The need for accurate and comprehensive DNA sequence databases to reliably identify species of policy concern

(Kenny Meganck and Sophie Gombeer, BopCo)

BopCo

LifeWatch BELGIUM

A Barcoding Facility for Organisms and Tissues of Policy Concern

SOPHIE GOMBEER, KENNY MEGANCK, NATHALIE SMITZ, ANN VANDERHEYDEN, THIERRY BACKELJAU & MARC DE MEYER

BopCo jointly run

by RBINS and RMCA

Federal Contribution

to the ERIC LifeWatch

Species identifications

Morphological characteristics

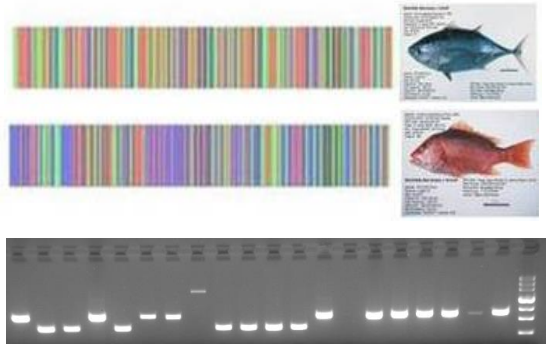
Monographs, identification keys, scientific periodic

Microscopy

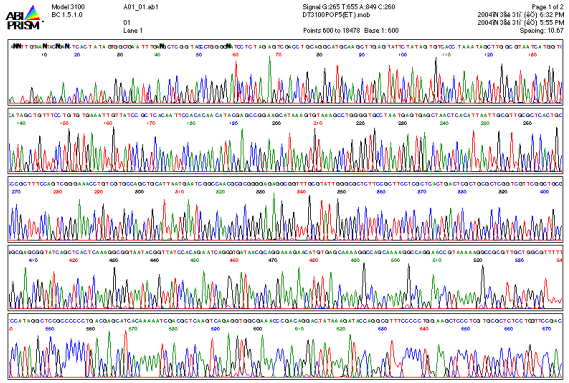
Network of in-house and external taxonomic experts

Specimen collections

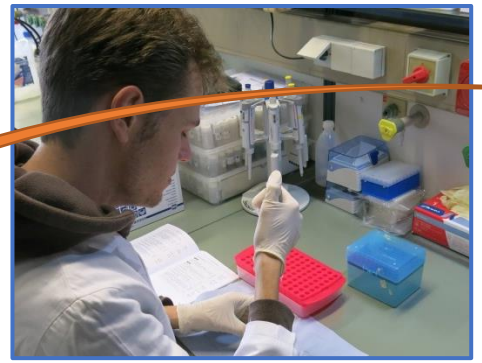
DNA-based techniques



Access to laboratory facilities and sequence databases

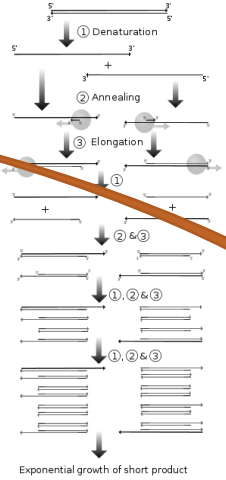


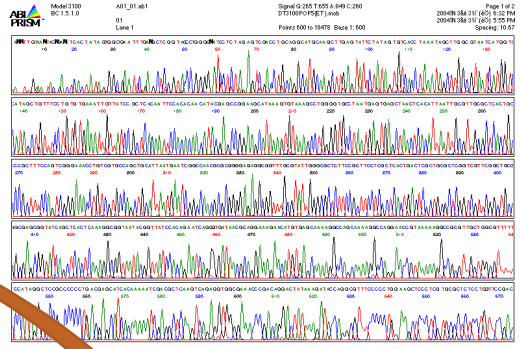
DNA barcoding



DNA Extraction

Amplification

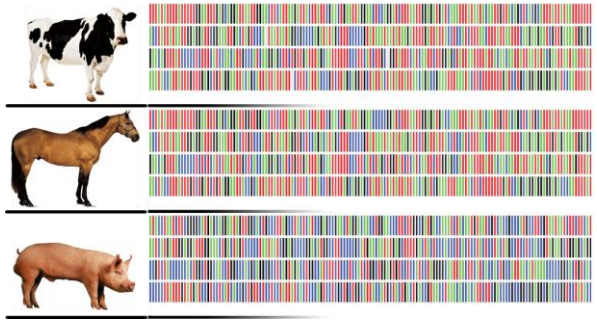




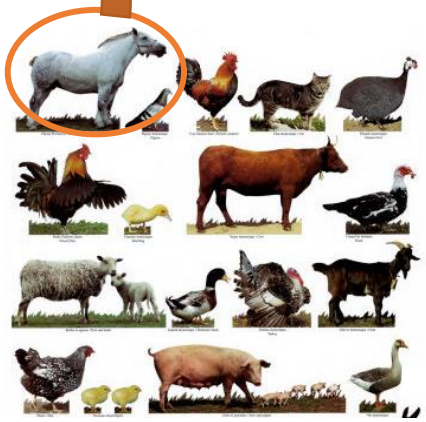
Sequencing

...TTGCCTTATGATCAGGAACC....

Comparison



Online DNA sequence databases



Species identification

Advantages

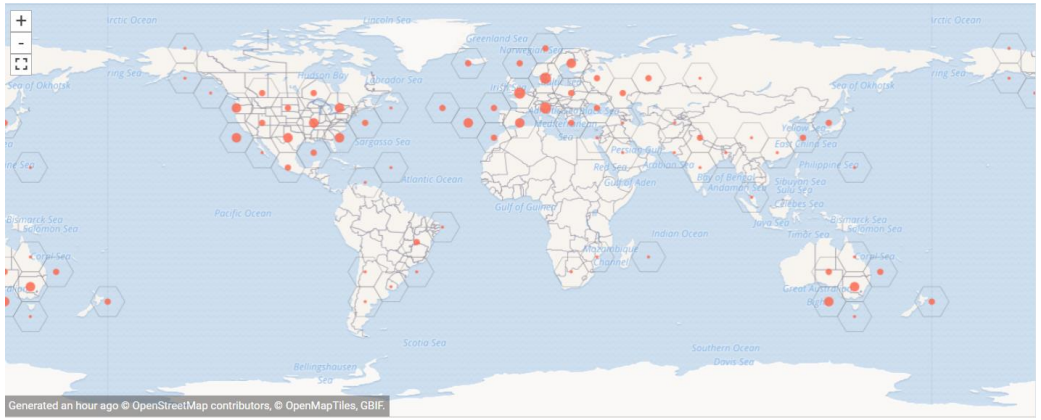
- All live stages
- No morphological characteristics
- Processed samples

