Curating BOLD

Automated pipeline to focus expert intervention

Ben Price

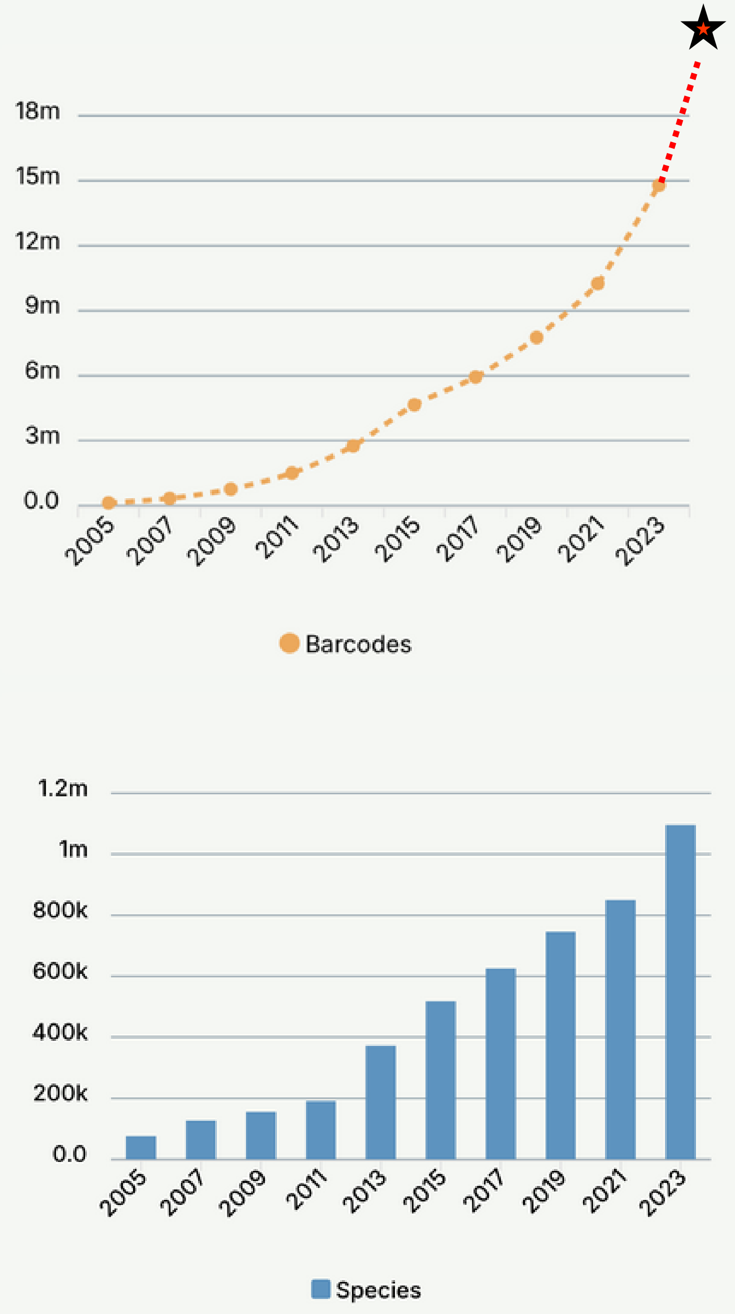
Fabian Deister

Rutger Vos

What’s the issue?

- BOLD growing at an impressive rate
- Variable sample data quality (incl. ID)
- Is BIN sharing real or mistaken ID?
- No easy curation mechanism

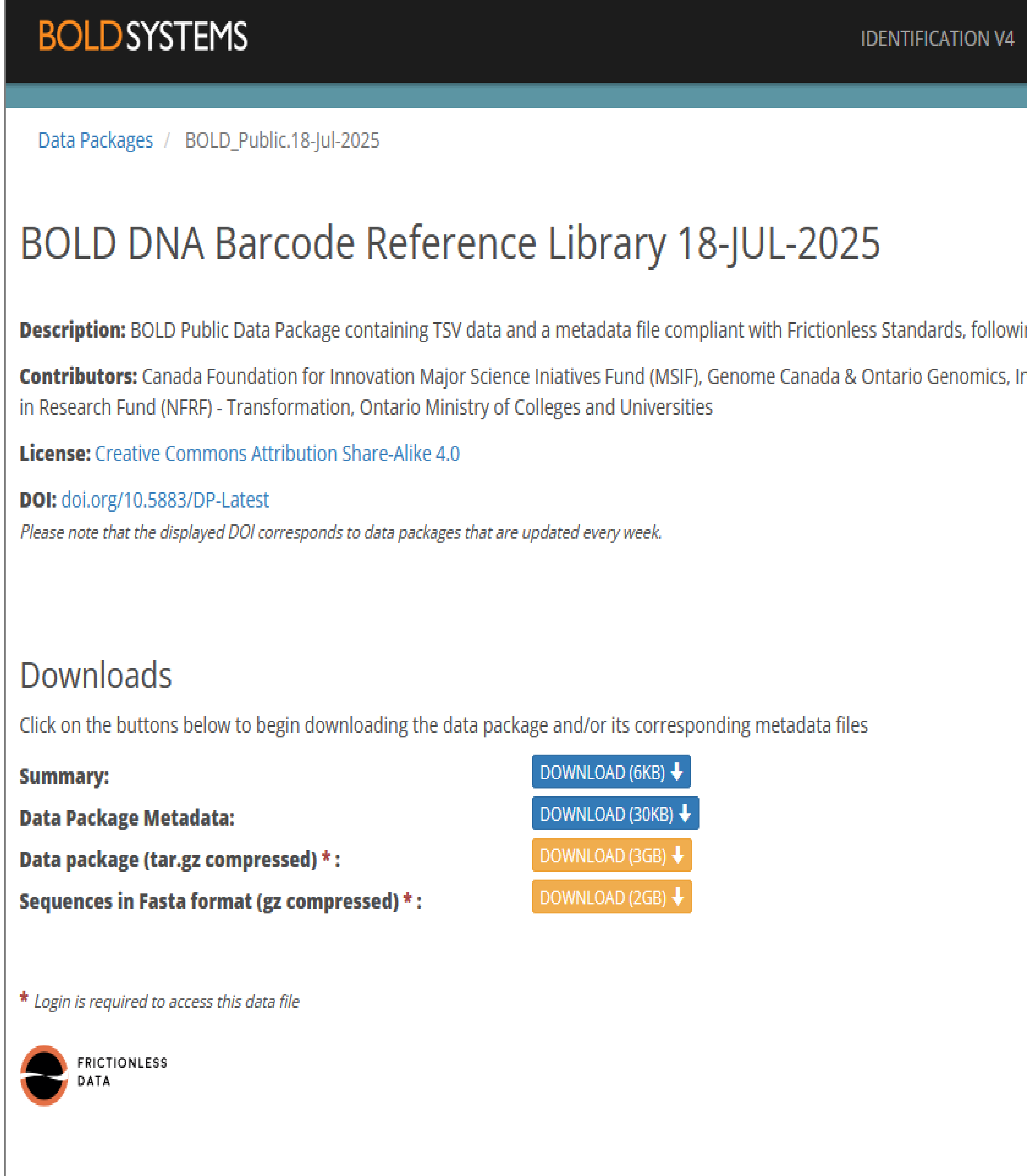
Growth of Barcodes & Species on BOLD



What’s the approach?

