map 0.008 2
Pleurosigma inscriptura 708204 Heterokontophyta Map 0.008 2
Neovahlkampfia damariscottae 1734589 Percolozoa Map 0.01 3
Chroomonas placoidea 573829 Cryptophyta Map 0.012 4
Polydora onagawaensis 1611501 Annelida Map 0.094 0.254 0.955 6
Hypereteone heteropoda 333652 Annelida Map 0.16 0.057 0.82 89
Coordinates: 5.00016, 52.97320
Sequence 1 200 bp
ATTGTCTAGGAATATTTCACATGCTGGCCCTTCTGTAGATTTGGCTATTTTTTCTTTACAT
TTAGCTGGTGTTTCATCTATTTTGGCCTCAATTAATTTTATTACTACGGCTATAAATATGC
GTTCAGTAGGAATGCGATTGGAGCGAGTTCCTCTTTTTGTTTGATCAGTTGGAATTACCGC
TCTTTTATTATTATTAT
Sequence 2 313 bp
ATTGTCTAGGAATATTTCACATGCTGGCCCTTCTGTAGATTTGGCTATTTTTTCTTTACAT
TTAGCTGGTGTTTCATCTATTTTGGCCTCAATTAATTTTATTACTACGGCTATAAATATGC
GTTCAGTAGGAATGCGATTGGAGCGAGTTCCTCTTTTTGTTTGATCAGTTGGAATTACCGC
TCTTTTATTATTATTATCTTTACCTGTTTTAGCTGGTGCAATTACTATACTTTTAACAGAT
CGTAATTTAAATACTTCGTTTTTTGATCCAGCAGGTGGAGGGGATCCTATTTTATATCAAC
ACTTGTTT
Coordinates: 8.80931, 54.48879
Sequence 1 200 bp
ATTGTCTAGGAATATTTCACATGCTGGCCCTTCTGTAGATTTGGCTATTTTTTCTTTACAT
TTAGCTGGTGTTTCATCTATTTTGGCCTCAATTAATTTTATTACTACGGCTATAAATATGC
GTTCAGTAGGAATGCGATTGGAGCGAGTTCCTCTTTTTGTTTGATCAGTTGGAATTACCGC
TCTTTTATTATTATTAT
Sequence 2 313 bp
ATTGTCTAGGAATATTTCACATGCTGGCCCTTCTGTAGATTTGGCTATTTTTTCTTTACAT
TTAGCTGGTGTTTCATCTATTTTGGCCTCAATTAATTTTATTACTACGGCTATAAATATGC
GTTCAGTAGGAATGCGATTGGAGCGAGTTCCTCTTTTTGTTTGATCAGTTGGAATTACCGC
TCTTTTATTATTATTATCTTTACCTGTTTTAGCTGGTGCAATTACTATACTTTTAACAGAT
CGTAATTTAAATACTTCGTTTTTTGATCCAGCAGGTGGAGGGGATCCTATTTTATATCAAC
ACTTGTTT


eDNA QC
localhost:3000/?dataset=wadden
eDNA QC results: Wadden COI
Search by scientific na
g 220 of 220 records
Scientific Name
lity
Cells
Lepinotus patruelis
1
Tenebrio molitor
1
Limnomonas gaiensis
0
Medakamo hakoo
0
Medakamo hakoo
0
Trebouxia aggregata
0
Pleurosigma inscriptura
2
Pleurosigma inscriptura
2
Neovahlkampfia damar
3
Chroomonas placoidea
4
Polydora onagawaensis
6
Hypereteone heteropo
89
Coordinates: 5.00016, 52.97320
Congenerics analysis: Hypereteone heteropoda
TaxonID: 333652 | Coordinates: Map
ATTGTCTAGGAATATTTCACATGCTGGCCCTTCTGTAGATTTGGCTATTTTTTCTTTACATTTAGCTGGTGTTTCATCTATTTTGGCCTCAATTAATTTTATTACTACGGCTATAAATATGCGTTCAGT
AGGAATGCGATTGGAGCGAGTTCCTCTTTTTGTTTGATCAGTTGGAATTACCGCTCTTTTATTATTATTATCTTTACCTGTTTTAGCTGGTGCAATTACTATACTTTTAACAGATCGTAATTTAAATAC
TTCGTTTTTTGATCCAGCAGGTGGAGGGGATCCTATTTTATATCAACACTTGTTT
Scientific Name
Taxon ID
pident
Score
Density
Identity
Suitability
Reference DB
Cells
Hypereteone lighti
333653
99.4
0.454
0.997
0.784
0.964
Yes
16
Hypereteone heteropoda
333652
99.4
0.131
0.057
0.784
0.856
Yes
89
Hypereteone alba
333649
0.009
-
7
Hypereteone barantollae
333650
0.005
-
3
Hypereteone tingara
328483
0.005
-
3
Hypereteone aestuarina
333648
0.004
-
1
Hypereteone otati
328482
0.004
-
2
Hypereteone foliosa
152250
0.933
0.986
Yes
192
Close



Hypereteone heteropoda



Hypereteone lighti

Image by Dean Janiak

Dutch long term monitoring of macrobenthos in the Dutch Continental Economical Zone of the North Sea

EurOBIS | MeasurementOrFact | Open in mapper

Overview | Data quality | Measurement types

Long term measurements of benthos composition in the Dutch national part of the North Sea. Samples have been taken by benthic sledge, box-corer, video and Hamon sampler. Samples were taken annually between 1991-2010, in 2012 and in 2015 and analysed to the genus or species level. For video sampling, coverage percentages were measured. The dataset has been used for assessment of the marine environment and habitat quality in the scope of the OSPAR Quality Status Reports and the Marine Strategy Framework Directive.