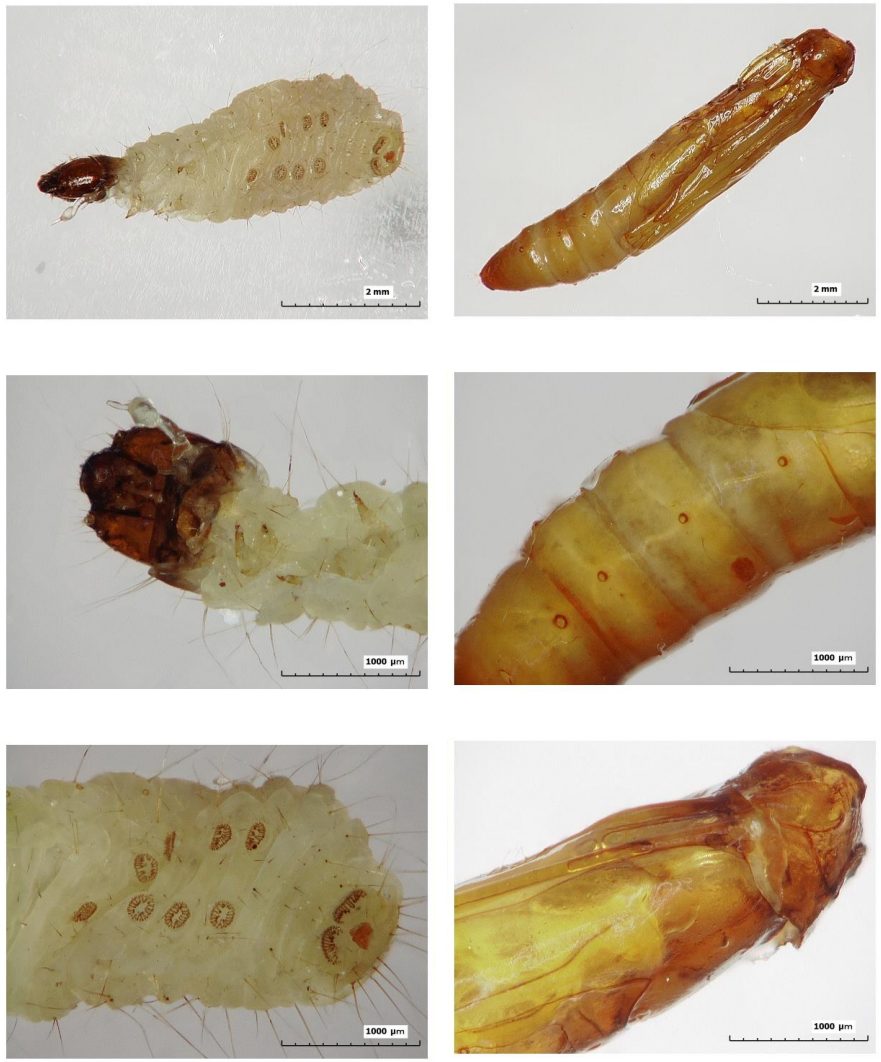
Advantages of DNA-based technologies

- All live stages, e.g. eggs, larvae, seeds, etc.

Identification request from an **International Pharmaceutical Company** to identify insect larva and insect pupa found in drum of chemical raw product which turned out to be *Plodia interpunctella*, a world-wide pest of stored products.

Indian meal moth, *Plodia interpunctella* By CSIRO [CC BY 3.0]





Advantages of DNA-based technologies

- No morphological characteristics, e.g. piece of tissue, cryptic species, etc.


Bird Strike By flightlog [CC BY 2.0]

Recurrent identification request from the **Belgian Air Force** and **Brussels Airport Company** to identify remains from birds that were involved in a bird strike. Several bird species have already been identified.

B-6543 By NMOS332 [CC BY-SA 2.0]

Advantages of DNA-based technologies

- Processed samples & derived products, e.g. cooked, ground, dried, smoked, etc.

Request from the **Agency for Nature and Forests (ANB)** to help with the identification of feaces to determine if they belong to a native species or the invasive muntjac deer, *Muntiacus reevesi*. DNA-based analyses confirmed the presence of the invasive species.

Deer feces By Krzysztof Maria Różański [CC BY-SA 3.0]

Muntjac Deer (*Muntiacus reevesi*) By Peter O'Connor aka anemoneprojectors [CC BY-SA 2.0]

Limitations

- DNA integrity
- Insufficient sequence divergence
- Reference database dependency

Limitations of DNA-based technologies

- DNA integrity might be affected by age and conservation method of the sample

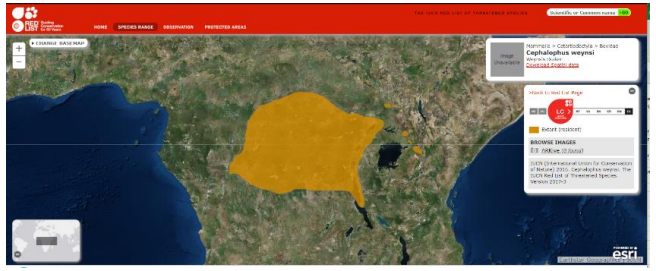
An **Antiques Dealer** requested the identification of a piece of animal skin used in an African ceremonial mask. A DNA-based identification, however, was not possible due to the low quality of the extracted DNA, which prevented all downstream analyses.

Limitations of DNA-based technologies

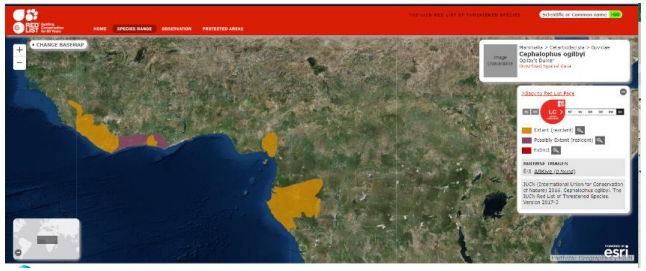
- Insufficient sequence divergence, e.g. recently diverged species

C. callipygus

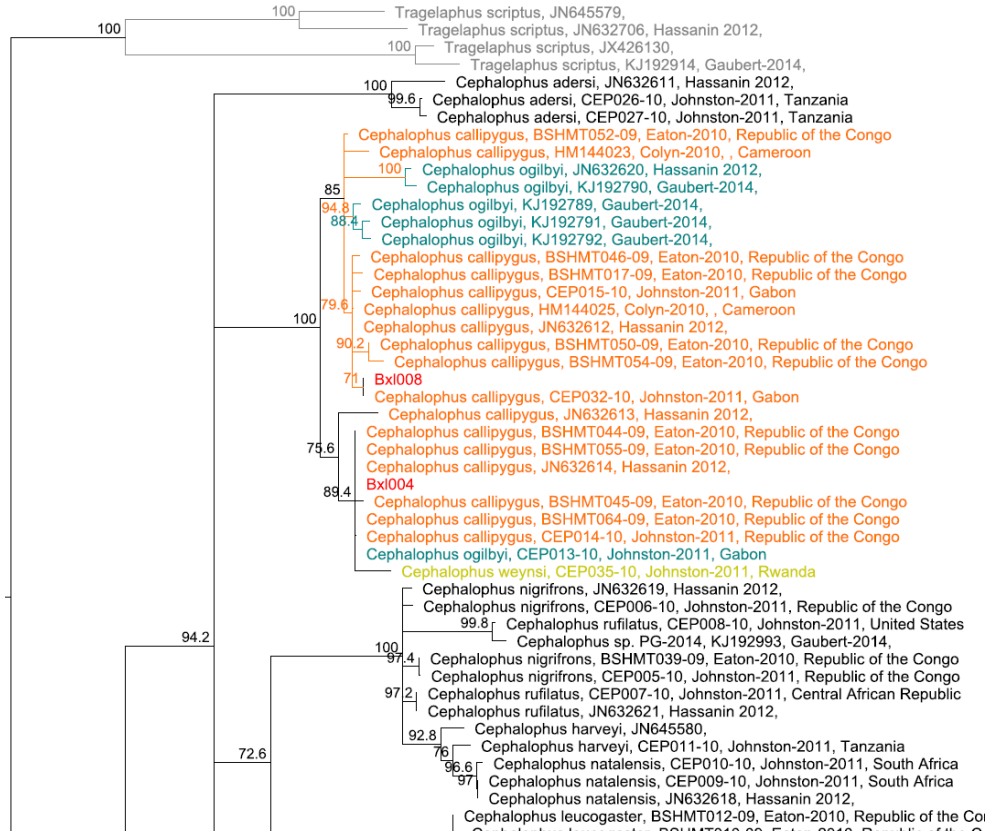
C. weynsi



C. ogilby



Within the framework of a **BopCo Research Project** investigating the bushmeat market in Brussels, two meat samples could not be identified to the species level, due to a recent divergence of three duiker species: *Cephalophus callipygus, C. weynsi* and *C. ogilbyi.* Only *C. ogilbyi* is CITES-listed.



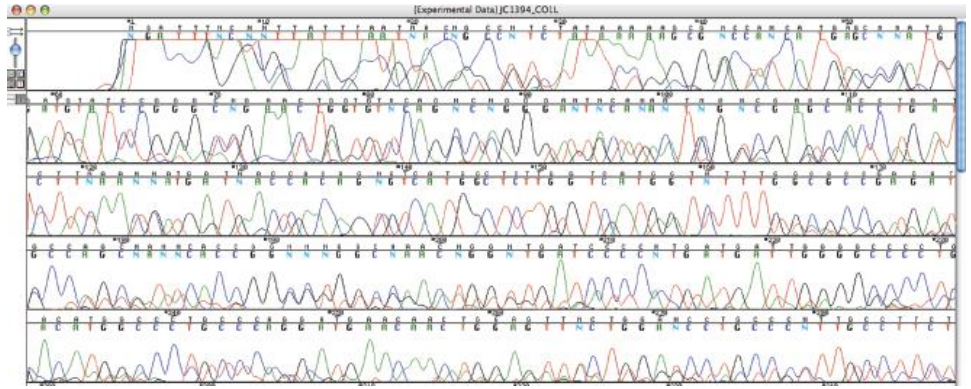
COI NJ-tree of Bovidae, zoomed in on *Cephalophus* cluster

Limitations of DNA-based technologies

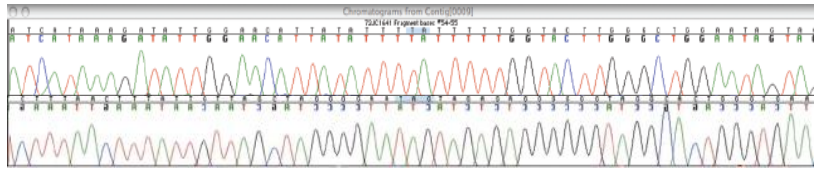
- Errors in the database lowering success rates and reliability of identifications

Technical errors: can often be detected by alignment with reference sequences and translation into amino acids.

