SWISS GENETIC DATA FOR THE BIODIVERSITY INVENTORY

MANDATORY AND RECOMMENDED FIELDS FOR SHARING DATA WITH GBIF.CH

July 2019

INTRODUCTION

The genetic data module developed within the GBIF Swiss node is an extension of the database designed for the compilation and provision of biodiversity data available in Switzerland. The GBIF.ch system offers optimal conditions for the storage and management of genetic data related to the Swiss inventory of biodiversity, and supports the links of these genetic data (publicly available or not) to the general biodiversity information usually processed by the node.

The purpose of the suggested procedure is to allow synergies with the most common modules used to manage biodiversity data in our country, and to provide users with an interface to integrate DNA data with traditional biodiversity databases that offer accuracy of taxonomic and voucher information or any other specimen documentation. The proposed scheme facilitates the retrieval and exchange of structured information, describing data related to the inherited genetic components of a natural specimen.

The elements proposed here use terms that are common to the Data Standards Access to Biological Collection Data (ABCD) and DarwinCore (DwC), both supported by the Global Genome Biodiversity Network (GGBN). The module allows registering repeatable elements related to a single specimen record, such as several DNA extractions and potentially thousands of DNA sequences. The database also allows entering numerous primers for the same marker. Finally, the procedure described hereafter is also prepared for the diffusion of the genetic data and information into the international networks, and we recommend the user to map as many elements as possible (four languages are supported – French, German, Italian and English).

THE DATA

Genetic data stands for the nucleic acid sequence (order of nucleotides) within a deoxyribonucleic acid (DNA) molecule. It results from the analysis of a biological sample that represents the entire specimen. Documenting these data with useful information is fundamental for the correct processing and the subsequent diffusion and generalised use of the data. Part of this documentation implies providing information on the methodology employed to obtain the DNA and produce the sequences. Such information allows to retrace every step of the process and to evaluate the quality of the data submitted to the GBIF Swiss Node.

The genetic data are organised by projects, each project grouping a data set from a common study. Documentation of the project follows infoSpecies' deontology and the user defines the dissemination date for the data within the project.

The genetic module integrates data related to the DNA with the information of the underlying biological specimen, i.e. the original occurrence in space and time, the taxonomical determination, voucher depository, and ecological information. In order to support the association among varied genetic data and to ensure the link to the general biodiversity information, the definition of a

VOUCHERID is obligatory. This element is therefore common to every table described in the sections A and B.

Adding a barcode tag (unique specimen identifier on label) in the moment of data capture permanently associates voucher specimen and electronic record, and hence facilitates the linkage with new, successively generated data.

BEST PRACTICE

The VOUCHERID is the identifier that allows the linkage between all genetic data and the association to the general biodiversity information. Hereafter, the order of preference to define the VOUCHERID:

1) VOUCHERID = BARCODE TAG ; unique specimen identifier on label provided by GBIF.ch
2) VOUCHERID = INFOSPECIES ID ; occurrenceID provided by a data centre
3) VOUCHERID = MUSEUM ID ; catalogNumber provided by the institution
4) VOUCHERID = FIELD ID ; arbitrary identifier attributed temporarily (chorological data must be provided)

A. DNA EXTRACTION

A.1 Indications about the organic sample

Organic sample stands for the biological material from which DNA is extracted – leaf, blood, muscle, hair, buccal swab. It is hereafter generalised as TISSUE.

REQUIRED INFORMATION

- TISSUETYPE → Type of organic material at the origin of DNA extraction

In the case of availability of the tissue sample, also provide:

- TISSUEINS → Institution where the tissue is stored
- TISSUEID → Number/code/identifier attributed by the institution to the tissue sample

A.2 Indications about the genetic sample

Genetic sample stands for the biological material related to DNA.