Citation: *Ministery of Infrastructure and Environment, The Netherlands; (2018): Dutch long term monitoring of macrobenthos in the Dutch Continental Economical Zone of the North Sea.*

Published: September 17, 2025 at 09:52

URL: http://ipt.vliz.be/eurobis/resource?r=deltaresbenthos

Ministerie van Infrastructuur en Waterstaat; Rijkswaterstaat
Ministerie van Infrastructuur en Waterstaat; Rijkswaterstaat

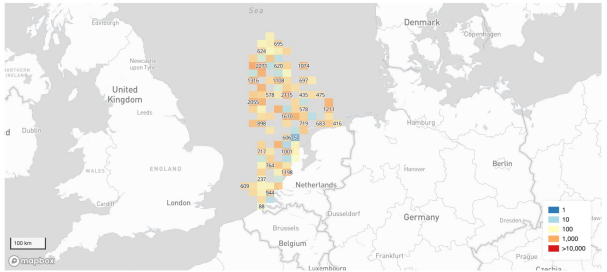
Willem Stolte
Deltares

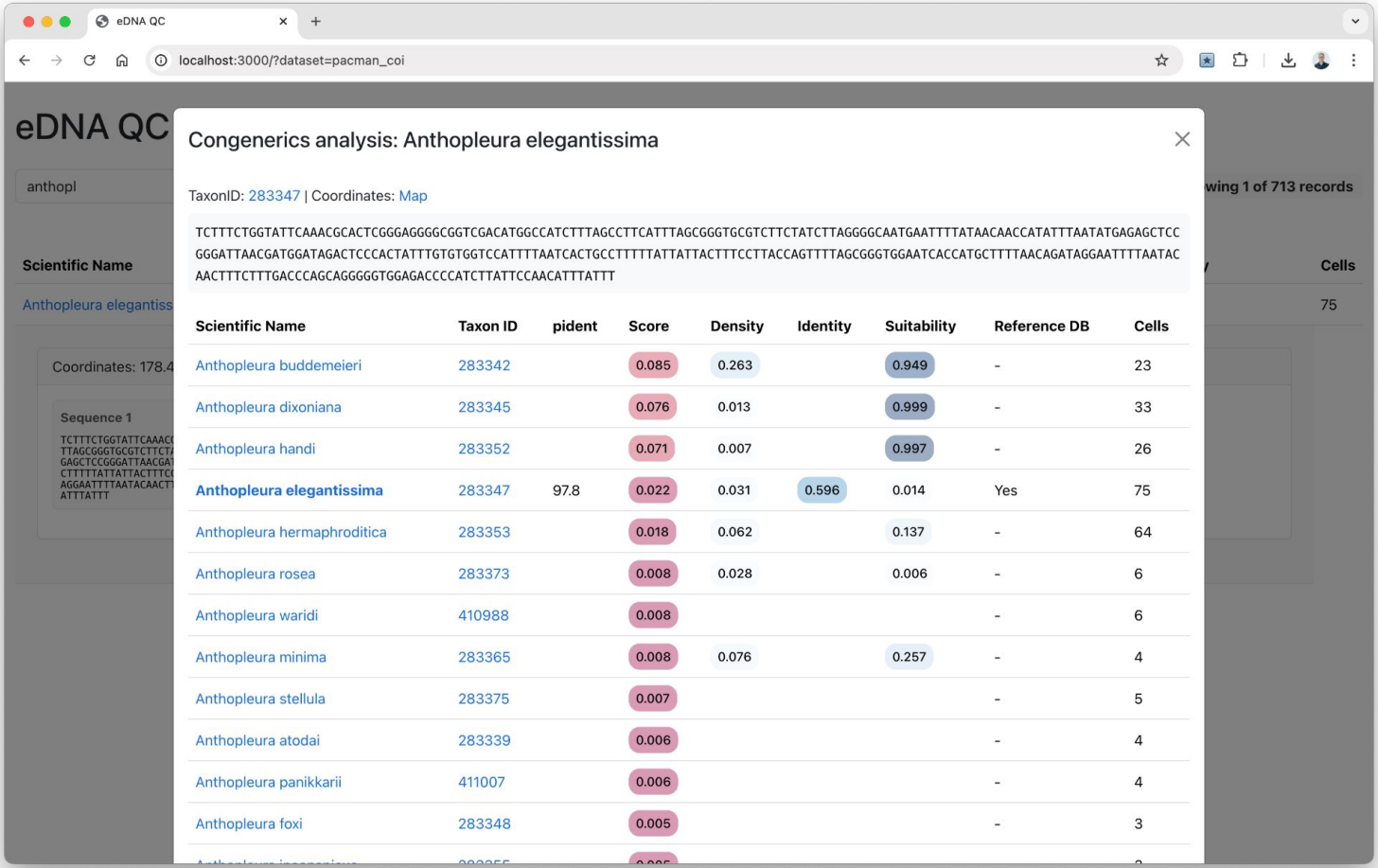
58,386
occurrence records

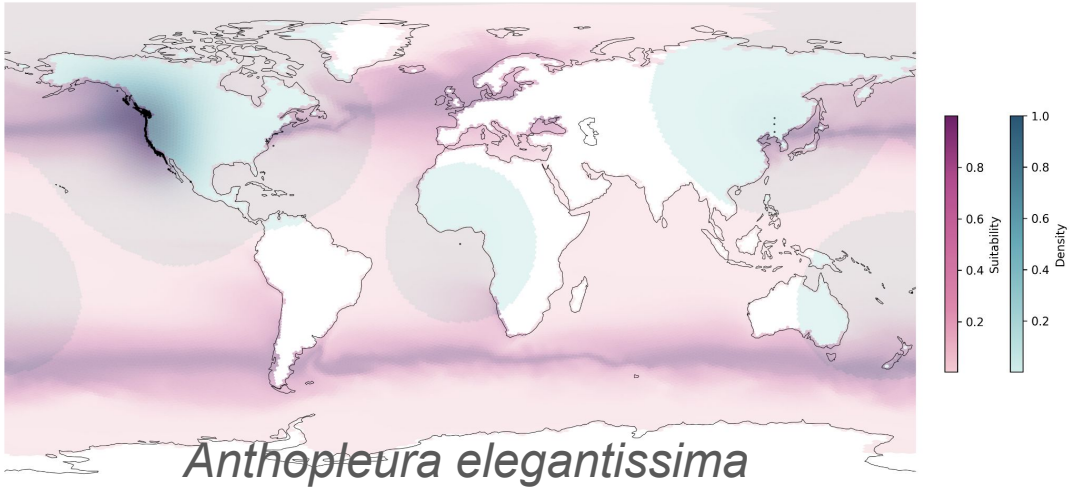
136,934
measurements and facts

811
taxa

582
species

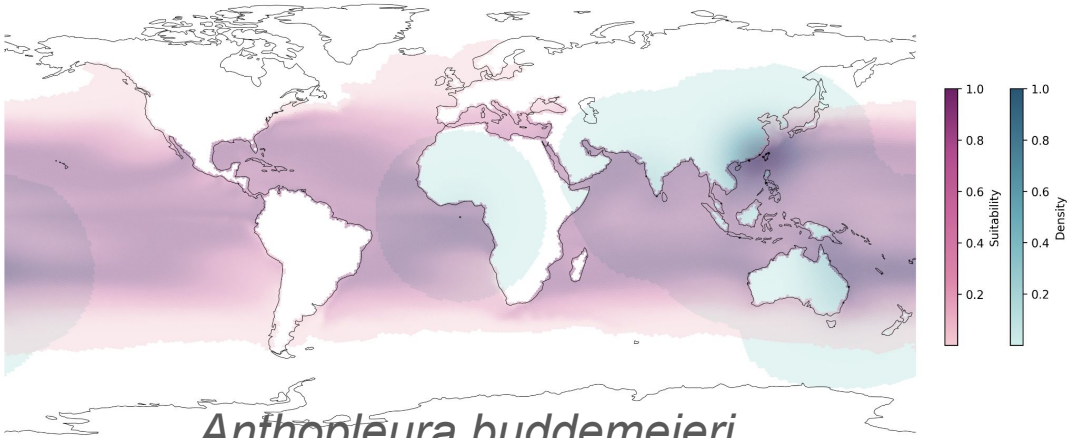



eDNA QC
localhost:3000/?dataset=pacman_coi
eDNA QC
anthopl
wing 1 of 713 records
Scientific Name
Cells
Anthopleura elegantiss
75
Coordinates: 178.4
Sequence 1
Congenerics analysis: Anthopleura elegantissima
TaxonID: 283347 | Coordinates: Map
TCTTTCTGGTATTCAAACGCACTCGGGAGGGGCGGTCGACATGGCCATCTTTAGCCTTCATTTAGCGGGTGCGTCTTCTATCTTAGGGGCAATGAATTTTATAACAACCATATTTAATATGAGAGCTCC