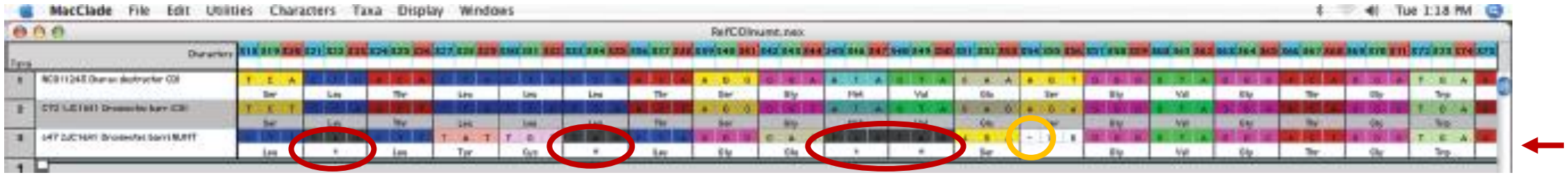
Bad quality sequences should be discarded.



Figures and examples from Buhay, 2009



Sequencing errors are more common at the edges of the sequenced fragments.



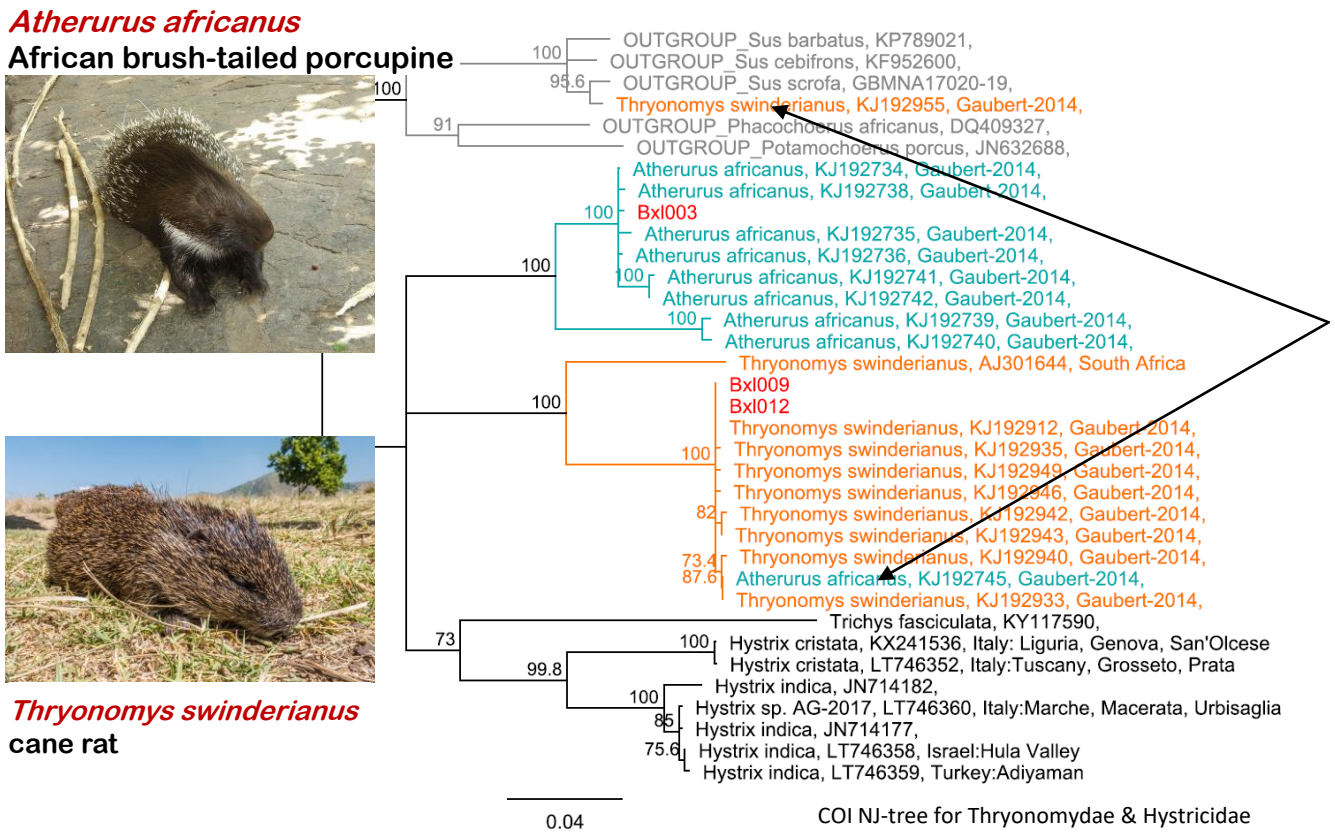
NUMTs or “nuclear mitochondrial DNA".

Limitations of DNA-based technologies

- Errors in the database lowering success rates and reliability of identifications

Wrongful species assignments: misidentification of specimens often occurs when species have overlapping distributions, or when the identifier has limited taxonomic expertise (e.g. specimens collected as outgroup). But mislabelling can also be due to errors during data uploading.

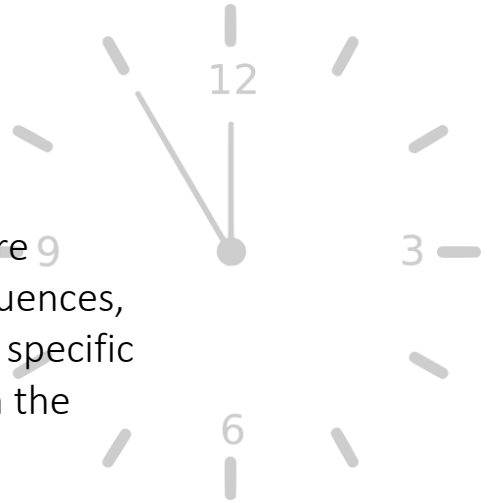
Clustering analyses and Neighbour-Joining trees useful tools for detecting errors due to mislabelling.



Potential misidentified sequences = species identifications were based on secondary morphological features of smoked bushmeat samples.

Limitations of DNA-based technologies

- Errors in the database lowering success rates and reliability of identifications



Time consuming: checking literature associated with the published sequences, details on the collection locations, specific details on specimens published on the sequence databases…..

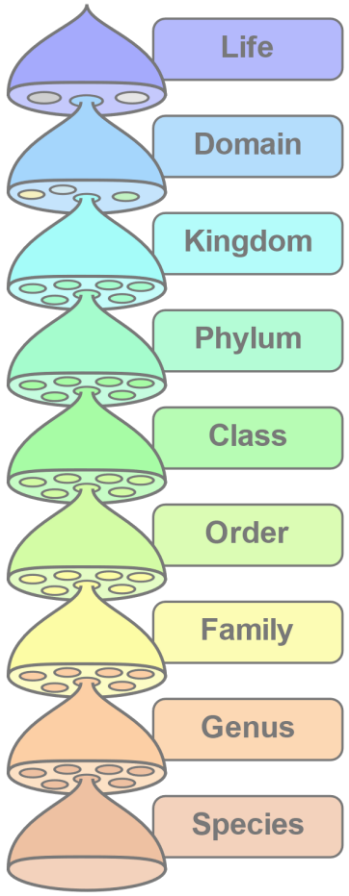
Case-by-case: each potential mislabelled sequence demands a tailored investigation depending on the available data

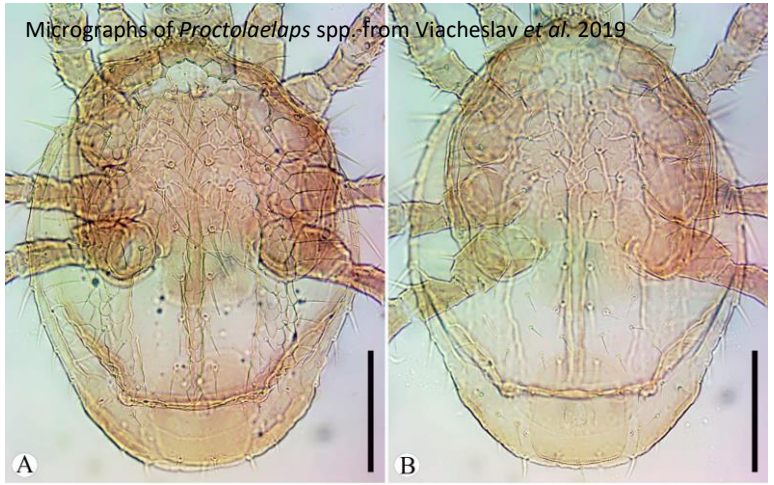
Requires extensive knowledge: interpretation of taxonomic revisions, geographical distribution data, species complexes, cryptic species, …..