- Automated pipeline
- Public data packages
  - No private data ☹
- Flexible inputs
  - Species AND / OR geography

Biodiversity
Genomics
Europe



Data Packages / BOLD_Public.18-Jul-2025

BOLD DNA Barcode Reference Library 18-JUL-2025

Description: BOLD Public Data Package containing TSV data and a metadata file compliant with Frictionless Standards, followi

Contributors: Canada Foundation for Innovation Major Science Iniatives Fund (MSIF), Genome Canada & Ontario Genomics, Ir in Research Fund (NFRF) - Transformation, Ontario Ministry of Colleges and Universities

License: Creative Commons Attribution Share-Alike 4.0

DOI: doi.org/10.5883/DP-Latest

Please note that the displayed DOI corresponds to data packages that are updated every week.

Downloads

Click on the buttons below to begin downloading the data package and/or its corresponding metadata files

Summary:	DOWNLOAD (6KB) ⬇
Data Package Metadata:	DOWNLOAD (30KB) ⬇
Data package (tar.gz compressed) * :	DOWNLOAD (3GB) ⬇
Sequences in Fasta format (gz compressed) * :	DOWNLOAD (2GB) ⬇

* *Login is required to access this data file*

Pipeline summary

GitHub →

1. Record selection (species + geography + BIN)
2. Scoring against criteria
3. Ranking based on criteria
4. OTU clustering & phylogeny
5. BAGS grading
6. Representative selection
   - Best rank for (country + species + OTU)
7. Curation package (each family)
   - Database file
   - Curation checklist
   - Phylogeny
8. Summary report

	specimen rank					
Criteria	**1**	**2**	**3**	**4**	**5**	**6**
Species level ID	✓	✓	✓	✓	✓	✓
Type specimen	✓					
Good quality sequence		✓	✓	✓	✓	
Image(s) available		✓				
Collection (Country)		✓	✓	✓		
ID (identifier named)		✓	✓			
ID method (method is not BIN match)		✓	✓			
Public voucher (has museum ID)		✓(or)	✓(or)			
Public voucher (agreed institution)		✓(or)	✓(or)			
Public voucher (agreed voucher type)		✓(or)	✓(or)			
Collection (Date)		✓				
Collection (Site)		✓(or)				
Collection (GPS coordinates)		✓(or)				
Collection (Sector)		✓(or)				
Collection (Region)		✓(or)				
Collector named		✓				

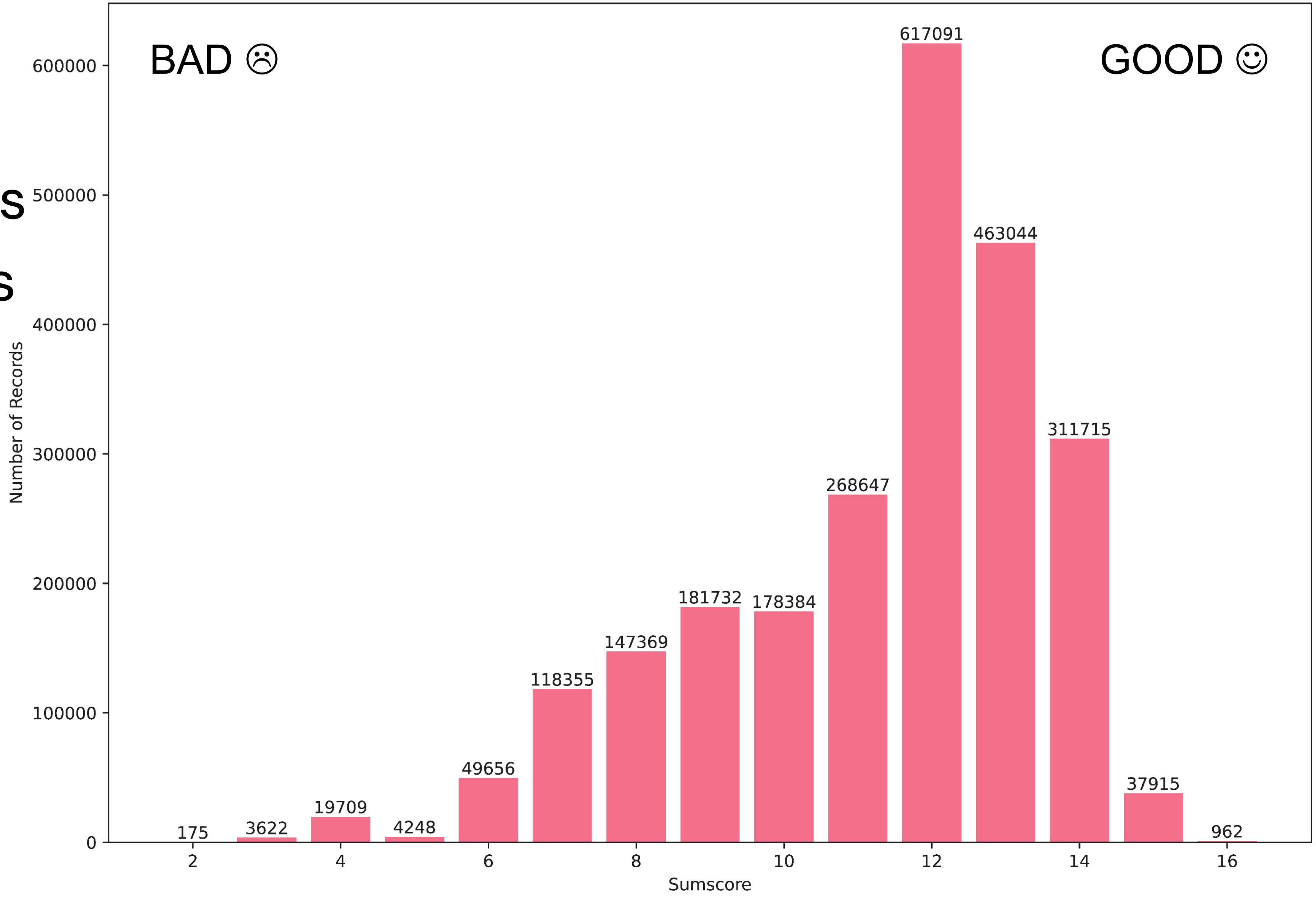
Results (score)