GGGATTAACGATGGATAGACTCCCACTATTTGTGTGGTCCATTTTAATCACTGCCTTTTTATTATTACTTTCCTTACCAGTTTTAGCGGGTGGAATCACCATGCTTTTAACAGATAGGAATTTTAATAC
AACTTTCTTTGACCCAGCAGGGGGTGGAGACCCCATCTTATTCCAACATTTATTT
Scientific Name | Taxon ID | pident | Score | Density | Identity | Suitability | Reference DB | Cells
Anthopleura buddemeieri | 283342 | | 0.085 | 0.263 | | 0.949 | - | 23
Anthopleura dixoniana | 283345 | | 0.076 | 0.013 | | 0.999 | - | 33
Anthopleura handi | 283352 | | 0.071 | 0.007 | | 0.997 | - | 26
Anthopleura elegantissima | 283347 | 97.8 | 0.022 | 0.031 | 0.596 | 0.014 | Yes | 75
Anthopleura hermaphroditica | 283353 | | 0.018 | 0.062 | | 0.137 | - | 64
Anthopleura rosea | 283373 | | 0.008 | 0.028 | | 0.006 | - | 6
Anthopleura waridi | 410988 | | 0.008 | | | | - | 6
Anthopleura minima | 283365 | | 0.008 | 0.076 | | 0.257 | - | 4
Anthopleura stellula | 283375 | | 0.007 | | | | - | 5
Anthopleura atodai | 283339 | | 0.006 | | | | - | 4
Anthopleura panikkarii | 411007 | | 0.006 | | | | - | 4
Anthopleura foxi | 283348 | | 0.005 | | | | - | 3