Limitations of DNA-based technologies

- Incomplete databases: success of identification depends on available sequence data

Reference databases are **not complete**, and often reference sequences for species or even entire genera are missing from these online databases. In such cases, specimens can only be identified to a higher taxonomic level. This lack of available reference data prevented us from providing a species name and a geographical origin of the specimen.

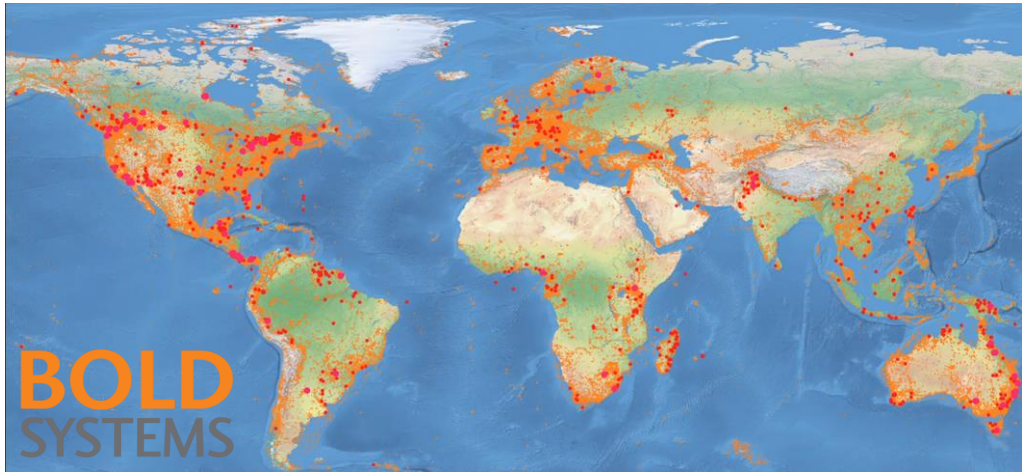




The identification of mite specimens that were found in a **shipment of cork** insulation products was only completed to the genus level, i.e. *Proctolaelaps*.

The identification of an insect larvae discovered in a **shipment of pharmaceutical products** was only completed to the subfamily level, i.e. Nematinae.

Sequence reference databases



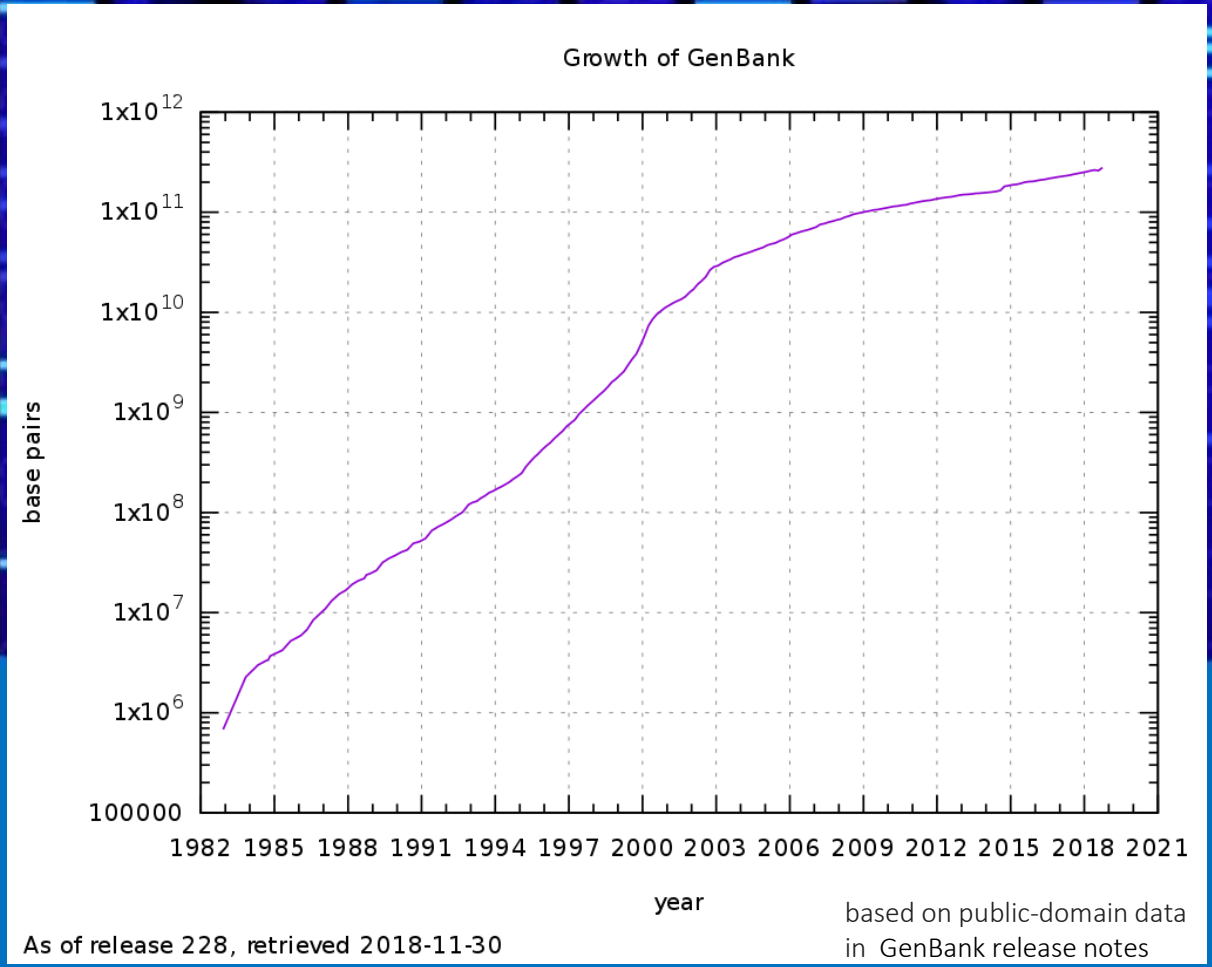
GenBank

- from 1980's
- USA National Institutes of Health

GenBank

GenBank

- Public data (open access)



GenBank

- Public data (open access)
- Different markers available

Sequence Data Submissions to GenBank via BankIt Can Include:

Single or multiple sequences of:

- Complete (or partial) sequences > 200 nt long
- Protein coding genes
- Ribosomal RNA genes (16S, 5S, …)
- Internal transcribed spacers (ITS)
- Microsatellite markers (but NOT sequence tagged sites, STS)
- Complete viral or phage genomes
- Complete mitochondrial genomes
- Complete chloroplast or other plastids genomes

