Improving the link between DNA data and museum specimens

Mathieu Perret

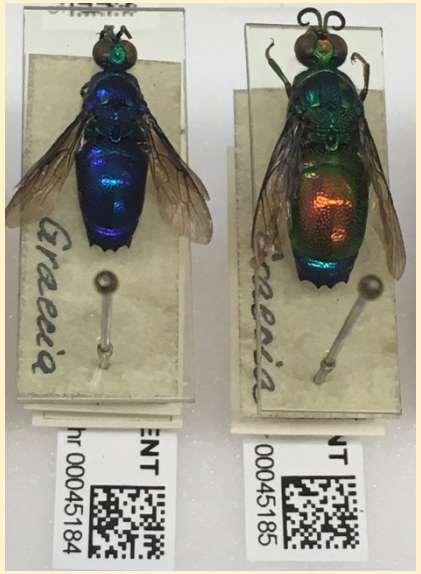
SwissCollNet Annual Workshop 2023
Bern, 20 January 2023

Conservatoire et Jardin botaniques Genève

scnat
swiss academy of sciences

The Swiss Natural History collections: a source and reference for genetic data

Swiss Natural History Collections
= 60 M specimens in CH



DNA

Genetic studies / museomics

Tree of life & systematics

Barcoding of life (SwissBOL, iBOL, BIOSCAN)

Monitoring of genetic diversity (DECLINE, GenDiv, Phylogenomics of GE flora)

Genomics - European Reference Genome Atlas (ERGA), Biodiversity Genomics Europe project

GBIF / SwissCollNet

GBIF – iBOL
GBIF - INSDC Sequences
GeOMe

BOLD / GBIF.ch (DNA barcodes)
EMBL/NCBI /ENA (raw data)
DRYAD and others (processed data)

The questions we want to solve

DNA data linked to specific specimens or collections:

- *How many type specimens hosted in a Swiss collection have been sequenced?*

DNA data linked to specific locality or regions:

- *What are the genetic data available for the alpine fauna? Where are the gaps?*

DNA or tissues available:

- *Are there DNA extractions available for Leontopodium alpinum in CH?*

→ There is no current easy/automatic way to answer these questions

Example of the broken link between specimens and DNA data

Specimen in G

Litterature

Received: 14 September 2021 | Revised: 7 April 2022 | Accepted: 19 April 2022
DOI: 10.1111/geb.13521

RESEARCH ARTICLE

Global Ecology and Biogeography | WILEY

Recent and local diversification of Central American understorey palms

Ángela Cano[1,2,3,4] | Fred W. Stauffer[2,3] | Tobias Andermann[4,5] | Isabel M. Liberal[6] | Alexander Zizka[7,8] | Christine D. Bacon[4,5] | Harri Lorenzi[9] | Camille Christe[3] | Mats Töpel[4,5] | Mathieu Perret[2,3] | Alexandre Antonelli[4,5,10,11]

Supplementary Tables

Table S1.1 Sampling. Palm samples included in our phylogenetic analyses and their reference specimens in botanical institutions (herbaria in bold, living collections plain text) whose acronyms are defined as: CJB - Conservatoire et Jardin botaniques de la Ville de Genève; **COL** - Colombia National Herbarium; DH - Donald Hodel's private collection; FTBG - Fairchild Tropical Botanical Garden; **FTG** - Fairchild Herbarium; **G** - Herbarium of Geneva; JBP - Jardim Botânico Plantarum; JBQ - Jardín Botánico del Quindío; **K** - Herbarium of Royal Botanic Gardens, Kew; MBC - Montgomery Botanical Center.