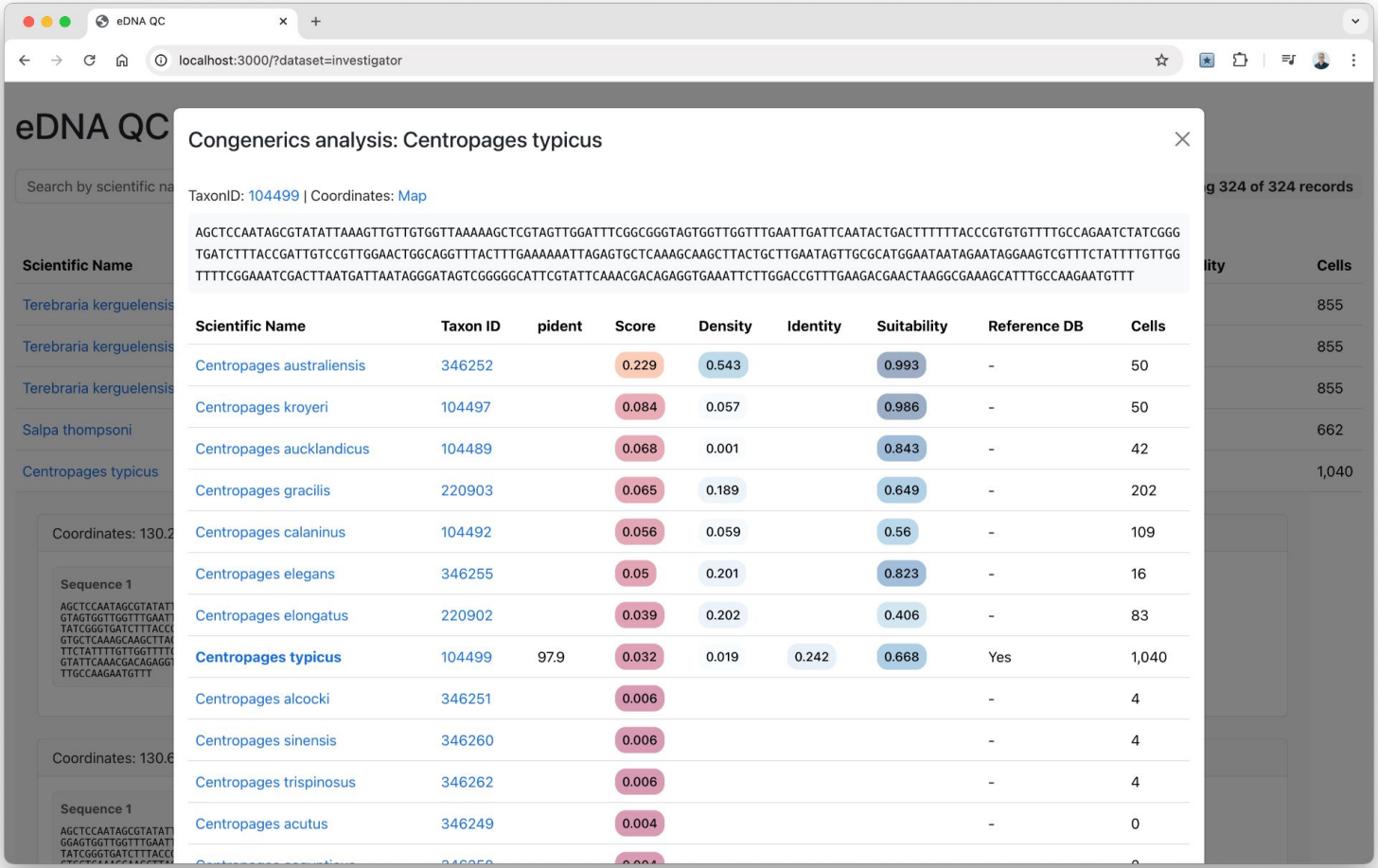
Anthopleura elegantissima

Image by Heather Fulton-Bennett



Anthopleura buddemeieri

Image by Langzi

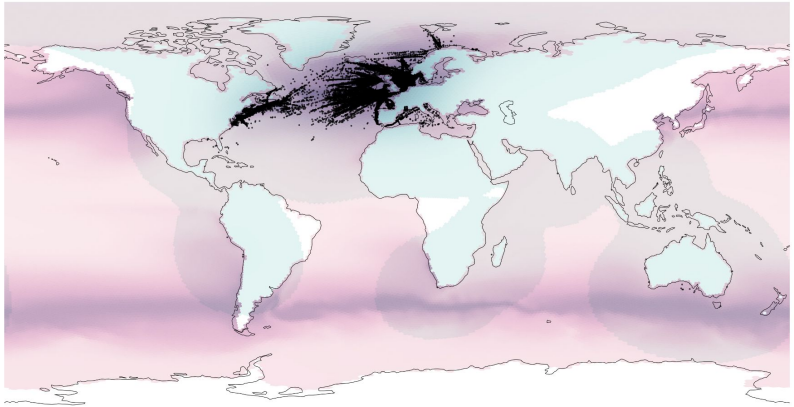


Congenerics analysis: Centropages typicus

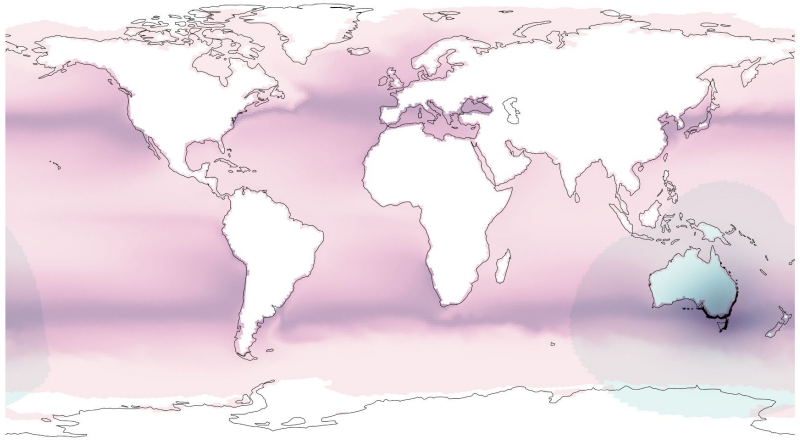
TaxonID: 104499 | Coordinates: Map

```
AGCTCCAATAGCGTATATTAAAGTTGTTGTGGTTAAAAAGCTCGTAGTTGGATTTCGGCGGGTAGTGGTTGGTTTGAATTGATTCAATACTGACTTTTTTACCCGTGTGTTTTGCCAGAATCTATCGGG
TGATCTTTACCGATTGTCCGTTGGAACTGGCAGGTTTACTTTGAAAAAATTAGAGTGCTCAAAGCAAGCTTACTGCTTGAATAGTTGCGCATGGAATAATAGAATAGGAAGTCGTTTCTATTTTGTTGG
TTTTCGGAAATCGACTTAATGATTAATAGGGATAGTCGGGGGCATTCGTATTCAAACGACAGAGGTGAAATTCTTGGACCGTTTGAAGACGAACTAAGGCGAAAGCATTTGCCAAGAATGTTT
```

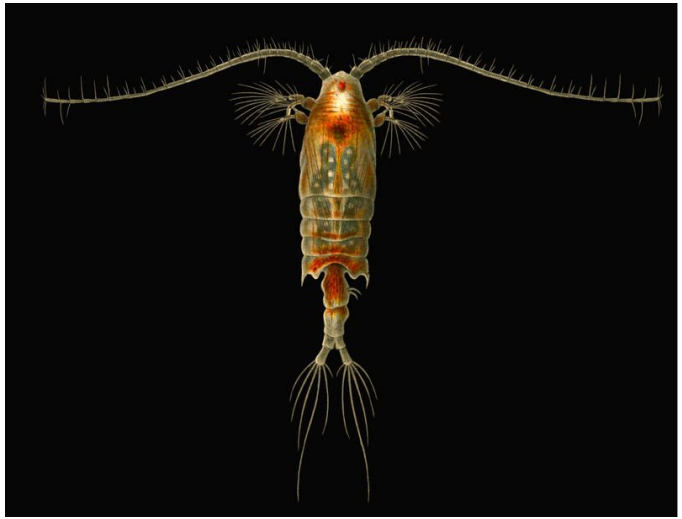
Scientific Name	Taxon ID	pident	Score	Density	Identity	Suitability	Reference DB	Cells
Centropages australiensis	346252		0.229	0.543		0.993	-	50
Centropages kroyeri	104497		0.084	0.057		0.986	-	50
Centropages aucklandicus	104489		0.068	0.001		0.843	-	42
Centropages gracilis	220903		0.065	0.189		0.649	-	202
Centropages calaninus	104492		0.056	0.059		0.56	-	109
Centropages elegans	346255		0.05	0.201		0.823	-	16
Centropages elongatus	220902		0.039	0.202		0.406	-	83
Centropages typicus	104499	97.9	0.032	0.019	0.242	0.668	Yes	1,040
Centropages alcocki	346251		0.006				-	4
Centropages sinensis	346260		0.006				-	4
Centropages trispinosus	346262		0.006				-	4
Centropages acutus	346249		0.004				-	0