GenBank

- Public data (open access)
- Different markers available
- Limited requirements for sequence uploads

https://www.ncbi.nlm.nih.gov/WebSub/html/requirements.html

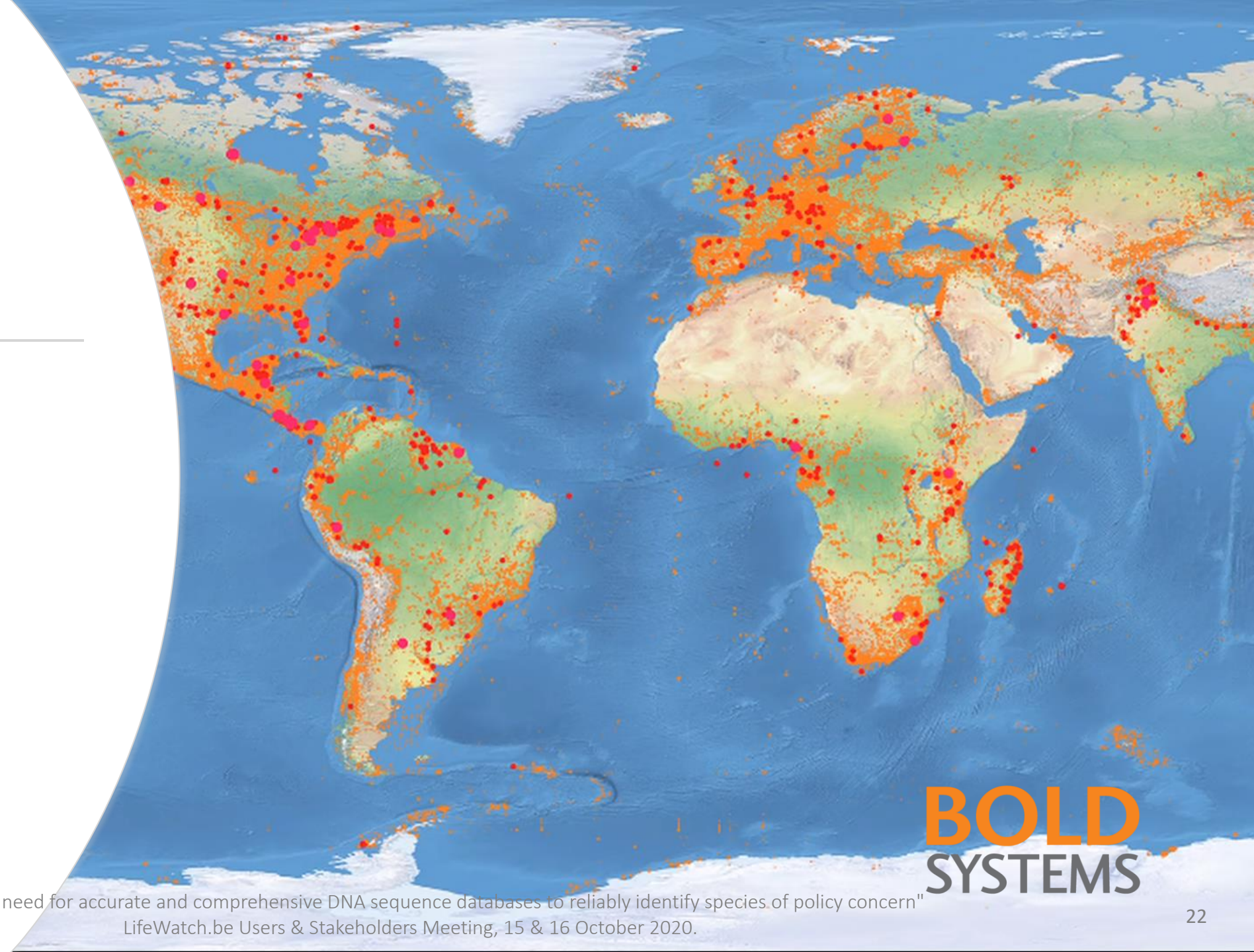
- No curation of data

GenBank is filled with
- direct submissions from individual laboratories and
- bulk submissions from large-scale sequencing centers.

There is no peer-review of these sequences, but a lot of reference material is generated in this way

BOLD

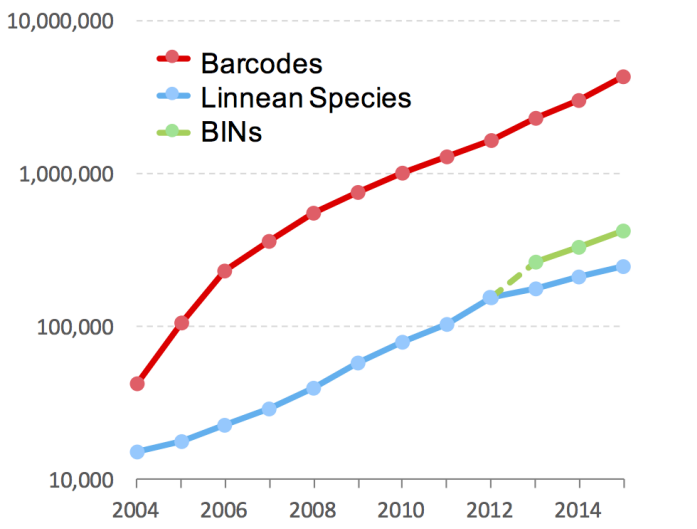
- from 2007
- Barcode Of Life Data System



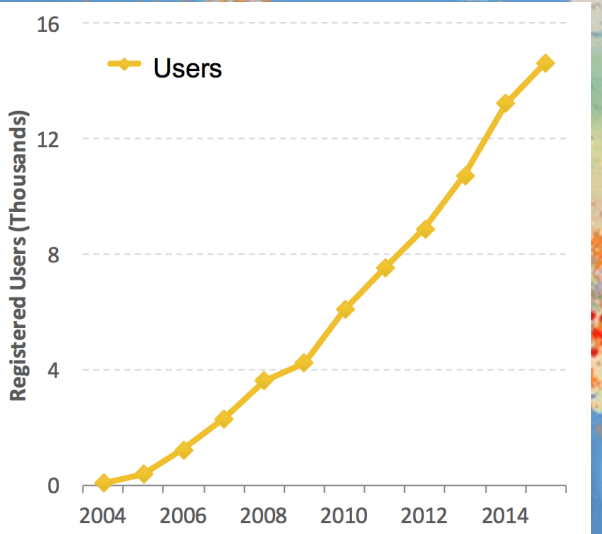
BOLD

- Barcodes

COI (animals),
matK, and rbcL (plants)
ITS (fungi)



Increase of number of barcodes deposited in BOLD (red)



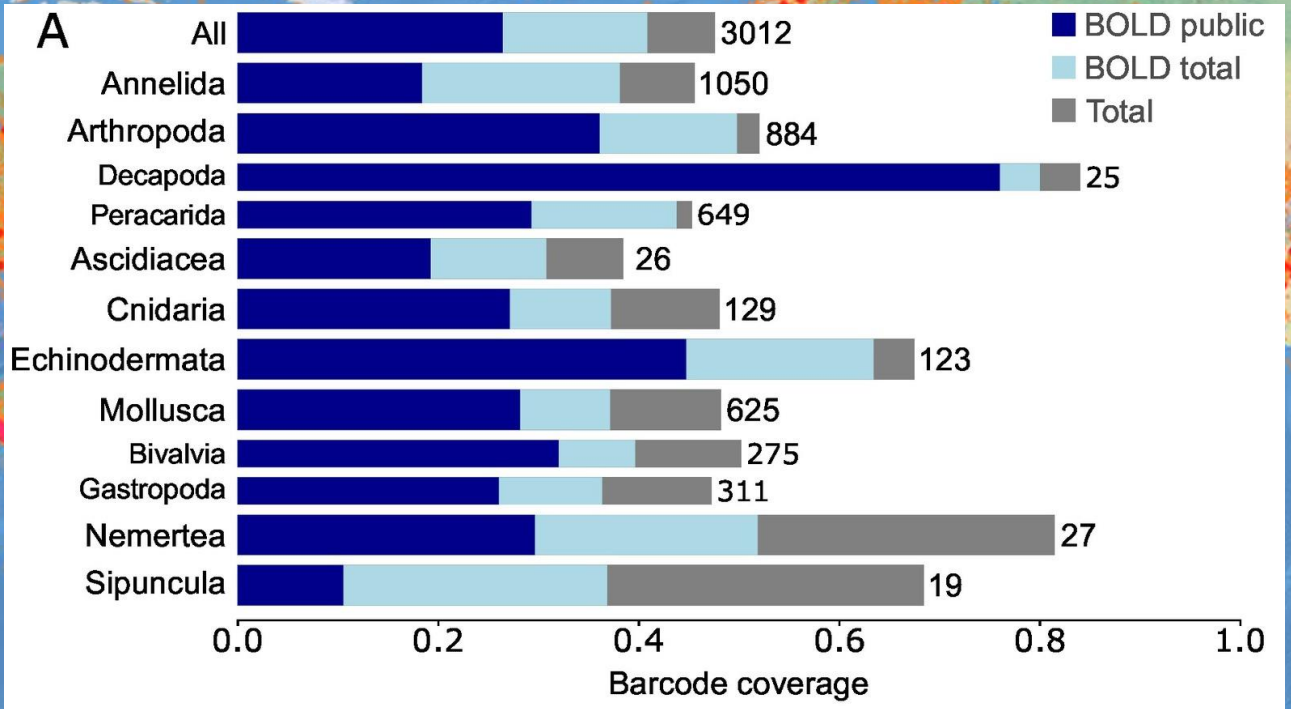
Increase of number of users of the BOLD platform

BOLD SYSTEMS

BOLD

- Barcodes
- With private data

still open-access, but more community-based



Cumulative barcode coverage for marine invertebrates. Barcode coverage of at least one reference sequence. Numbers on bars refer to total number of species in checklist. BOLD public library, the BOLD private library, and subsequently GenBank.