https://species.iboleurope.org/

- 150k species
- European countries
- 20M BOLD records

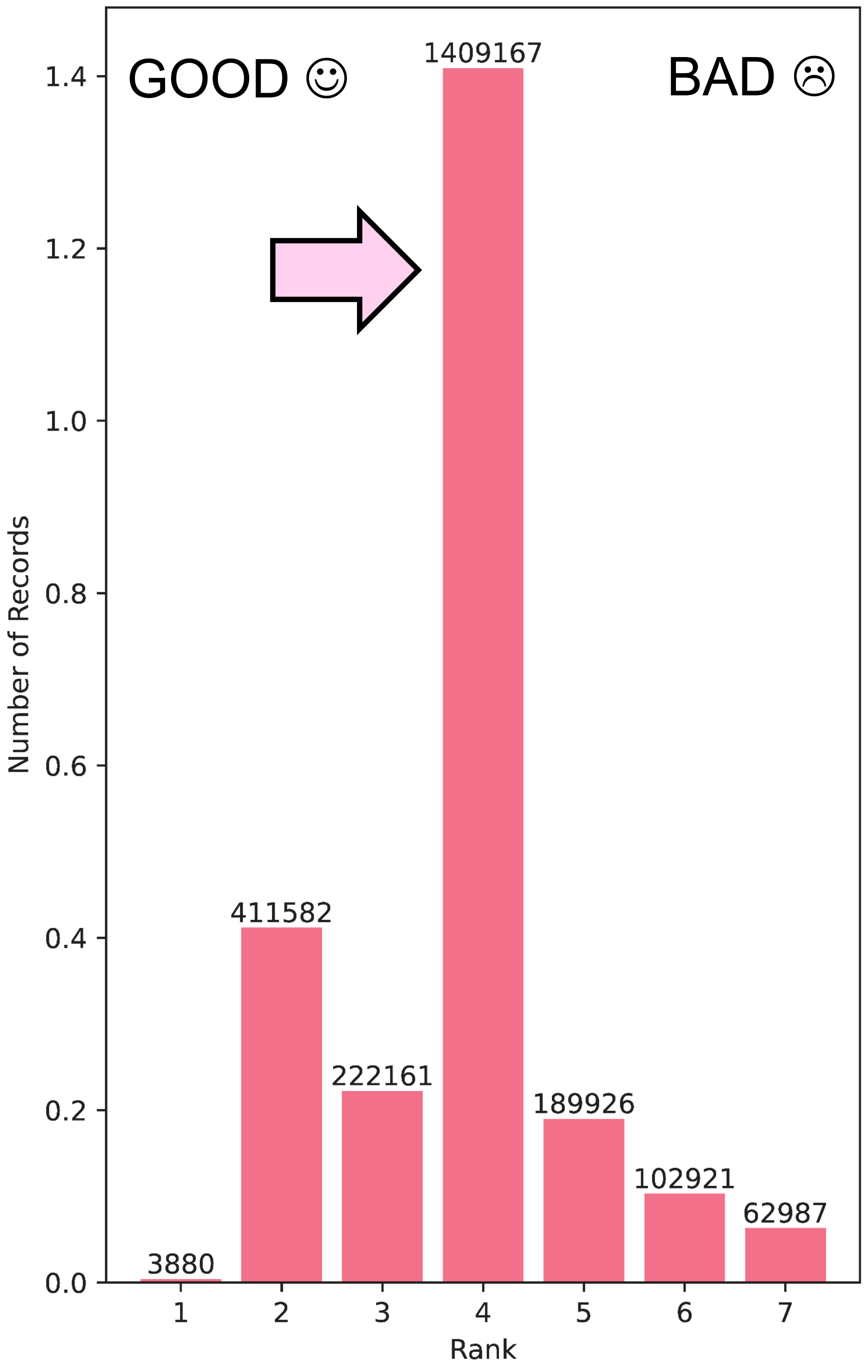
<u>Database stats:</u>

- 2.4 M records
- 87k species
- 93k BINs
- 300k OTUs

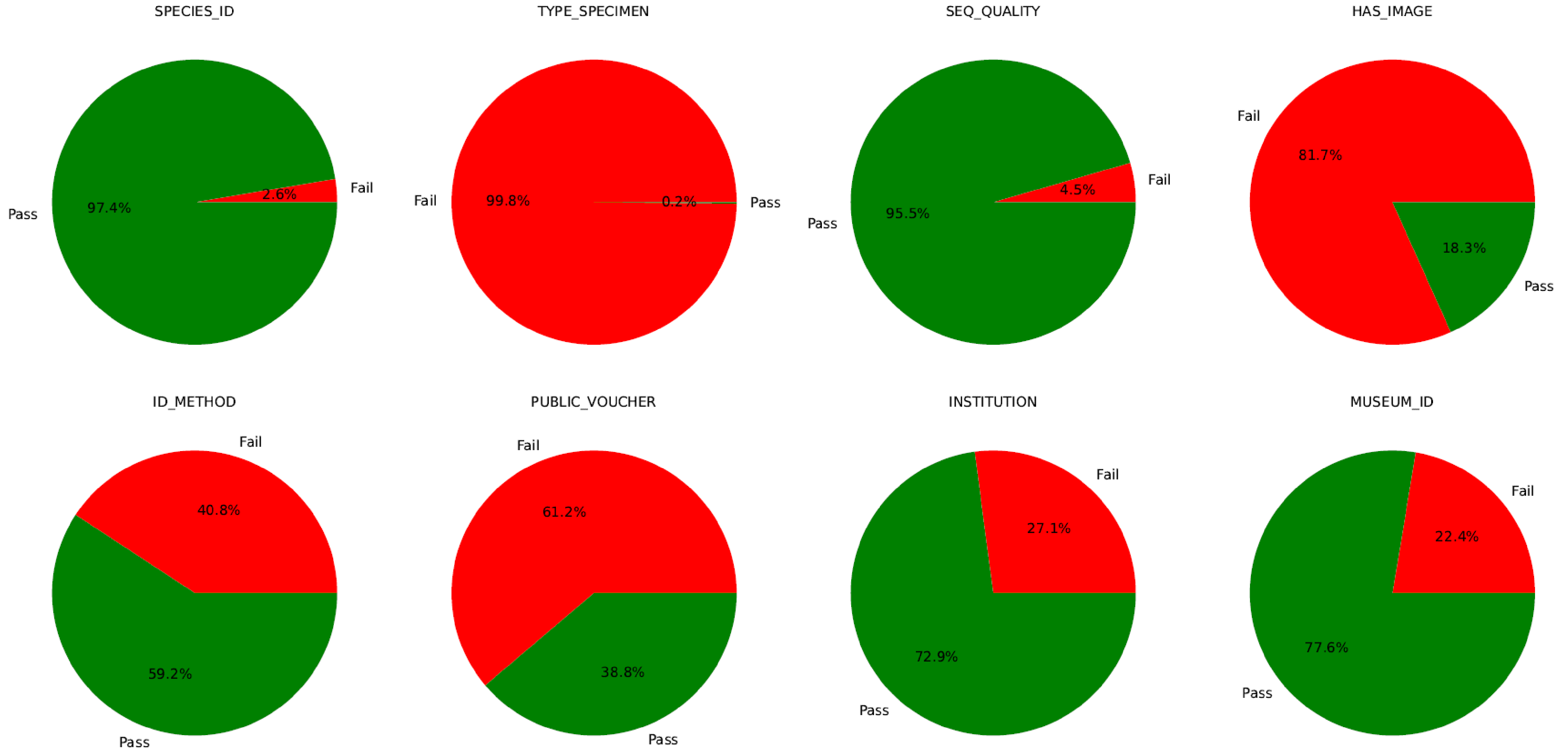


Results (rank)