Centropages typicus

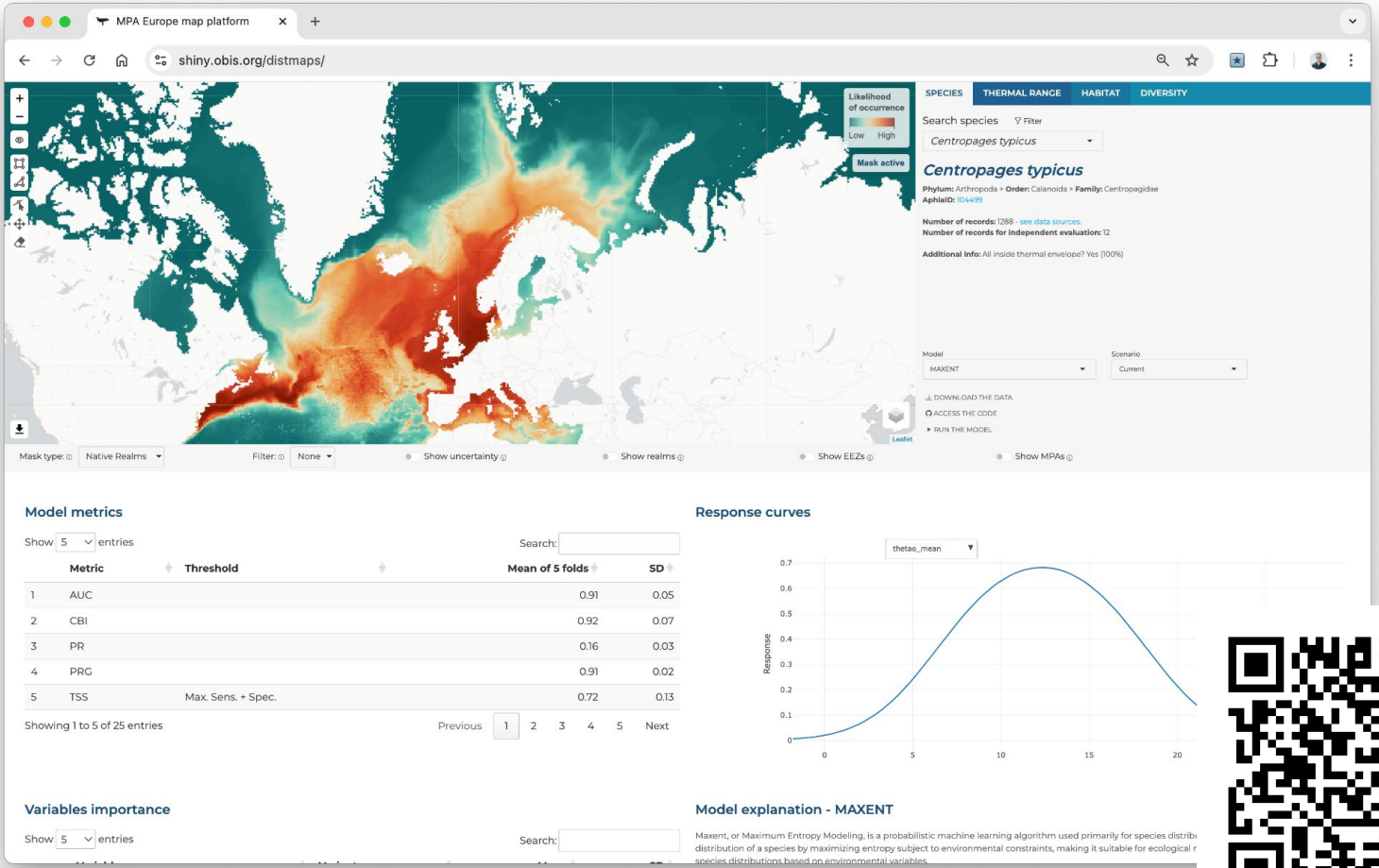


Centropages australiensis



What's next?

- Make the tooling available for data providers and node managers
- Improve the algorithm and integrate more data sources (more sophisticated models, curated species distributions)
- Dataset tagging and filtering
- Data standards improvements and guidelines
- Integration into QC procedures
- Provide mechanisms for annotating records


MPA Europe map platform
shiny.obis.org/distmaps/
Likelihood of occurrence
Low High
Mask active
SPECIES
THERMAL RANGE
HABITAT
DIVERSITY
Search species
Filter
Centropages typicus
Centropages typicus
Phylum: Arthropoda > Order: Calanoida > Family: Centropagidae
AphiaID: 104499
Number of records: 1288 - see data sources.
Number of records for independent evaluation: 12
Additional info: All inside thermal envelope? Yes (100%)
Model
MAXENT
Scenario
Current
DOWNLOAD THE DATA
ACCESS THE CODE
RUN THE MODEL
Leaflet
Mask type:
Native Realms
Filter:
None
Show uncertainty
Show realms
Show EEZs
Show MPAs
Model metrics
Show 5 entries
Search:
Metric
Threshold
Mean of 5 folds
SD
1 AUC 0.91 0.05
2 CBI 0.92 0.07
3 PR 0.16 0.03
4 PRG 0.91 0.02
5 TSS Max. Sens. + Spec. 0.72 0.13
Showing 1 to 5 of 25 entries
Previous 1 2 3 4 5 Next
Response curves
thetao_mean
Response
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 5 10 15 20
Variables importance
Show 5 entries
Search:
Model explanation - MAXENT
bitly

What's next?

- Make the tooling available for data providers and node managers
- Improve the algorithm and integrate more data sources (more sophisticated models, curated species distributions)
- Dataset tagging and filtering
- Data standards improvements and guidelines
- Integration into QC procedures
- Provide mechanisms for annotating records

Search

Search for: Dataset

Search term: https://rs.obis.org/obis/vocabulary/Data

Search

Dataset search syntax: + signifies AND operation, | signifies OR operation, - negates a single token, " wraps a number of tokens to signify a phrase for searching, * at the end of a term signifies a prefix query, (and) signify precedence.

