





Shining a light on 'dark taxa': naming and discovering DNA-based biodiversity data

Urmas Kõljalg, University of Tartu, Estonia Dmitry Schigel, GBIF Secretariat

17 OCTOBER 2018

OUTLINE



- Taxonomic bias in GBIF
- Fungi in GBIF
- UNITE and GBIF
- Latin and non-Latin names
- Going beyond fungi











TAXONOMIC BIAS IN GBIF MEDIATED DATA



vertical line at x = 0

'ideal' N occurrences per class, where each class is sampled proportionally to its number of known species

Green and red bars

classes that are over- and under-represented in GBIF compared to 'ideal' sampling

Insects: <200M occurrences Birds: >200M occurrences





GLOBAL BIODIVERSITY DATA: 2018



Number of species worldwide:

Purvis & Hector 2000 Nature 405 doi:10.1038/35012221



Number of species occurrences GBIF.org, 9 October 2018



FUNGI IN GBIF

14,940,711 occurrences:
8M+ observations / 5M+ specimens
11,667,185 georeferenced
830,508 with images
217,699 species



As for 9 October 2018

WHY ARE FUNGI UNDERREPRESENTED?



- Fungi are difficult to identify
- Few citizen science portals deal (seriously) with fungi
- Molecular researchers generally don't publish to GBIF (yet?)
 - GBIF portal starts to expose molecular occurrences named by stable OTU identifiers



CELEBRATING 15 YEARS IN SERVICE





UNITE 2003 (UNITE ITS SEQUENCES

	and the			*	U		:+		
()							IE	2	
1			LA	melecular da	tabase for the identi	fication of a	ctores central	cal funga	
	fenne	About	database		Primer motes	Cor	itributors	Acknow ledgements	
		CEEGG TGCAGC TAATGT GCTATT TGATGA GCTTGC ACCACC GACACC	ICATGTGCA CTGGGGGG GAATTGCAA CCGTGGAC ICAGTGTTTC TAACCTTCI CTTTTTGAAC	CGCTICTGTTT CTCTGGCCCC (AATTCAGTG ATGCCCCTGTT TTTGGCGCTG (GTGTGGTAT CCCACAAAI CGTTTGATCT	IGCACATCCATTCAC CCTGCCGTGGTTCT/ IAATCATCGAATCTT IGAGTATCATCGAACA TTTGCTGCCTGCGC ICATGGGTGTGGTATAA SCTTCGCTGCGGCCT ICAAATCACGTAGGA	ACCTGTGC ATGTATTTA TGAACGCA ICCTCAACTI ICCTCAACT ICTACCTCC ICTATCTACI ICTACCCGC	ACCETETGTA CACACACACA CETTGEGECE ETCATGGCTT TETCAAATGA GETTGTGGGTT CTCTCCTCCC TGAACTTAAG	RESET ATAAG TTTG GCCA MATTA TTCC GTCAG CC	
		ice#LAS	T (belp)	and cho	ose relevant m	tethod fo	r the analy	yses: blasm	
Analy	sis method	Output to	Evalue	Action	Analysis method	Output to	Action	BLAST Evalue	Action
	1	 Screen 	 Unique 	Submit	O NJ	⊖ Screen	Submit	BLAST e value: 1 (default)	(Submit)
() P	arximony .	O File	O AII		O Parsimony	⊖ File			

UNITE 2005 (UNITE + INSDC ITS SEQUENCES)



UNITE 2010 (COMMUNICATION SYSTEM FOR SPECIES - NAMES OR NO NAMES AVAILABLE)



UNITE 2018 (SPECIES HYPOTHESES - IDENTIFICATION AND COMMUNICATION SYSTEM)





UNITE - SPECIES HYPOTHESES (SH)





UNITE - SPECIES HYPOTHESES (SH)





UNITE - SPECIES HYPOTHESES (SH)



UNITE - EVERY SH HAS DOI (DIGITAL OBJECT IDENTIFIER)





PLUTOF - DATA MANAGEMENT AND PUBLISHING PLATFORM



Workbench Modules

Projects

Manage all your projects online and connect them to different data and datasets. Every project may include unlimited number of localities and associated data. It may be public, private or shared with selected workgroups.