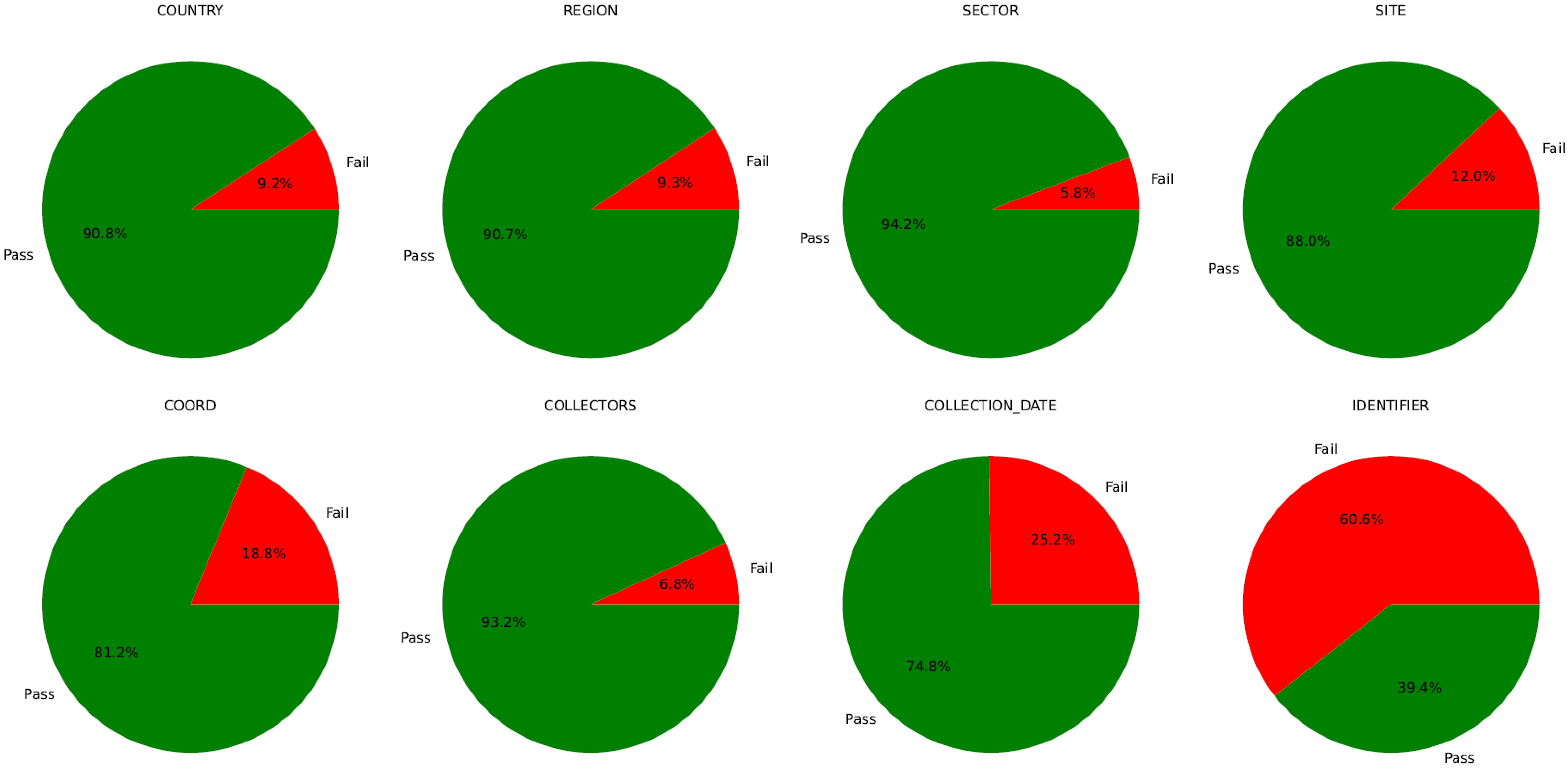
	specimen rank					
Criteria	**1**	**2**	**3**	**4**	**5**	**6**
Species level ID	✓	✓	✓	✓	✓	✓
Type specimen	✓					
Good quality sequence		✓	✓	✓	✓	
Image(s) available		✓				
Collection (Country)		✓	✓	✓		
ID (identifier named)		✓	✓			
ID method (method is not BIN match)		✓	✓			
Public voucher (has museum ID)		✓(or)	✓(or)			
Public voucher (agreed institution)		✓(or)	✓(or)			
Public voucher (agreed voucher type)		✓(or)	✓(or)			
Collection (Date)		✓				
Collection (Site)		✓(or)				
Collection (GPS coordinates)		✓(or)				
Collection (Sector)		✓(or)				
Collection (Region)		✓(or)				
Collector named		✓				



Results (score)