Marine Microbes from RV Investigator voyage IN2016_V03, 170W, Pacific Ocean (2016) 4,583,637 records
https://www.marine.csiro.au/ipt/resource?r=bioplatforms_mm_in2016_v03

Marine metagenomes Metagenome 4,582,446 records
https://hosted-datasets.gbif.org/mgnify/MGYS00003194.zip

Marine Microbes from the North Stradbroke Island National Reference Station (NRS), Queensland, Australia (2012-2020) 3,032,979 records
https://www.marine.csiro.au/ipt/resource?r=bioplatforms_mm_nrs_nsi

Marine Microbes from the Port Hacking National Reference Station (NRS), New South Wales, Australia (2012-2020) 3,071,809 records
https://www.marine.csiro.au/ipt/resource?r=bioplatforms_mm_nrs_phb

Marine Microbes from the Maria Island National Reference Station (NRS), Tasmania, Australia (2012-2021) 2,797,790 records
https://www.marine.csiro.au/ipt/resource?r=bioplatforms_mm_nrs_mai

Marine Microbes from RV Investigator voyage IN2016_V04, Australia (2016) 2,539,104 records
https://www.marine.csiro.au/ipt/resource?r=bioplatforms_mm_in2016_v04

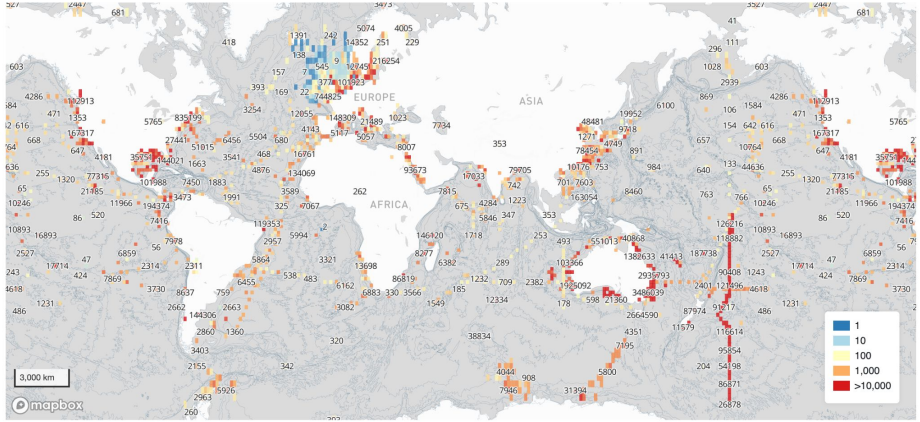
Marine Microbes from the Rottnest Island National Reference Station (NRS), Western Australia, Australia (2015-2020) 1,786,909 records
https://www.marine.csiro.au/ipt/resource?r=bioplatforms_mm_nrs_rot

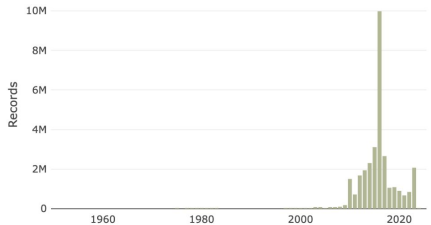
Marine Microbes from the Yongala National Reference Station (NRS), Queensland, Australia (2015-2020) 1,247,306 records
https://www.marine.csiro.au/ipt/resource?r=bioplatforms_mm_nrs_yon

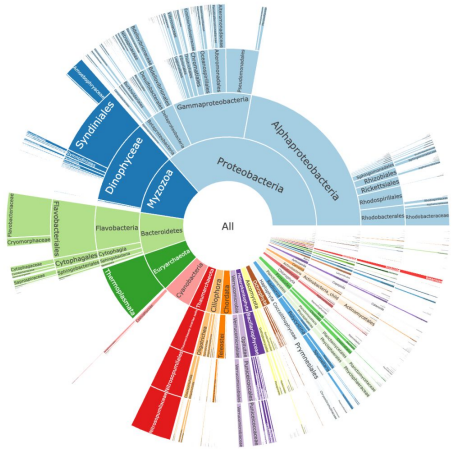
Data associated with 'A 17-year time-series of fungal environmental DNA from a coastal marine ecosystem reveals long-term seasonal-scale and inter-annual diversity patterns' 1,068,660 records
https://www.dassh.ac.uk/ipt/resource?r=l4-edna

Nahant Collection 1,261,908 records
https://hosted-datasets.gbif.org/mgnify/MGYS00002376.zip