Collections

Biological collections may create and manage their digital archives. There are specific modules for the printing of labels, managing of loans, compiling of reports, publishing of datasets in GBIF or other places. Types of the collection datasets include preserved or living specimens and environmental samples.

Monitoring and Conservation

PlutoF may also help in creating and managing your projects of monitoring and nature conservation. Single project may combine different records of taxon occurrences, human and DNA based observations, samples, specimens, references, etc. Datasets

File Repository

Most data may easily be linked to the files which are uploaded and managed through the file repository module. For example, images or videos may be linked to the sampling area, observation, specimen, reference, institution, person, project, etc. Sound, raw DNA sequences, digitized notebooks are also good examples of the files managed through this module.

Analyses

Specific analyses can be run online and the number of analyses accessible is growing. Examples of current analyses include sequence analyses with different programs (e.g. ITSx, ATOSH, massBLASTer).

Publishing

Here you can ask for a new Digital Object Identifier (DOI) for your dataset. Datasets may be published via GBIF also. Data of references and keywords is managed







DU are browsing: UNITE - Unified system for the DNA based ingal species linked to the classification OVERVIEW VERBATIM	sp.)
	incation -
Ingoin Fungi : hylum Ascomycota This is the interpretation of the species as published in UNITE - Unified system for the DNA ballinked to the classification. lass Dothideomycetes issues: trank unknown, could not be matched to GBIF backbone inder Capnodiales CITATION shked SH200601.07FU (Capnodiales sp.) SH200601.07FU (Capnodiales sp.) in Natural History Museum, University of Tartu (2017). UNIT the DNA based fungal species linked to the classification. Checklist Dataset https://doi.org/10	TE - Unified system for
No children	.su/z/dorayq

GBIF

	Classification	
You are brows fungal specie	rsing: UNITE - Unified system for the DNA based as linked to the classification	SPECIES SYNONYM SH246432.07FU (Amauroderma schomburgkii) in UNITE - Unified system for the DNA based fungal species linked to the classification Synonym Of Amauroderma schomburgkii (Mont. & Berk.) Torrend, 1920
Kingdom	Fungi :	OVERVIEW VERBATIM
Phylum	Basidiomycota	This is the interpretation of the species as published in UNITE - Unified system for the DNA based fungal species
Class	Agaricomycetes	linked to the classification. issues: rank unknown could not be matched to GBIF backbone
Order	Polyporales	CITATION
Family	Ganodermataceae	OUO 46 400 OTELL (A many dama a sharehumbil) in Natural Ulatan (Aturaum University of Tarty (2017), UNITE
Genus	Amauroderma Murrill, 1905	Unified system for the DNA based fungal species linked to the classification. Checklist Dataset https://doi.org/10.5072/d5rayq accessed via GBIF.org on 2017-12-07.
Species	Accepted Name Amauroderma schomburgkii (Mont. & Berk.) Torrend, 1920 Synonym = SH246432.07FU (Amauroderma schomburgkii)	



DU are browsing: UNITE - Unified system for the DNA based ingal species linked to the classification OVERVIEW VERBATIM	sp.)
	incation -
Ingoin Fungi : hylum Ascomycota This is the interpretation of the species as published in UNITE - Unified system for the DNA ballinked to the classification. lass Dothideomycetes issues: trank unknown, could not be matched to GBIF backbone inder Capnodiales CITATION shked SH200601.07FU (Capnodiales sp.) SH200601.07FU (Capnodiales sp.) in Natural History Museum, University of Tartu (2017). UNIT the DNA based fungal species linked to the classification. Checklist Dataset https://doi.org/10	TE - Unified system for
No children	.su/z/dorayq

GBIF

GBIF - NEWS



NEWS 29 AUGUST 2018

Adding sequence-based identifiers to backbone taxonomy reveals 'dark taxa' fungi

Pilot project with northern European researchers enables inclusion of non-Linnaean 'species hypotheses' aimed at advancing scientific understanding of mycology and functional biodiversity



Hygrocybe conica, observed in Trondheim, Norway by Ole Reitan, via Norwegian Species Observation Service. Photo licensed under CC BY 4.0.