REQUIRED INFORMATION

- DNATYPE → Type of DNA extracted

In the case of availability of the DNA sample, also provide:

- DNAINS → Institution where the DNA is stored
- DNAID → Number/code/identifier attributed by the institution to the DNA sample

A.3 Indications about the extraction of the DNA

This section concerns the methodology and protocols used in the lab for the obtention of the DNA. The information refering to these fields are grouped under the prefix EXT.

REQUIRED INFORMATION

- EXTID → Number/code/identifier attributed by the lab to the DNA extraction
- EXTINS → Institution where the DNA was extracted
- EXTCONTACT → Email of the person responsible for the DNA extraction
- EXTMETHOD → Commercial kit or protocol used to extract the DNA
- EXTMETHODSUP → Supplier of the kit or bibliographic reference of the protocol
- EXTSTAFF → Person who performed the DNA extraction
- EXTYEAR → Date when DNA extraction was performed

B. SANGER SEQUENCING

B.1 Indications about the DNA sequence

The DNA sequence stands for the sequence of nucleotides obtained for the targeted region of the genome. It is derived from the assemblage of at least two single reads or from the edition of a sole single read. This element is hereafter generalised as SEQ.

REQUIRED INFORMATION

- SEQGENOME → Organelle source of the sequence
- SEQMARKER → Fragment of the genome or locus targeted
- SEQNCBI → Accession number given by GenBank to the sequence
- SEQPROCESSID → BOLD identifier attributed to each specimen entry
- SEQBIBLIOREF → Bibliographic reference of the article where the sequence was published
- SEQBIBLIOREFDOI → Digital identifier associated to the published article

In the case of having different DNA extractions for the same specimen, also provide:

- EXTID → Number/code/identifier attributed by the lab to the DNA extraction
- EXTINS → Institution where the DNA was extracted

BEST PRACTICE

The format accepted for the DNA sequence files is fasta (.fasta, .fst, .fas). Two possibilities exist to construct the files: 1) one file per individual edited sequence or 2) one file per marker regrouping all the edited sequences obtained for that marker under this project. Hereafter, the mandatory procedure to name the files and the sequences:

1) File Name = VOUCHERID_SEQMARKER
2) File Name = SEQMARKER; Header of each individual sequence = VOUCHERID

In the case of having different sequences from different DNA extractions for the same specimen, name your file and sequences as follows:

3) File Name = EXTID_SEQMARKER
4) File Name = SEQMARKER; Header of each individual sequence = EXTID

B.2 Indications about the single reads

Single reads refer to the chromatogram or trace files issued from Sanger sequencing. The fields that document the files and the sequencing process are grouped under the prefix TRA.

REQUIRED INFORMATION

- TRAFILENAME → Name of the chromatogram trace file of the individual sequence (.ab1)
- TRAMARKER → Fragment of the genome or locus targeted
- TRAPRIMER → Name of the primer used for the sequencing
- TRAINS → Institution where the DNA was amplified and sequenced
- TRAYEAR → Date when DNA individual sequence was produced
- TRAPCRPRIMER1 → Name of the forward primer used in the amplification reaction
- TRAPCRPRIMER2 → Name of the reverse primer used in the amplification reaction

BEST PRACTICE

The format accepted for the DNA trace files is .ab1. Consider naming the files as:

1) VOUCHERID_TRAMARKER_TRAPRIMER

In the case of having different traces obtained for the same marker, but for different DNA extractions of the same specimen, name your files as follows:

2) EXTID_ TRAMARKER_TRAPRIMER

B.3 Indications about the primers

This section concerns the documentation of the primers used for sanger sequencing and PCR amplification. Information is gathered under the prefix PRI.

REQUIRED INFORMATION

- PRINAME → Name given to the primer by the authors who published it or designed it
- PRIMARKER → Fragment of the genome or locus targeted
- PRIDIRECTION → Forward or reverse direction of the primer
- PRISEQUENCE → DNA sequence of the primer 5'–3'
- PRIREFERENCE → Bibliographic reference of the article where the primer was published
- PRIREFDOI → Digital identifier associated to the published article