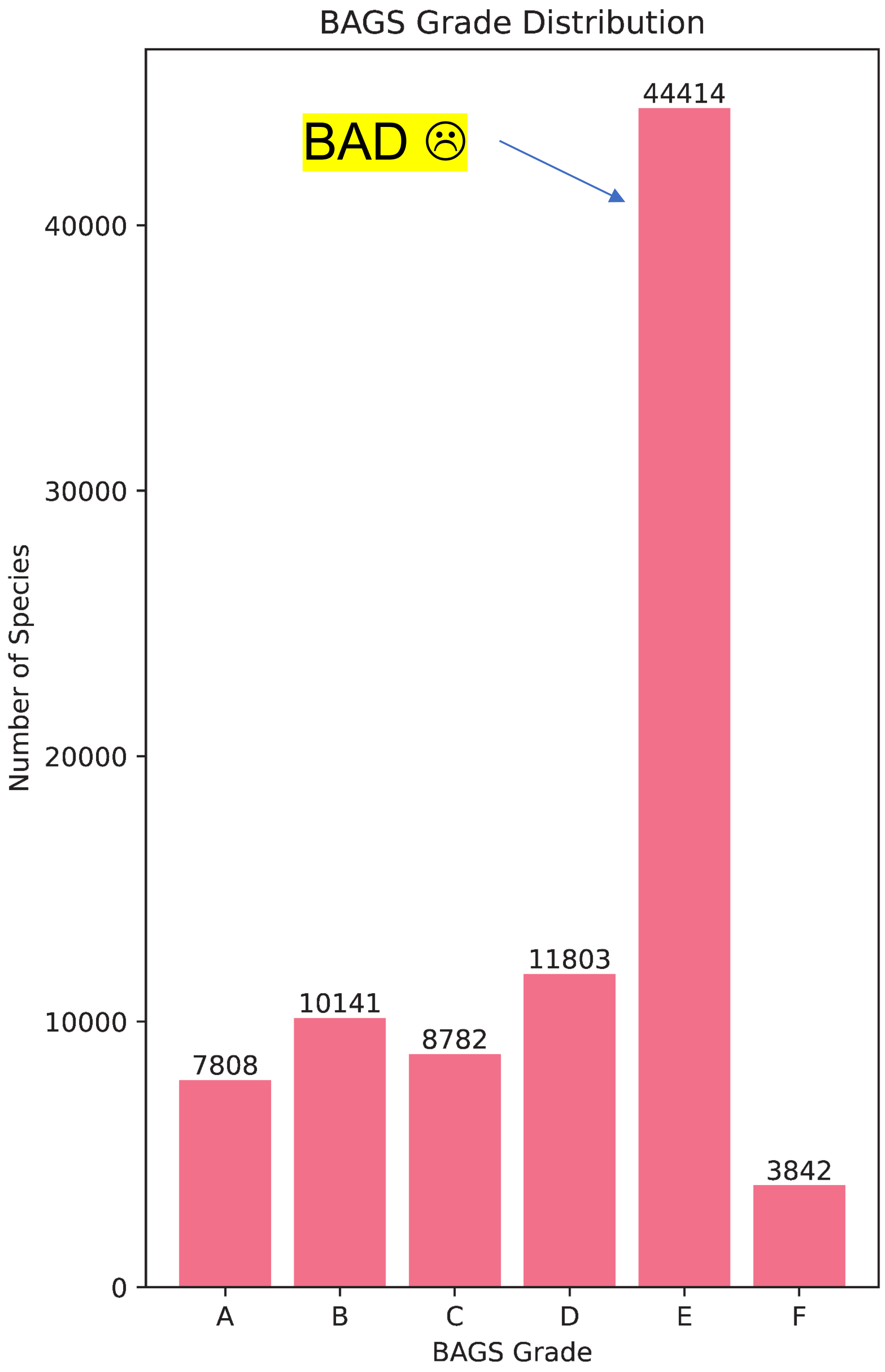


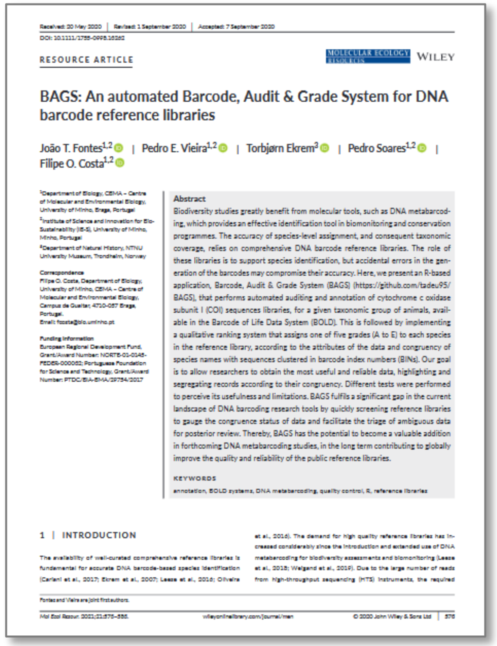
Results (score)



Results (BAGS)

- A: >10 specimens + 1 BIN
- B: 3–10 specimens +1 BIN
- C: >1 BIN
- D: <3 specimens + 1 BIN
- E: >1 name in a BIN




Received: 20 May 2020 | Revised: 1 September 2020 | Accepted: 7 September 2020
DOI: 10.1111/1755-0998.13262

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries

João T. Fontes[1,2] | Pedro E. Vieira[1,2] | Torbjørn Ekrem[3] | Pedro Soares[1,2] | Filipe O. Costa[1,2]

[1] Department of Biology, CBMA – Centre of Molecular and Environmental Biology, University of Minho, Braga, Portugal
[2] Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Minho, Portugal
[3] Department of Natural History, NTNU University Museum, Trondheim, Norway

Correspondence
Filipe O. Costa, Department of Biology, University of Minho, CBMA – Centre of Molecular and Environmental Biology, Campus de Gualtar, 4710-057 Braga, Portugal.
Email: fcosta@bio.uminho.pt

Funding information
European Regional Development Fund, Grant/Award Number: NORTE-01-0145-FEDER-000062; Portuguese Foundation for Science and Technology, Grant/Award Number: PTDC/BIA-BMA/29754/2017

Abstract
Biodiversity studies greatly benefit from molecular tools, such as DNA metabarcoding, which provides an effective identification tool in biomonitoring and conservation programmes. The accuracy of species-level assignment, and consequent taxonomic coverage, relies on comprehensive DNA barcode reference libraries. The role of these libraries is to support species identification, but accidental errors in the generation of the barcodes may compromise their accuracy. Here, we present an R-based application, Barcode, Audit & Grade System (BAGS) (https://github.com/tadeu95/BAGS), that performs automated auditing and annotation of cytochrome c oxidase subunit I (COI) sequences libraries, for a given taxonomic group of animals, available in the Barcode of Life Data System (BOLD). This is followed by implementing a qualitative ranking system that assigns one of five grades (A to E) to each species in the reference library, according to the attributes of the data and congruency of species names with sequences clustered in barcode index numbers (BINs). Our goal is to allow researchers to obtain the most useful and reliable data, highlighting and segregating records according to their congruency. Different tests were performed to perceive its usefulness and limitations. BAGS fulfils a significant gap in the current landscape of DNA barcoding research tools by quickly screening reference libraries to gauge the congruence status of data and facilitate the triage of ambiguous data for posterior review. Thereby, BAGS has the potential to become a valuable addition in forthcoming DNA metabarcoding studies, in the long term contributing to globally improve the quality and reliability of the public reference libraries.

KEYWORDS
annotation, BOLD systems, DNA metabarcoding, quality control, R, reference libraries

1 | INTRODUCTION

The availability of well-curated comprehensive reference libraries is fundamental for accurate DNA barcode-based species identification (Carlani et al., 2017; Ekrem et al., 2007; Leese et al., 2016; Oliveira et al., 2016). The demand for high quality reference libraries has increased considerably since the introduction and extended use of DNA metabarcoding for biodiversity assessments and biomonitoring (Leese et al., 2018; Weigand et al., 2019). Due to the large number of reads from high-throughput sequencing (HTS) instruments, the required