Subfamily - Tribe - Subtribe *Species*	Species Author	Collector	Collector and/or accession number	Botanical institution
Arecoideae - Chamaedoreeae				
Chamaedorea adscendens	(Dammer) Burret	Cano, Á.	99680*A/ACS312	MBC/**FTG**
Chamaedorea allenii	L.H. Bailey	Bernal, R. et al.	RB4948	**COL**
Chamaedorea anemophila	Hodel	Cano, Á. et al.	ACS426	**G**
Chamaedorea benziei	Hodel		2000822E	FTBG/**FTG**
Chamaedorea brachypoda	Standl. & Steyerm.		P442B	FTBG
Chamaedorea cataractarum	Mart.	Cano, Á.	7248D/ACS321	MBC/**FTG**
Chamaedorea correae	Hodel & N.W. Uhl	Cano, Á. et al.	ACS411	**G**
Chamaedorea costaricana	Oerst.	Cano, Á. et al.	ACS354	**G**
Chamaedorea costaricana	Oerst.	Cano, Á. et al.	ACS356	**G**

DNA data (SRA)

NIH National Library of Medicine
National Center for Biotechnology Information

Sequence Read Archive | Search | Run Browser | Analyses | Study | Provisional SRA

Run Browser › SRR8983928

NGS Sequence Capture of Central American Palms (SRR8983928)

Metadata | Analysis | Reads | Data access | FASTA/FASTQ download

Run

Run	Spots	Bases	Size	GC Content	Published	Access Type
SRR8983928	56.3k	455.3M	186.6M	43.3%	2020-04-19	public

Quality graph (bigger)

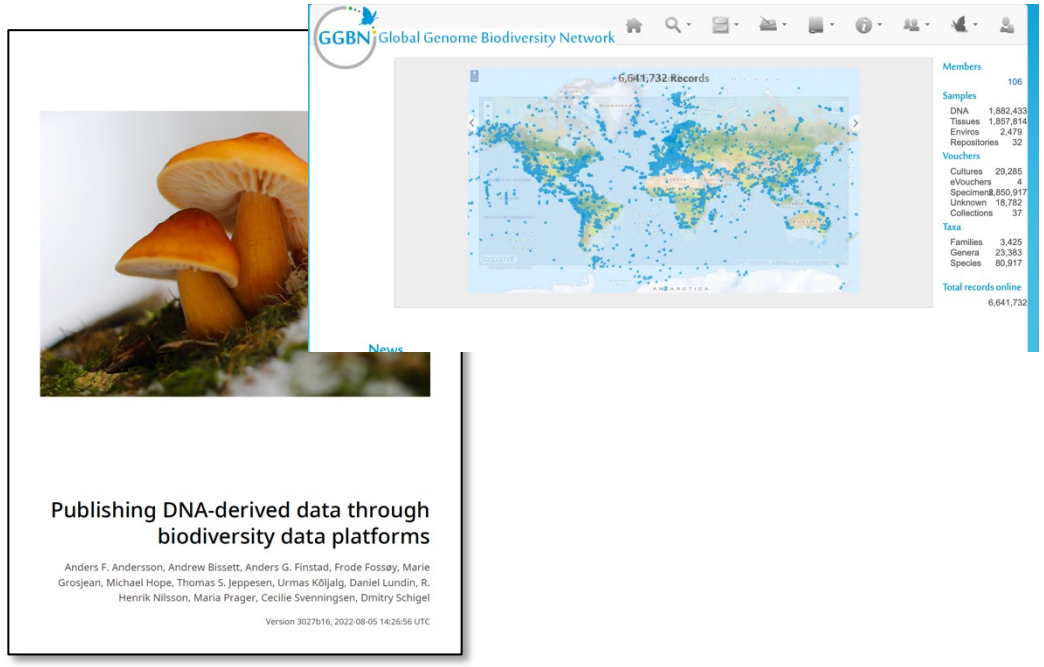
This run has 2 reads per spot:

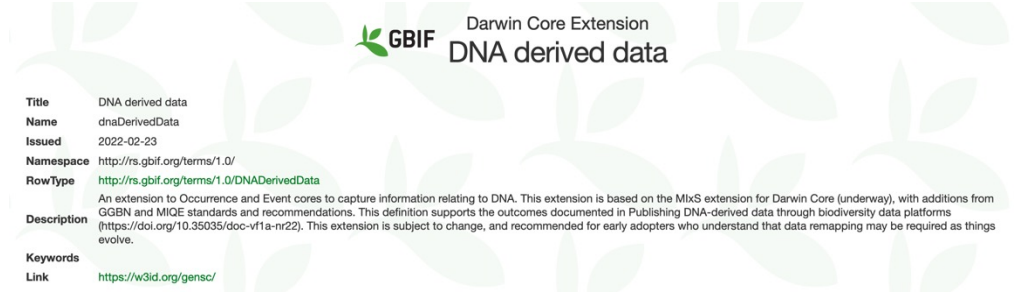
L=301, 100% | L=301, 100%

Legend

Existing/proposed solutions

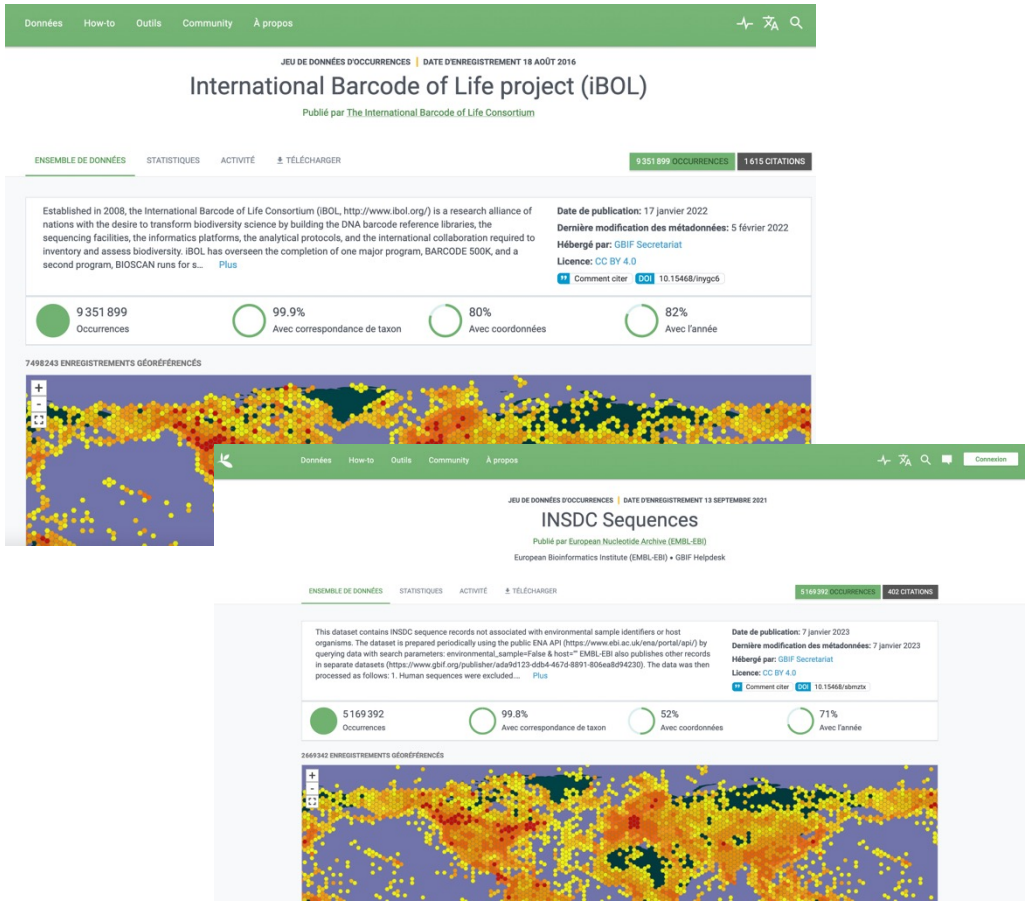
- Guidelines prepared by the GBIF and GGBN
- Dwc standards for DNA-derived data