Until a few weeks ago, the GBIF backbone taxonomy fit snugly within a traditional model, classifying and ranking organisms' names using the system that Carl Linnaeus first outlined in *Systema Naturae* in 1735. By combining name-based information from dozens of different authoritative sources like the Catalogue of Life, IRMNG and the World Register of Marine Species (affectionately known as 'WoRMS'), the backbone provides a consistent means of organizing all species-related content on GBIF.org-like



GBIF - DANISH BIOWIDE PROJECT UPLOADED FIRST DNA DATASET WHERE SPECIES ARE COMMUNICATED VIA UNITE SH DOI





GBIF: COURSE IN COPENHAGEN JANUARY 2019 IDENTIFICATION AND PUBLISHING HTS/SANGER DNA SEQUENCE DATASETS



EXISTING DATA MODEL AND DNA EVIDENCE

Ŧ

Ŧ

	<pre>"associatedSequences": "KP194575",</pre>
basisOfRecord:	"MATERIAL_SAMPLE"
country:	"Denmark"
dataGeneralizations:	"100% identity match with species hypothesis"
decimalLatitude:	"56.06375636169536"
decimalLongitude:	"9.731862068229827"
eventDate:	"2014-10-16"
eventID:	"BIOWIDE-ES066"
extensions:	
<pre>whttp://data.ggbn.org/schemas/ggb</pre>	n/terms/Amplification:
▼0:	
consensusSequence:	"TAATCTCTCAACTTTTGGTCTTTTGGTTTCTTCTTAACCCAAGGCCTGGAGTTTGGATGGTGGAGGTGTGCTGG
marker:	"ITS2"
footprintWKT:	"POLYGON((9.731601299999966 56.0639614,9.731487457013827 56.063607736393486
id:	"14193"
kingdom:	"Fungi"
locality:	"Odderholm"
nameAccordingTo:	"http://dx.doi.org/10.15156/BIO/SH221291.07FU"
occurrenceID:	"14193"
preparations:	"DNA extract"
recordedBy:	"Tobias Frøslev"
samplingProtocol:	"eDNA sampling from soil"
scientificName:	"SH221291.07FU"
taxonConceptID:	"SH221291.07FU"
Note that defaults The same in-a parameters is presented as if days, 190-190, you address: Investor the approximation is presented as if days, 190-190, you address: the same is the approximation is and an address is the same is a same interval table in the days is a same interval table interval table in the days is a same interval table interval table in the days is a same interval table interval table in the days is a same interval table in the days is a same interval table int	<pre>"scientificName": null, "sex": null, "stateProvince": "Queensland" "taxonID": "BOLD:AAF9888"</pre>
	Le GBIF



LATIN IS RULED BY THE CODES

INTERNATIONAL CODE OF NOMENCLATURE FOR ALGAE, FUNGI, AND PLANTS (SHENZHEN CODE)

2018





International Commission on Zoological Nomenclature

INTERNATIONAL CODE OF ZOOLOGICAL NOMENCLATURE

Fourth Edition

adopted by the International Union of Biological Sciences







THINGS CHANGE



THE FIRST MICROSCOPES











Wikipedia; Museo Galileo/Institute and Museum of the History of Science; Wellcome Library, London; Granger/NYC; Discover, ABI, Roche

DNA IS NEW (?) MICROSCOPY



Quarterly journal of microscopical science 1853, 1857







MULTIPLE WAYS TO DOCUMENT BIODIVERSITY





Latin: Linnaean nomenclature

OTU codes

Content of Content of Content and Content and Content of Content



Images: Noah Strycker, Biodiversity Heritage Library, Dmitry Schigel

GOING BEOYND FUNGI

BOLD: Trondheim 20195.7MNCBI: EBI and ELIXIR4BnSILVA – bacterial OTUs6MR syst – Diatoms??



Suggest **key reference libraries of OTUs**: New names -> indexing dark biodiversity



THANKS

Identification and publishing HTS/Sanger DNA sequence datasets

Copenhagen, Denmark 14–15 January 2019



Global Biodiversity Information Facility





https://sisu.ut.ee/publish-sequence-datasets