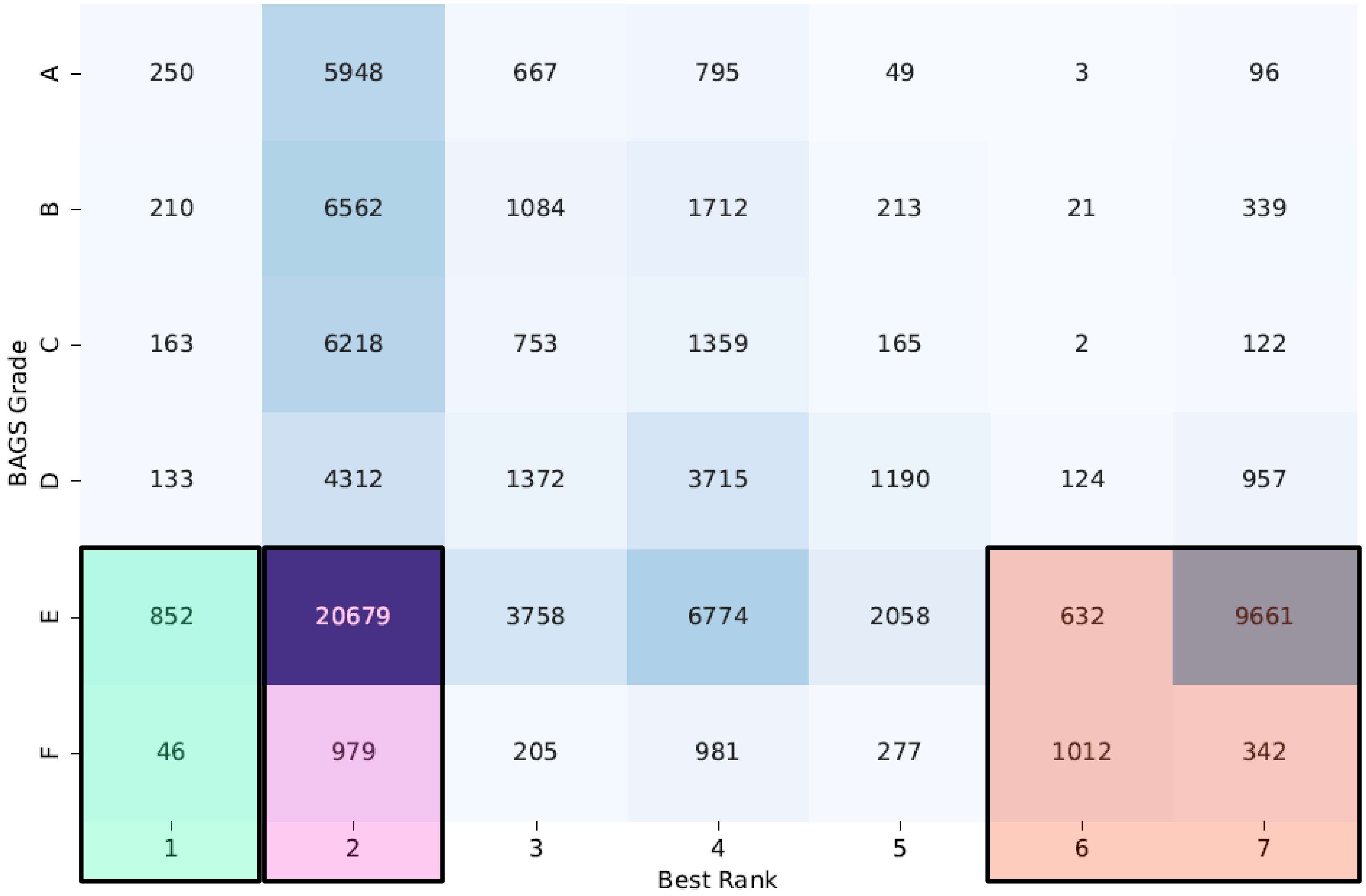
Fontes and Vieira are joint first authors.

iBOL EUROPE

Results

- Types
- Good metadata
- Difficult

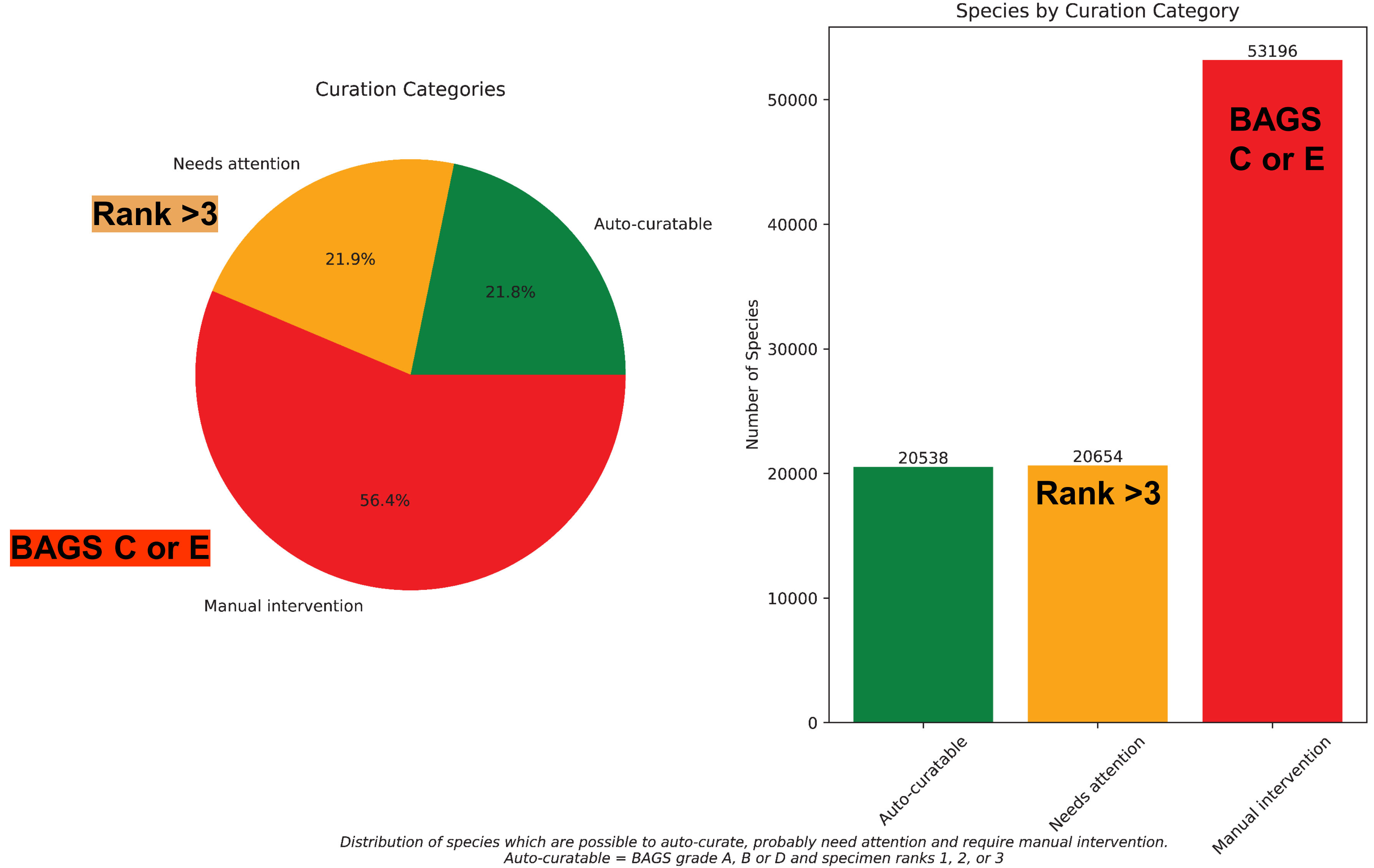
Count of species



BAGS Grade	1	2	3	4	5	6	7
A	250	5948	667	795	49	3	96
B	210	6562	1084	1712	213	21	339
C	163	6218	753	1359	165	2	122
D	133	4312	1372	3715	1190	124	957
E	852	20679	3758	6774	2058	632	9661
F	46	979	205	981	277	1012	342