Existing/proposed solutions

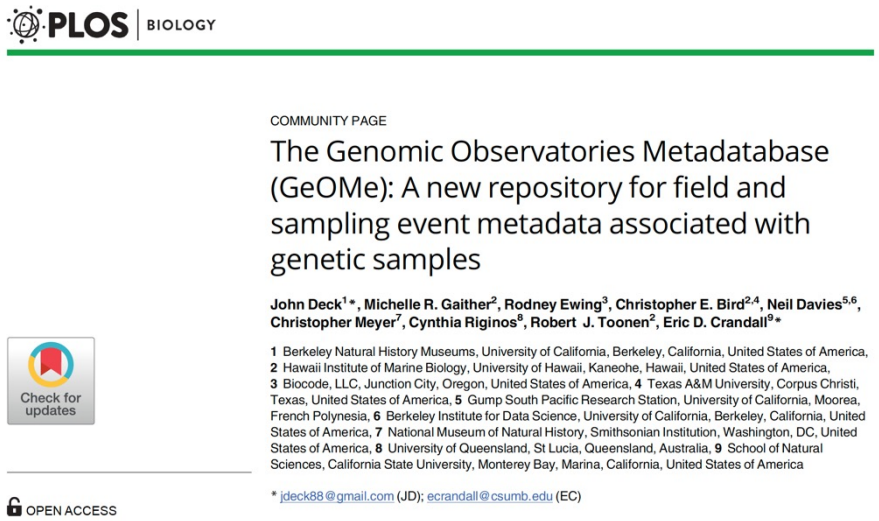
- Guidelines prepared by the GBIF and GGBN
- Dwc standards for DNA-derived data
- Access to DNA-derived data through GBIF.org



Existing/proposed solutions

- Guidelines prepared by the GBIF and GGBN
- Dwc standards for DNA-derived data
- Access to DNA-derived data through GBIF.org
- Genomic Observatories Metadatabase (GeOMe)
- GenDIB (WSL) – mapping genetic diversity in CH

→ Need to adopt same standards and vocabulary!!!


PLOS BIOLOGY

COMMUNITY PAGE

The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples

John Deck[1]*, Michelle R. Gaither[2], Rodney Ewing[3], Christopher E. Bird[2,4], Neil Davies[5,6], Christopher Meyer[7], Cynthia Riginos[8], Robert J. Toonen[2], Eric D. Crandall[9]*

1 Berkeley Natural History Museums, University of California, Berkeley, California, United States of America, 2 Hawaii Institute of Marine Biology, University of Hawaii, Kaneohe, Hawaii, United States of America, 3 Biocode, LLC, Junction City, Oregon, United States of America, 4 Texas A&M University, Corpus Christi, Texas, United States of America, 5 Gump South Pacific Research Station, University of California, Moorea, French Polynesia, 6 Berkeley Institute for Data Science, University of California, Berkeley, California, United States of America, 7 National Museum of Natural History, Smithsonian Institution, Washington, DC, United States of America, 8 University of Queensland, St Lucia, Queensland, Australia, 9 School of Natural Sciences, California State University, Monterey Bay, Marina, California, United States of America

* jdeck88@gmail.com (JD); ecrandall@csumb.edu (EC)

Check for updates

OPEN ACCESS


< Zurück

Machbarkeitsstudie für eine neue nationale Datenbank zur genetischen Vielfalt in Populationen wildlebender Arten

GenDiB

Hintergrund: Die genetische Vielfalt ist ein wesentlicher Bestandteil der biologischen Vielfalt und gilt als Schlüssel für den Fortbestand von Populationen in einer sich verändernden Umwelt sowie für die Anpassung an extreme Ereignisse. Das weltweite Übereinkommen über die biologische Vielfalt (CBD) verpflichtet die Länder, den Verlust der genetischen Vielfalt zu bewerten, zu überwachen und letztendlich zu stoppen. Es gibt jedoch keine systematische Erfassung innerartlicher genetischer Daten, die auf die Bedürfnisse der Naturschutzpraxis zugeschnitten ist und diesen Prozess unterstützen könnte.