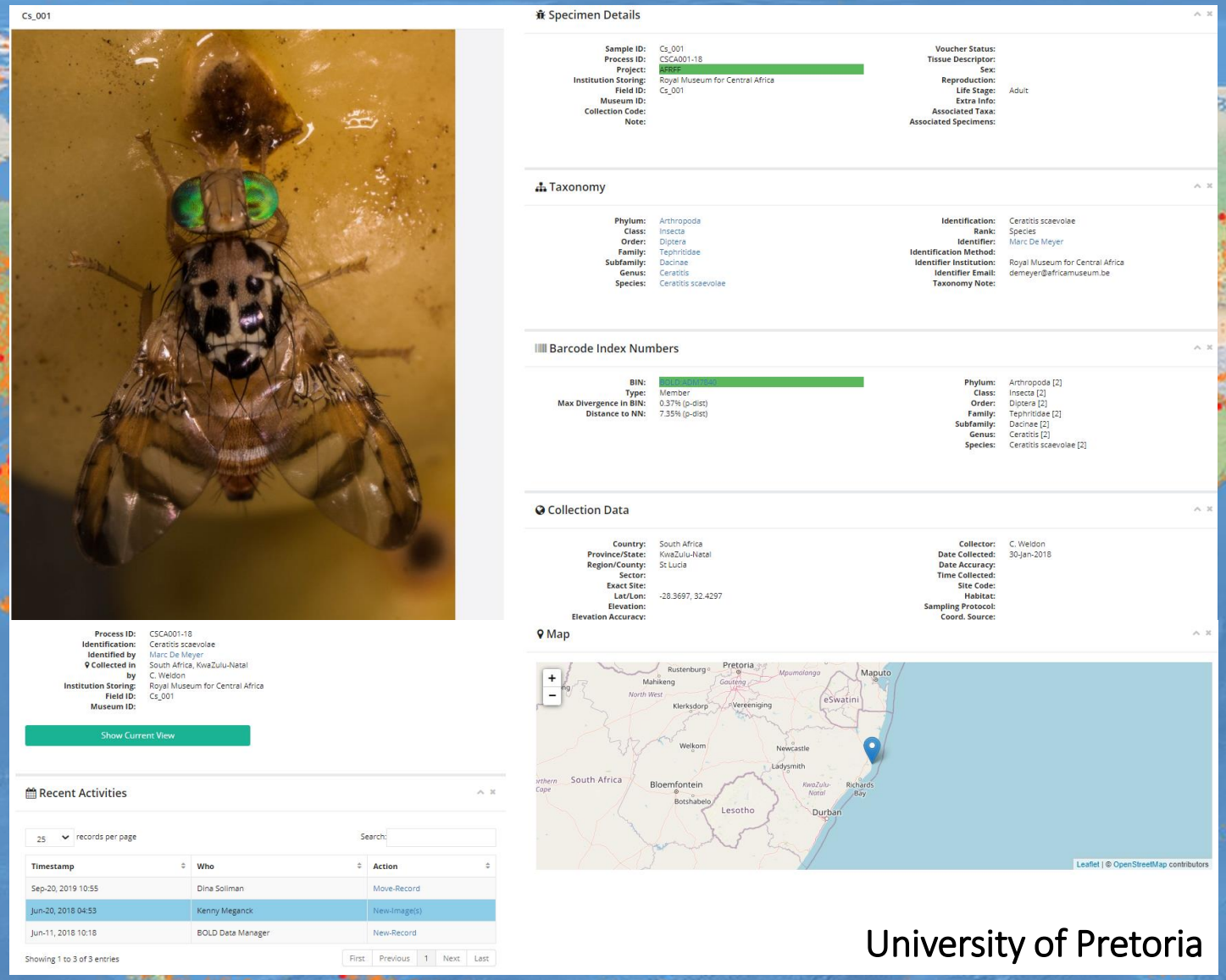
Weigand *et al*. Science of the Total Environment 678 (2019) 499–524

BOLD SYSTEMS

BOLD

- Barcodes
- With private data
- Strict requirements:

specimen vouchers, sampling data,



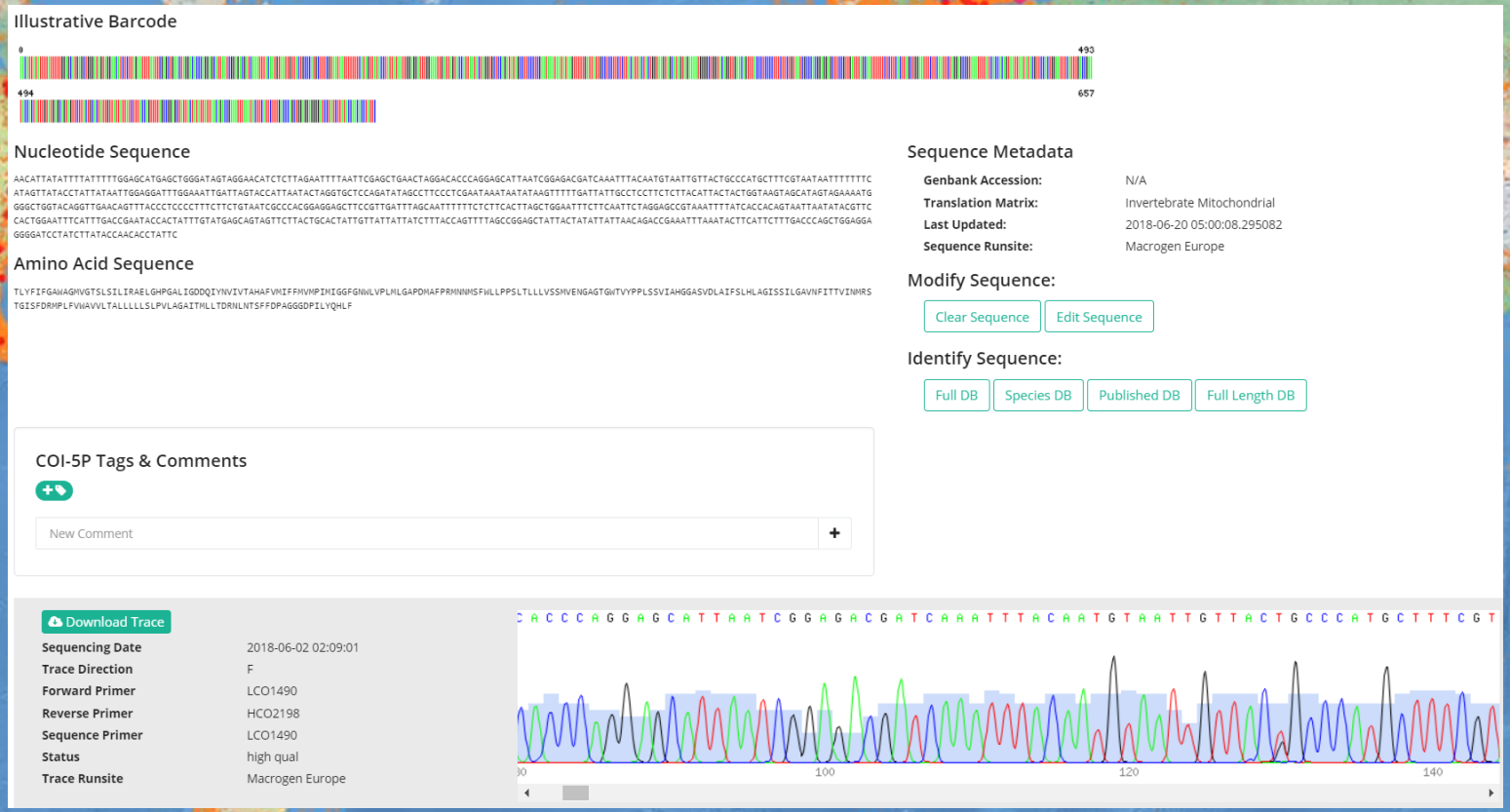
University of Pretoria

BOLD SYSTEMS

BOLD

- Barcodes
- With private data
- Strict requirements:

specimen vouchers, sampling data, sequence quality & length



BOLD SYSTEMS

Sequence reference databases

GenBank

- Public data
- Different markers available
- Limited requirements for sequence uploads
- No curation of data

Comprehensive

BOLD

- With private data
- Barcodes
- Strict requirements
- Community curation

Reliable

Combine both databases to compare and interpret own results, taking in to account the strengths and weaknesses of each.