Previous Next Showing 1-10 of 409 results

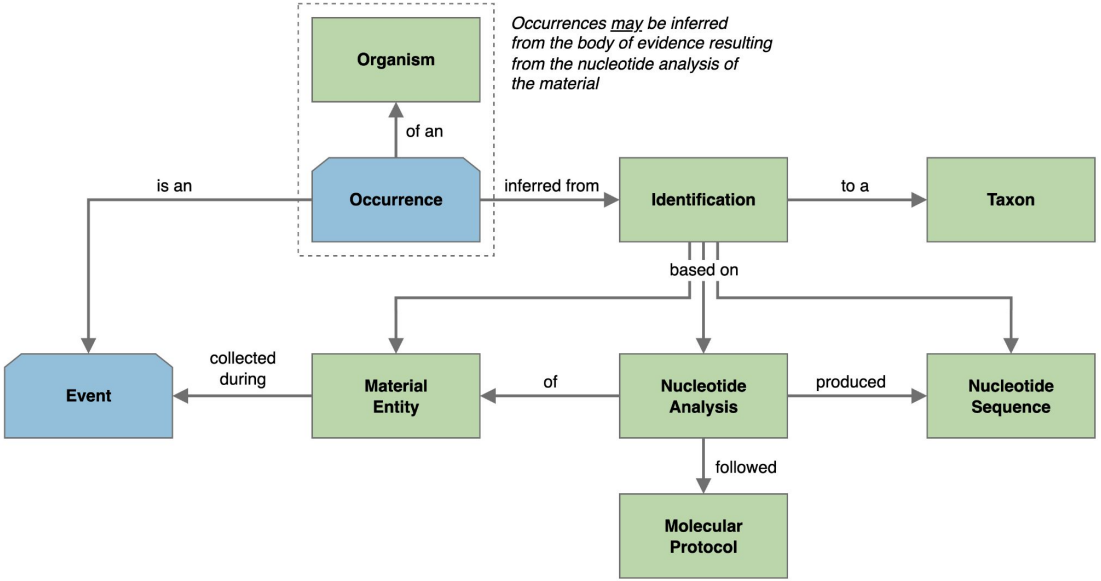






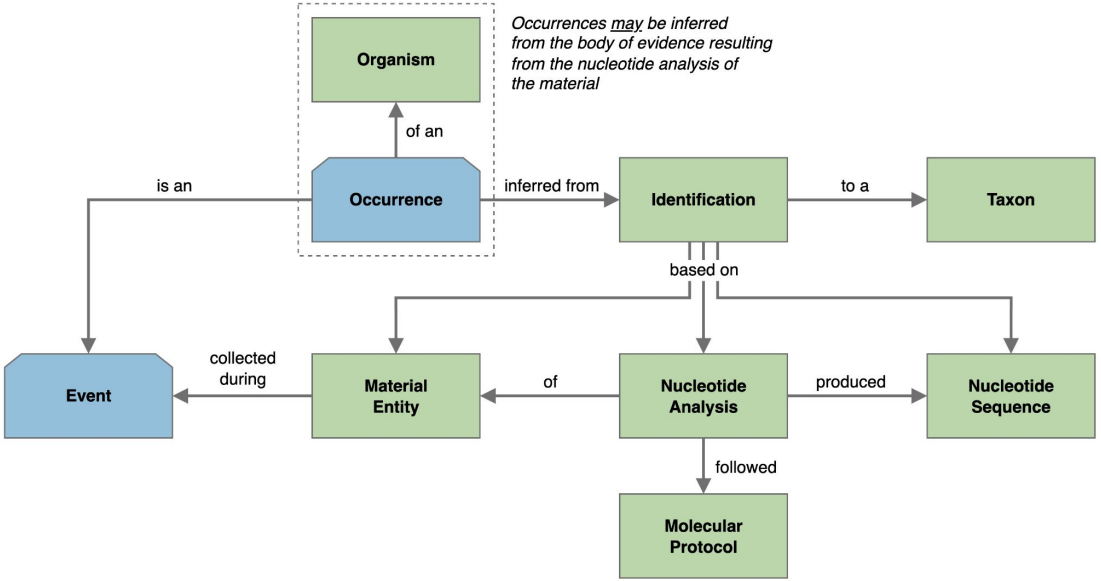
What's next?

- Make the tooling available for data providers and node managers
- Improve the algorithm and integrate more data sources (more sophisticated models, curated species distributions)
- Dataset tagging and filtering
- Data standards improvements and guidelines
- Integration into QC procedures
- Provide mechanisms for annotating records



https://gbif.github.io/dwc-dp/cm/

taxonID	nucleotideAnalysisID	taxonAssignmentMethod	identificationVerificationStatus	isAcceptedIdentification
genus_A	provider_analysis_X	pipeline_ncbi_rdp_08	accepted	true
species_B	provider_analysis_X	pipeline_silva_blast_top	unconfirmed	false
species_C	provider_analysis_X	pipeline_silva_blast_lca	unconfirmed	false



https://gbif.github.io/dwc-dp/cm/

taxonID	nucleotideAnalysisID	taxonAssignmentMethod	identificationVerificationStatus	isAcceptedIdentification
genus_A	provider_analysis_X	pipeline_ncbi_rdp_08	~~accepted~~ unconfirmed	~~true~~ false
species_B	provider_analysis_X	pipeline_silva_blast_top	unconfirmed	false
species_C	provider_analysis_X	pipeline_silva_blast_lca	unconfirmed	false
family _D	platform_qc_Y	pipeline_bold_rdp_09	accepted	true

Thank you!

PacMAN
Pacific Islands Marine Bioinvasions Alert Network

UK Research and Innovation

Flanders
State of the Art

Co-funded by the European Union

MPA
MARINE PROTECTED AREAS EUROPE

RICHARD LOUNSBERY FOUNDATION