Auto-curation

Curation Analysis

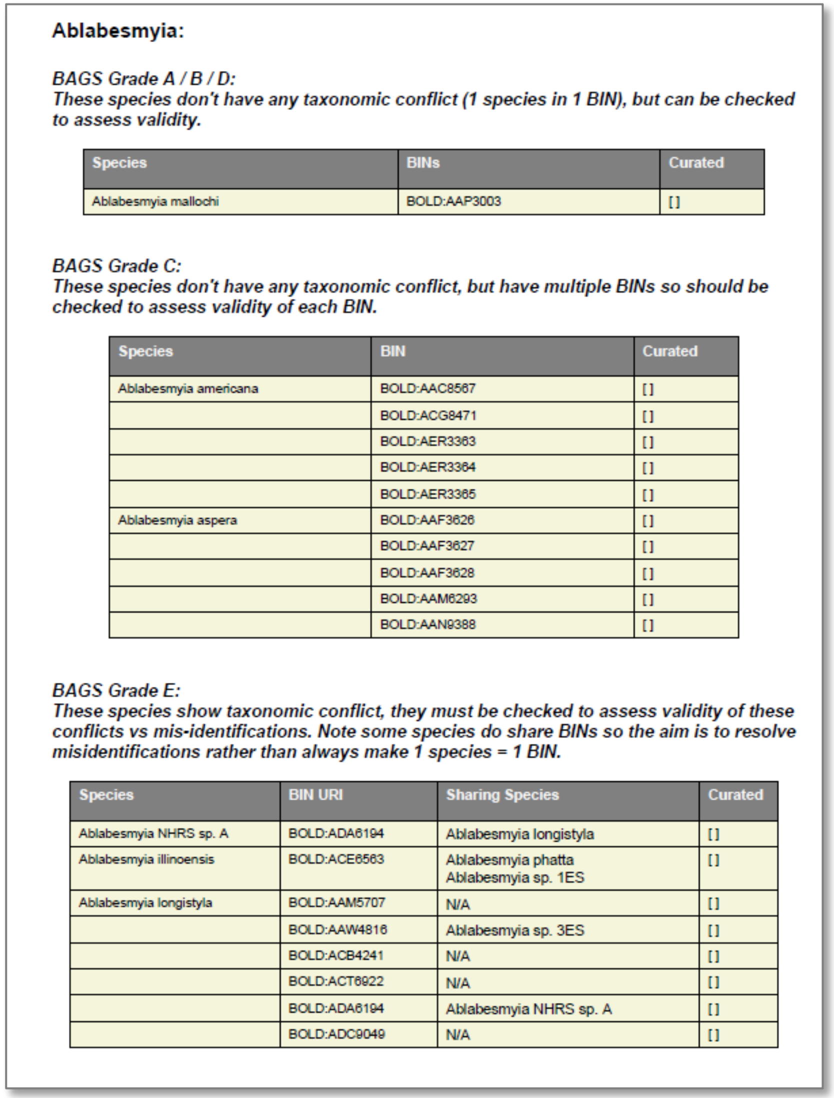


Distribution of species which are possible to auto-curate, probably need attention and require manual intervention.
Auto-curatable = BAGS grade A, B or D and specimen ranks 1, 2, or 3
Need attention = BAGS grade A, B or D and specimen ranks >=4
Manual intervention = BAGS grade C or E

Curation package

iBOL Europe BOLD curation datasets

Browse and download curated database files organized by taxonomic classification



Ablabesmyia:

BAGS Grade A / B / D:
These species don't have any taxonomic conflict (1 species in 1 BIN), but can be checked to assess validity.

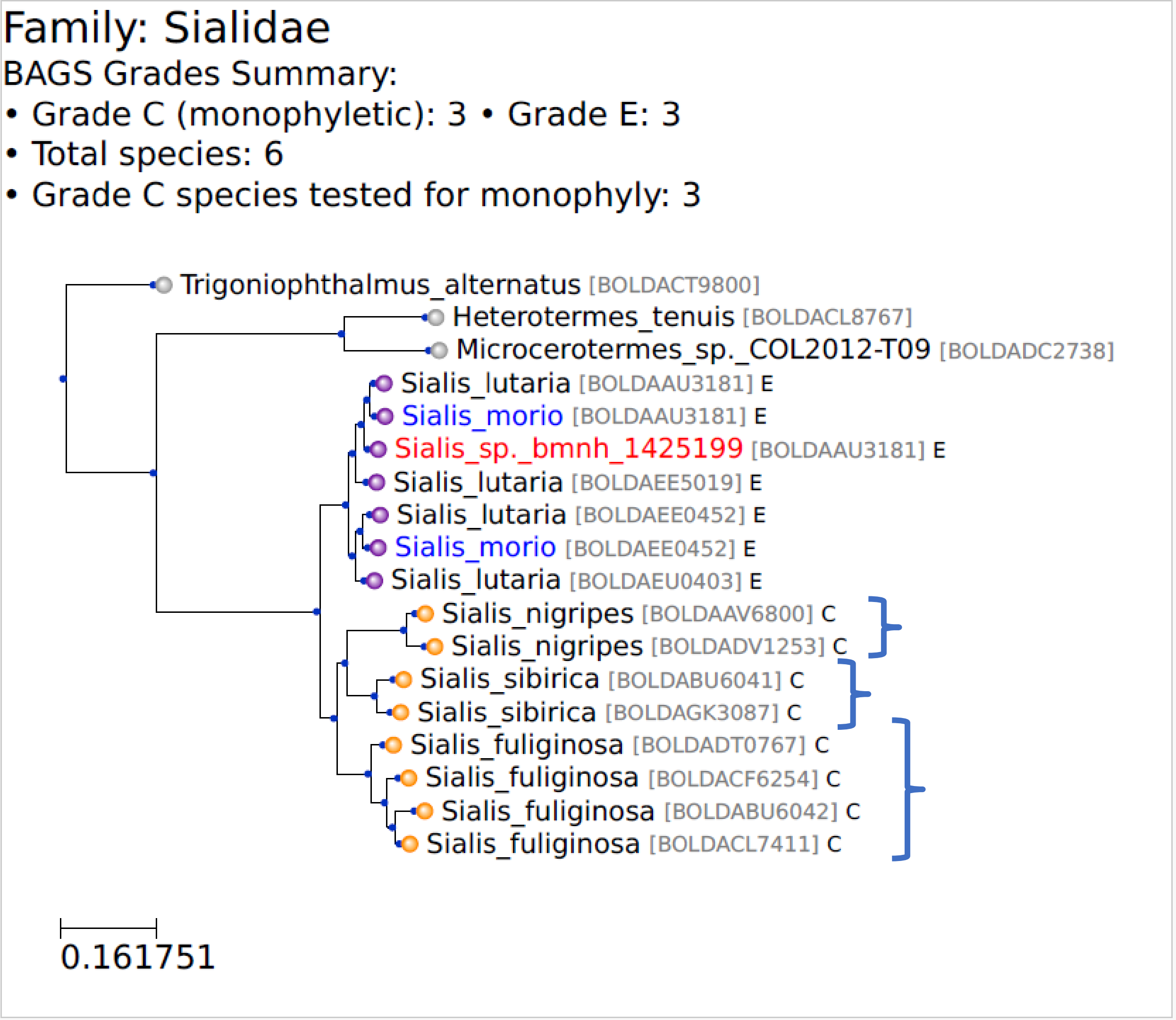
Species	BINs	Curated
Ablabesmyia mallochi	BOLD:AAP3003	[]

BAGS Grade C:
These species don't have any taxonomic conflict, but have multiple BINs so should be checked to assess validity of each BIN.