Sequence database construction

Museum stored samples as an under-appreciated source for DNA and biodiversity informatics

Recapturing (molecular) data and providing it to open, public databases

incl. Raw data and information on experiment, processing and sampling

National Institute for Criminalistics and Criminology

Building a barcode reference library for the Belgian rove beetle species (Staphylinidae) of forensic importance in collaboration with the **NICC**

National Institute for Criminalistics and Criminology

- working from a list of forensically important species (50 sp.) found on corpses

Tableau 2.3 (suite)

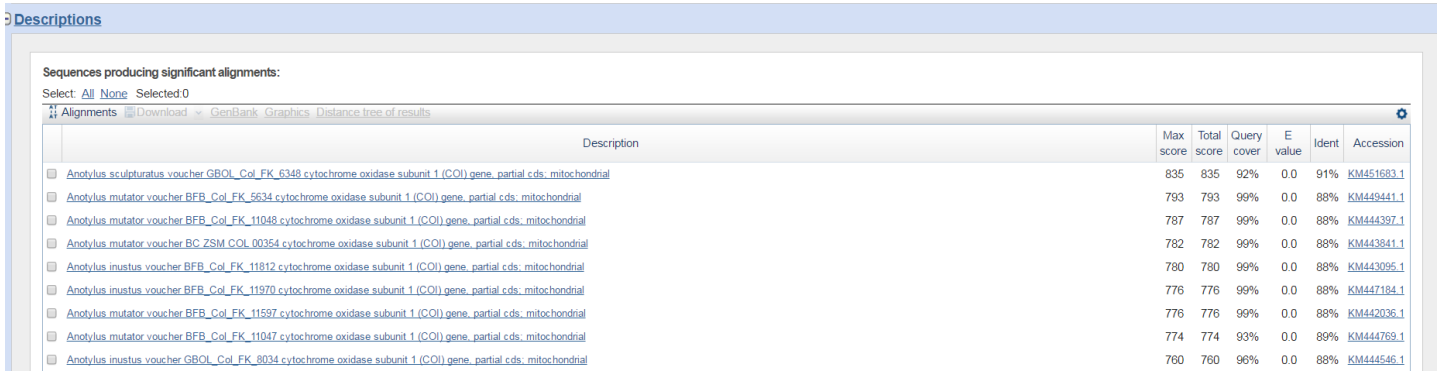
Espèce	Descripteur	Bibliographie associée	Cas concerné(s)	Intervalle *post mortem*
Staphylinidae				
Aleochara curtula	Goeze, 1777	Matuszewski et coll. 2008, Smith 1986, Kocarek 2003, Bourel et coll. 1999	46	-
Aleochara lata	Gravenhorst, 1802	Ozdemir et Sert 2009	46	-
Aleochara ruficornis	Gravenhorst, 1802	-	119	< 3 mois
Aleochara sp.		-	51	12 semaines
Amischa soror	Kraatz, 1856	-	98	> 6 mois
Anotylus sculpturatus	Gravenhorst, 1806	Matuszewski et coll. 2008	62	3-4 semaines
Atheta sp.		Kentner et Streit 1990, Nabaglo 1973	98	> 6 mois
Coprophilus striatulus	Fabricius, 1792	-	114	10 mois
Creophilus maxillosus	Linné, 1758	Turchetto et coll. 2001; Grassberger et Frank 2004; Matuszewski et coll. 2008, Smith 1986, Wyss et Cherix 2006, Ozdemir et Sert 2009, Matuszewski et coll. 2010, Kocarek 2003, Kentner et Streit 1990, Garcia-Rojo 2004	14, 15, 46, 72, 119	6 jours à plus de 3 mois
		Matuszewski et coll. 2008, Kocarek	63, 98, 113, 128,	19 jours à plus de

National Institute for Criminalistics and Criminology

- working from a list of forensically important species (50 sp.) found on corpses
- collect vouchered material from
  - RBINS (1970's-2000)
  - Gembloux (Université de Liège)
  - monitoring project (2015)

National Institute for Criminalistics and Criminology

- working from a list of forensically important species (50 sp.) found on corpses
- collect vouchered material from
- produce novel genetic reference data



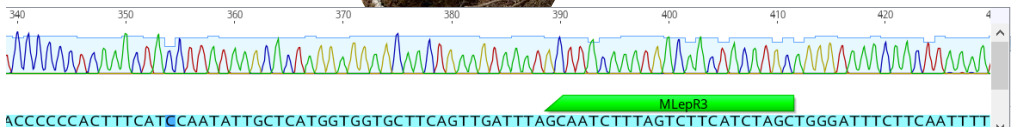
Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

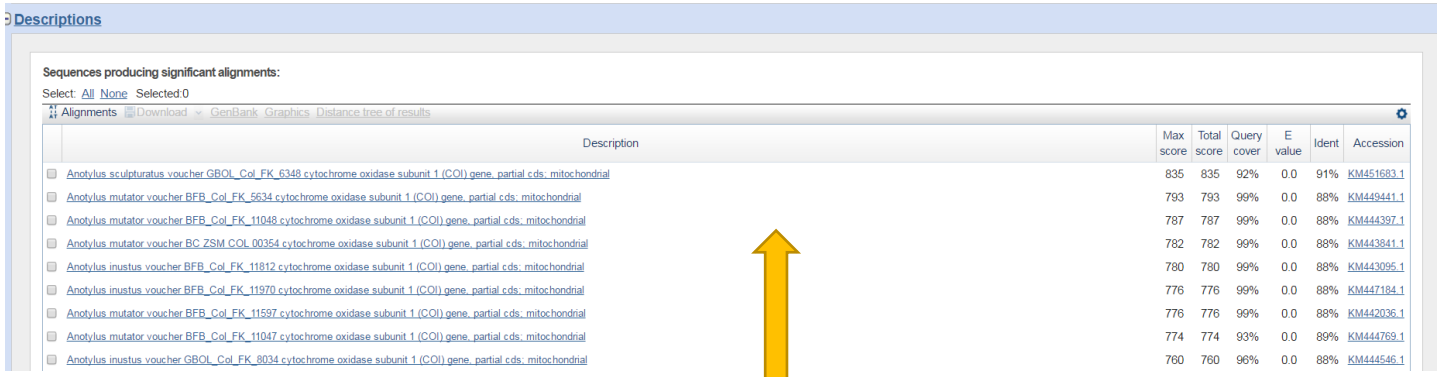
Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Anotylus sculpturatus voucher GBOL_Col_FK_6348 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	835	835	92%	0.0	91%	KM451683.1
Anotylus mutator voucher BFB_Col_FK_5634 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	793	793	99%	0.0	88%	KM449441.1
Anotylus mutator voucher BFB_Col_FK_11048 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	787	787	99%	0.0	88%	KM444397.1
Anotylus mutator voucher BC ZSM COL 00354 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	782	782	99%	0.0	88%	KM443841.1
Anotylus inustus voucher BFB_Col_FK_11812 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	780	780	99%	0.0	88%	KM443095.1
Anotylus inustus voucher BFB_Col_FK_11970 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	776	776	99%	0.0	88%	KM447184.1
Anotylus mutator voucher BFB_Col_FK_11597 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	776	776	99%	0.0	88%	KM442036.1
Anotylus mutator voucher BFB_Col_FK_11047 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	774	774	93%	0.0	89%	KM444769.1
Anotylus inustus voucher GBOL_Col_FK_8034 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	760	760	96%	0.0	88%	KM444546.1



National Institute for Criminalistics and Criminology

- working from a list of forensically important species (50 sp.) found on corpses
- collect vouchered material from
- produce novel genetic reference data



Descriptions

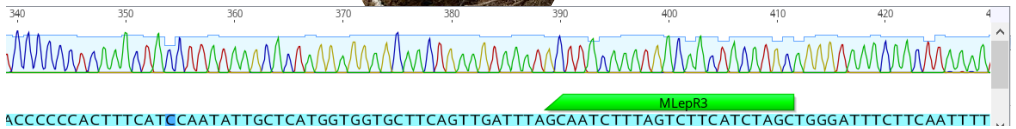
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

