2018  981,464,491  →  2020  1,602,089,347
DNA-DERIVED DATA IN GBIF
GUIDANCE AND TRAINING

Dmitry Schigel | Scientific officer
A WINDOW ON EVIDENCE ABOUT WHERE SPECIES HAVE LIVED, AND WHEN

- Observations
- Literature
- Remote-sensing
- Digitized specimens
- Data publishing and indexing
- Common standards (DwC)
- Environmental DNA

Data discovery and use
DATA STREAMS IN GBIF

**DNA-based evidence**

- Sequence names
  - "taxon hypotheses", OTUs
  - Indexing dark taxa

**DNA occurrences**

- Indexed described taxa

**Natural history collections**

**Human observations etc.**

**Latin names**

- UNITE: SH
- BOLD: BINs
- SILVA
- ...
Sequence names
“taxon hypotheses”, OTUs

DNA-based evidence
natural history collections
human observations etc.

Indexing dark taxa

Indexing described taxa

DNA occurrences

species mixes (samples)

Indexing

individuals / isolates

Amplicon (barcoding)
one or more marker genes

Genomic shotgun DNA

Transcriptomic shotgun RNA

Amplicon (metabarcoding)
one or more marker genes

Metagenomic shotgun DNA

Metatranscriptomic shotgun RNA

UNITE: SH
BOLD: BINs
SILVA
...

UNITE: SH
BOLD: BINs
SILVA
...

MGnify

INSDC / ENA

BOLD occurrences

UNITE
...

MGnify

UNITE
BIOWIDE
...

MGnify
DATA STREAMS IN GBIF

DNA-based evidence
- natural history collections
- human observations etc.

Sequence names
- "taxon hypotheses", OTUs
  - Indexing dark taxa
  - UNITE: SH
  - BOLD: BINs
  - SILVA

DNA occurrences
- Indexing described taxa
- sequencing
- individuals / isolates
- species mixes (samples)

Amplicon
- (metabarcoding)
- one or more marker genes
- Genomic shotguns DNA
- Transcriptomic shotguns RNA
- Metagenomic shotguns DNA
- Metatranscriptomic shotguns RNA

Targeted detection
- qPCR
- Norway

UNITE: SH
- BOLD: BINs
- SILVA

MGnify
- INSDC / ENA
- BOLD occurrences
- LEARN
- BIOWIDE
- ...
DATA STREAMS IN GBIF

DNA-based evidence

natural history collections
human observations etc.

Sequence names
“taxon hypotheses”, OTUs

Indexing dark taxa

DNA occurrences

Indexing described taxa

individuals / isolates

specie mixes (samples)

Targeted detection
qPCR

Amplicon (barcoding)
one or more marker genes

Geneomic shotgun DNA

Transcriptomic shotgun RNA

Amplicon (metabarcoding)
one or more marker genes

Metagenomic shotgun DNA

Metatranscriptomic shotgun RNA

UNITE: SH
BOLD: BINs
SILVA ...

INSDC / ENA

CHECKLISTS

BOLD occurrences
UNITE ...

TARGETED DETECTION

qPCR

TOTAL DNA occurrances

DNA occurrences

Species mixes (samples)

Latin names

CHECKLISTS

SILVA

BIOWIDE

TARGETED detection

qPCR

Norway ?

OCCURRENCE DATASETS

SAMPLING EVENT DATASETS

MGNify

MGNify

MGNify

UNITE BIOWIDE ...

MGNify
OPERATIONAL TAXONOMIC UNITS

OTU = SH, Species hypothesis

GBIF backbone taxonomy

SH ABC0001

OTU = BIN, Barcode identification number

BIN DEF0002
INDIVIDUAL SEQUENCES WITH COORDINATES

European Nucleotide Archive: 4.8M records

International Barcode of Life: 6.8M records
METABARCoding: MGnify and More

920 datasets: 20.1M new records
14.8M Bacteria
2.3M Chromista

Description: MGnify is a free-to-use analysis platform and database of analysed microbiome sequencing projects, hosted by the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI). Microbiome research typically involves the study of all genomes present within a specific environment, such as soil, seawater or a human body site. Underpinned by dramatically mainstream. The approach can provide unique insights into the complex of surroundings, each other and, in some cases, their host. MGnify offers an diversity and functional/metabolic potential of environmental samples. U held within the database. In addition, users can request analysis of any up. Following publication: EBI Metagenomics in 2017 enriching the analysis o Mitchell, Maximim Scheremetjei, Hubert Denise, Simon Porter, Aleksandra T Petra ten Hoope, Blaise Alako, Clara Arriz, Darren J. Wilkinson, Thomas F Endorsed by: National Biodiversity Network

Administrative contact: Robert Finn
Technical contact: MGnify Helpdesk

Country or area: United Kingdom of Great Britain and Northern Ireland

https://www.gbif.org/publisher/ab733144-7043-4e88-bd4f-fca7b7f858880
NEW GUIDE: PUBLISHING DNA-DERIVED DATA THROUGH BIODIVERSITY DISCOVERY PLATFORMS

Mapping and data publishing
Cross-platform
About 40 pages long "cookbook"
  ❖ Introduction
  ❖ Categorization
  ❖ Mapping*
  ❖ Visuals
  ❖ Future prospects
  ❖ Resources

*Darwin Core and MiXS based

https://docs.gbif-uat.org/publishing-sequence-derived-data/1.0/en
Introduction

- Rationale
- Audiences
- DNA derived occurrence data
- Biodiversity data
- Processing workflows
- Taxonomy of sequences

https://docs.gbif-uat.org/publishing-sequence-derived-data/1.0/en
PUBLISHING SEQUENCE-DERIVED DATA: THE “DO” SECTION

Data packaging and mapping

Categorisation of data

I  Sequence-based occurrences
II  Enriched occurrences
III  Targeted species detection (qPCR)
IV  Name references
V  Metadata-only

Mapping

Examples

https://docs.gbif-uat.org/publishing-sequence-derived-data/1.0/en
NEW TRAINING COURSE ON GBIF AND BOLD SKILLS – WITH BIODATA /GBIF NORWAY

Accelerating biodiversity research through DNA barcodes, collection and observation data

Work in progress

Why publish in GBIF?
Barcoding, data and biodiversity
Biodiversity databases in Georgia
Open data as a first-class research citizen
Databases as a research tool
Why do we need to identify species
Concept of DNA barcoding
Data in and data out: recognize and understand your data needs
Principles of data organization and personal data management
Data structure: standards
Data citation
Data exposure: why and when
Data papers
Barcode reference repositories
Quality control
Use case Bombus and legumes: uncovering pollination mysteries

KEY FACTS

- New GBIF course
- Tbilisi, Georgia
- 2021, 4 days
- Onsite or virtual
- GBIF and BOLD
- publish and use
- DNA-derived FAIR data
- Learn and practice

https://www.nhm.uio.no/english/research/projects/biodata/activities/202007-tbilisi.html
THANK YOU

Dmitry Schigel
dschigel@gbif.org