Species	BIN	Curated
Ablabesmyia americana	BOLD:AAC8567	[]
	BOLD:ACG8471	[]
	BOLD:AER3363	[]
	BOLD:AER3364	[]
	BOLD:AER3365	[]
Ablabesmyia aspera	BOLD:AAF3626	[]
	BOLD:AAF3627	[]
	BOLD:AAF3628	[]
	BOLD:AAM6293	[]
	BOLD:AAN9388	[]

BAGS Grade E:
These species show taxonomic conflict, they must be checked to assess validity of these conflicts vs mis-identifications. Note some species do share BINs so the aim is to resolve misidentifications rather than always make 1 species = 1 BIN.

Species	BIN URI	Sharing Species	Curated
Ablabesmyia NHRS sp. A	BOLD:ADA6194	Ablabesmyia longistyla	[]
Ablabesmyia illinoensis	BOLD:ACE6563	Ablabesmyia phatta Ablabesmyia sp. 1ES	[]
Ablabesmyia longistyla	BOLD:AAM5707	N/A	[]
	BOLD:AAW4816	Ablabesmyia sp. 3ES	[]
	BOLD:ACB4241	N/A	[]
	BOLD:ACT6922	N/A	[]
	BOLD:ADA6194	Ablabesmyia NHRS sp. A	[]
	BOLD:ADC9049	N/A	[]

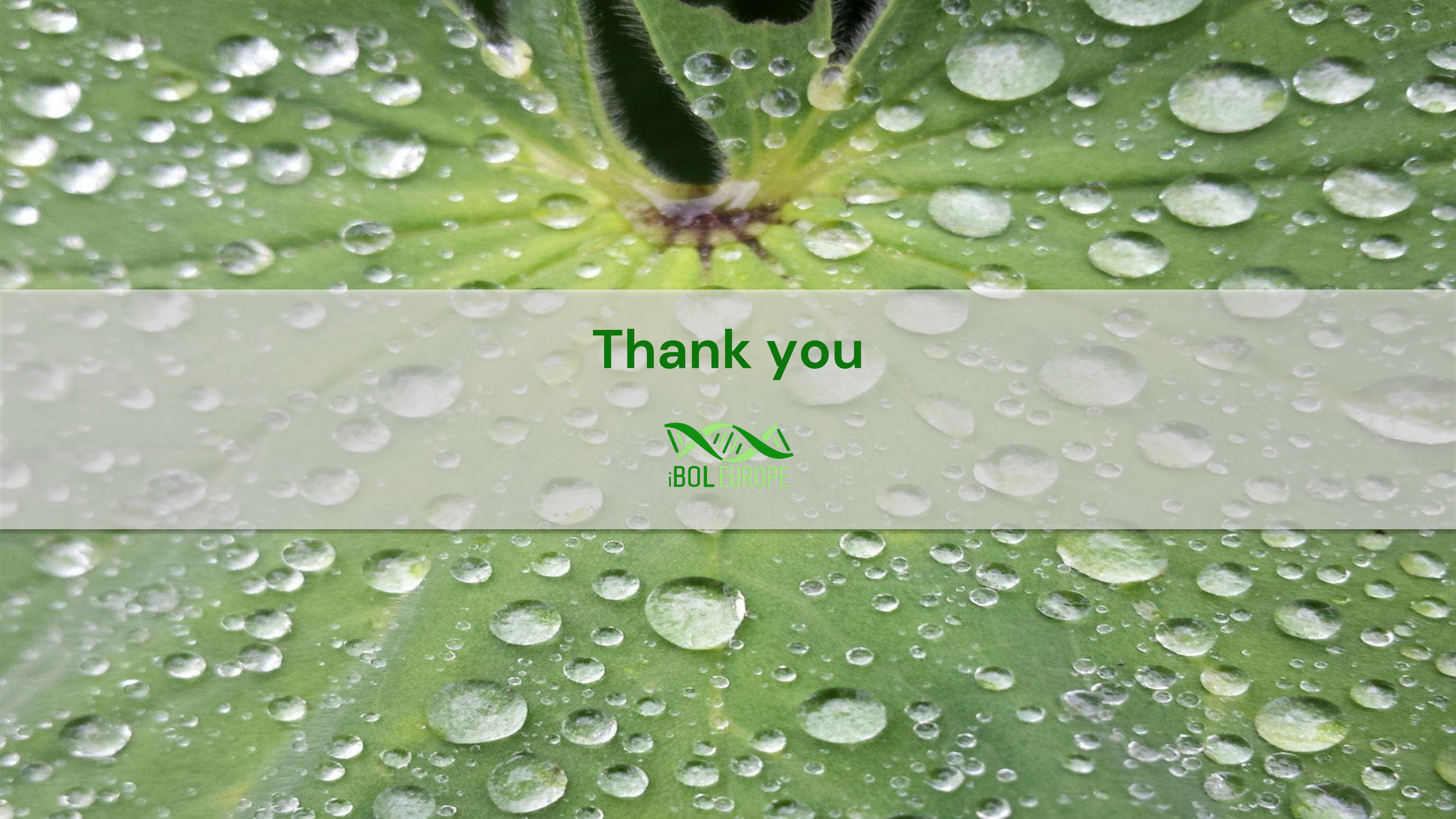


Family: Sialidae

BAGS Grades Summary:

- Grade C (monophyletic): 3 • Grade E: 3
- Total species: 6
- Grade C species tested for monophyly: 3

Trigoniophthalmus_alternatus [BOLDACT9800]
Heterotermes_tenuis [BOLDACL8767]
Microcerotermes_sp._COL2012-T09 [BOLDADC2738]
Sialis_lutaria [BOLDAAU3181] E
Sialis_morio [BOLDAAU3181] E
Sialis_sp._bmnh_1425199 [BOLDAAU3181] E
Sialis_lutaria [BOLDAEE5019] E
Sialis_lutaria [BOLDAEE0452] E
Sialis_morio [BOLDAEE0452] E
Sialis_lutaria [BOLDAEU0403] E
Sialis_nigripes [BOLDAAV6800] C
Sialis_nigripes [BOLDADV1253] C
Sialis_sibirica [BOLDABU6041] C
Sialis_sibirica [BOLDAGK3087] C
Sialis_fuliginosa [BOLDADT0767] C
Sialis_fuliginosa [BOLDACF6254] C
Sialis_fuliginosa [BOLDABU6042] C
Sialis_fuliginosa [BOLDACL7411] C
0.161751



Thank you

iBOL EUROPE