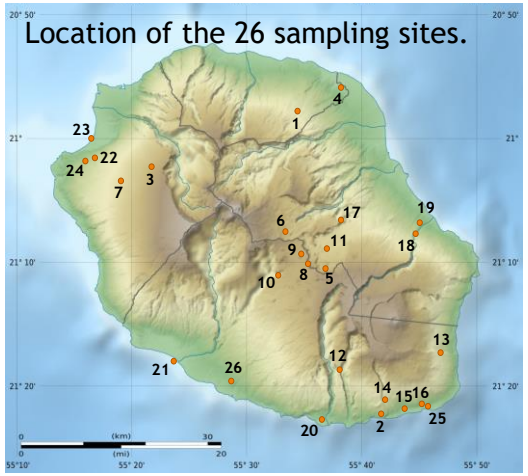
Description	Max score	Total score	Query cover	E value	Ident	Accession
Anotylus sculpturatus voucher GBOL_Col_FK_6348 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	835	835	92%	0.0	91%	KM451683.1
Anotylus mutator voucher BFB_Col_FK_5634 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	793	793	99%	0.0	88%	KM449441.1
Anotylus mutator voucher BFB_Col_FK_11048 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	787	787	99%	0.0	88%	KM444397.1
Anotylus mutator voucher BC ZSM COL 00354 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	782	782	99%	0.0	88%	KM443841.1
Anotylus inustus voucher BFB_Col_FK_11812 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	780	780	99%	0.0	88%	KM443095.1
Anotylus inustus voucher BFB_Col_FK_11970 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	776	776	99%	0.0	88%	KM447184.1
Anotylus mutator voucher BFB_Col_FK_11597 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	776	776	99%	0.0	88%	KM442036.1
Anotylus mutator voucher BFB_Col_FK_11047 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	774	774	93%	0.0	89%	KM444769.1
Anotylus inustus voucher GBOL_Col_FK_8034 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	760	760	96%	0.0	88%	KM444546.1

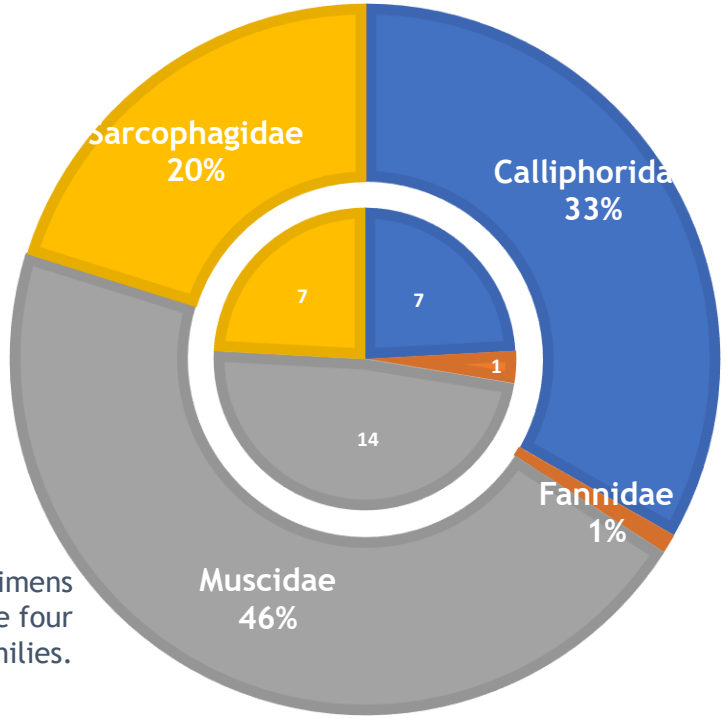
Comparison



Forensically important flies (Diptera) of the island of La Réunion

- even earlier colonizers of corpses
- local, representative, and comprehensive reference library
- 195 barcodes were generated for 29 species of which at least 10 have a forensic relevance





Number of species (inner circle) and proportion of specimens (n=337; outer circle) collected for each of the four forensically important fly families.

Belgian Air Force and Brussels Airport Company

- identify bird remains
- several species

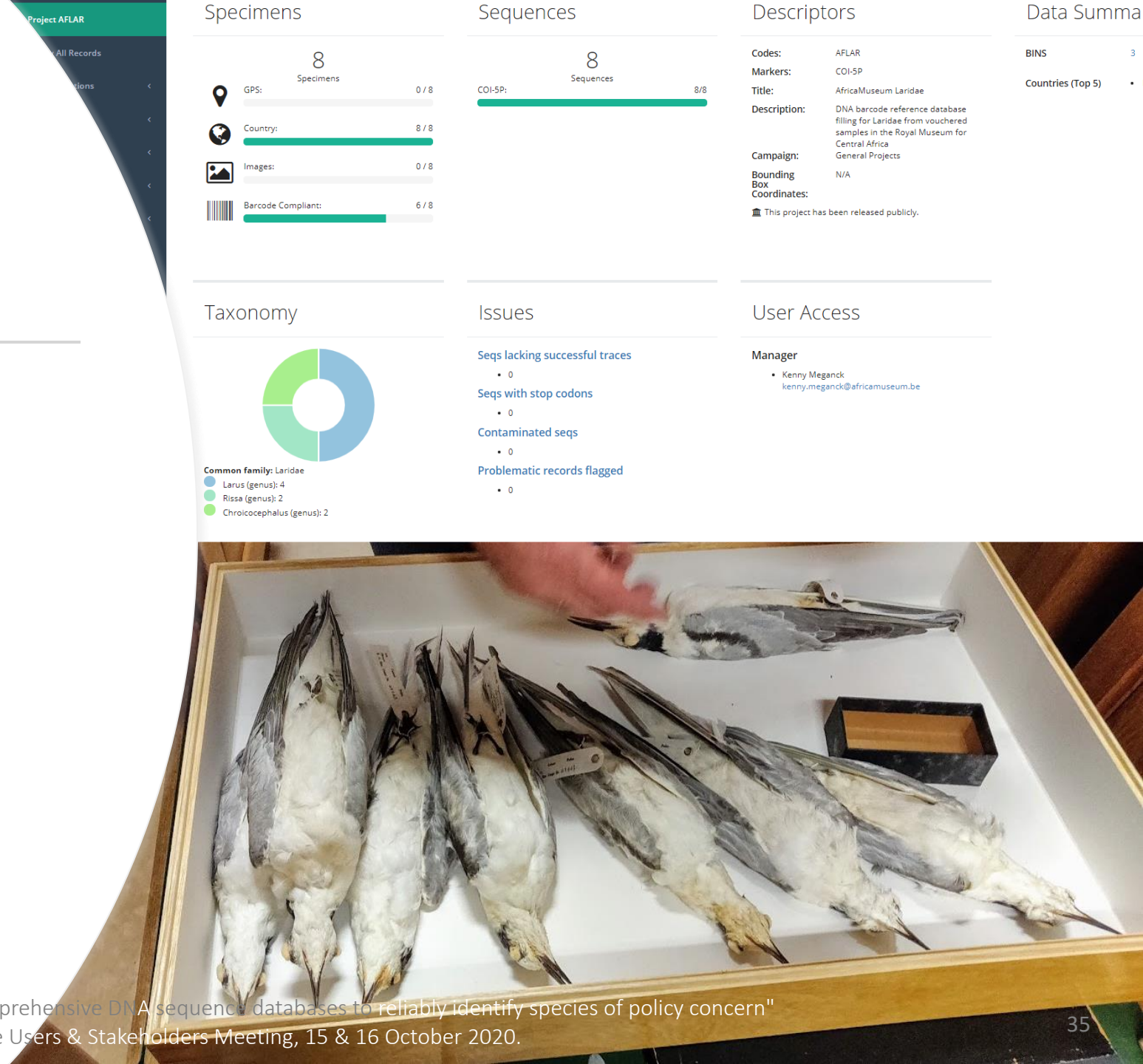
- some less represented
- databases filling with museum collection material

Database building

- Laridae*: Sterna, Larus, Rissa*

- *Falco*

Rock Kestrel (*Falco rupicolis*)
By Bernard Dupont [CC BY-SA 2.0]



Conclusion

To give a reliable answer to an identification request,

we need

a reference database that is representative,

trustable (vouchered) material,

check the data for errors,

and add reference sequences ourselves.

BopCo contact details

- **Royal Belgian Institute of Natural Sciences**

  Vautierstraat 29

  1000 Brussels

  +32 (0)2 627 41 23

- **Royal Museum for Central Africa**

  Leuvensesteenweg 13

  3080 Tervuren

  +32(0)2 769 58 54

bopco@naturalsciences.be http://bopco.myspecies.info/