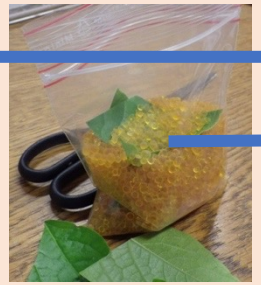
Ziel: Das WSL-interne Projekt GenDiB führt eine Bedarfs- und Machbarkeitsanalyse über eine neue nationale Datenbank mit georeferenzierten Daten zur intraspezifischen genetischen Vielfalt in Populationen wildlebender Arten in der Schweiz durch.

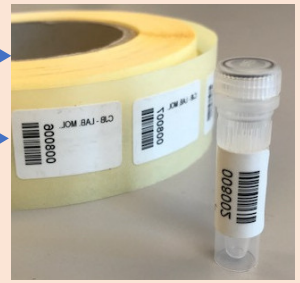
Herbier - G

*CatalogueNumber
(e.g., G00213456)

Info Flora / Info Fauna ID (CH)

SIBG
→ GBIF.ch / org





Tissue

Tissue identifier
*Tissue_materialSampleID
Tissue_basisOfRecord
Tissue_materialSampleType
Tissue_preparationType
Tissue_institutionCode
Tissue Preservation
DNA_preservationType
DNA_preservationTemperature
DNA_preservationDateBegin

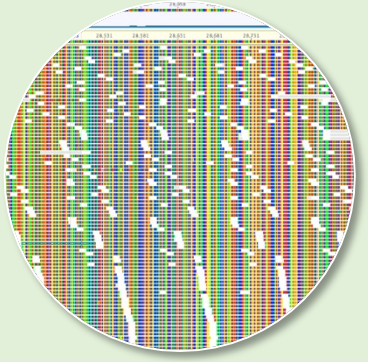
DNA extracts

DNA sample identifier
*DNA_catalogNumber
DNA_basisOfRecord
DNA_materialSampleType
DNA_preparationType
DNA_institutionCode
DNA Extraction
DNA_preparationDate
DNA_preparedBy
DNA_preparationMaterials
DNA Preservation
DNA_preservationType
DNA_preservationTemperature
DNA_preservationDateBegin
DNA_preservationRoom
DNA_preservationFreezer
DNA_preservationBox
DNA_availableInStock
DNA Quality
DNA_ratioOfAbsorbance260_280
DNA_ratioOfAbsorbance260_230
DNA_concentration
DNA_qualityCheckDate
DNA_qualityRemarks
DNA_volume
DNA_gelIdentifier

DNA – bank
→ GGBN
→ GBIF.ch/ org

PCR & Library

Library Identifier
*library_ID
Library metadata
library_title
library_layout
library_strategy
library_source
library_selection
library_preparedBy
library_preparationDate
library_concentration
Library_sonicationTreatment
library_weightOfDnaUsed
library_protocol
library_fragmentSize
library_fragmentSizeImage
library_success
library_baitsSetName
library_baitsSetReference
Library preservation
library_preservationType
library_preservationTemperature
library_preservationFreezer
library_preservationBox
library_preservationBoxPosition
library_availableInStock



Sequencing

Sequence Identifier
*seq_ID
Sequencing Method
platform
instrument_model
Sequencing Details
seq_preparationDate
seq_preparedBy
project_name
numberOfReads

Link to NCBI/EMBL/ENA/BOLD entry (raw data)
BOLDProcessID
geneticAccessionNumber (SRR)
associatedSequences

DNA – data
→ GBIF.ch/ org
→ ENA – EMBL

Next steps

- Complete the databasing of specimens used for genetic studies at G
- Structure our DNA data and fill in the tables !
- Define the mapping and transfer the data to GBIF.ch and GGBN
- Develop a *Botalista* module to facilitate the data capture and management
- Share the database structure with other users
- ? Integrate DNA-derived data within the « Swiss data aggregator » and the SVNHC ?

Acknowlegments

Guillaume Antonioli

Raoul Palese, Andreas Ensslin, Anouk Mentha, Yamama Naciri, PhyloLab team

Sofia Wyler, Pascal Tschudin