A GENOMIC LOOK INTO THE EVOLUTION AND BIOGEOGRAPHY OF THE ICHTHYOFAUNA OF THE ALBERTINE RIFT

A Data Management Plan created using DMPonline.be

Creators: Manon Mireille Geerts, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Manon Mireille Geerts

Project Administrator: n.n. n.n.

Grant number / URL: 11Q4724N

ID: 205877

Start date: 01-11-2023

End date: 24-10-2027

Project abstract:

Optimizing bioinformatic pipelines for the processing of high-throughput genomic data to explore questions related to ecology and biodiversity.

Last modified: 27-03-2024

A GENOMIC LOOK INTO THE EVOLUTION AND BIOGEOGRAPHY OF THE ICHTHYOFAUNA OF THE ALBERTINE RIFT

DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

A GENOMIC LOOK INTO THE EVOLUTION AND BIOGEOGRAPHY OF THE ICHTHYOFAUNA OF THE ALBERTINE RIFT

GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

A GENOMIC LOOK INTO THE EVOLUTION AND BIOGEOGRAPHY OF THE ICHTHYOFAUNA OF THE ALBERTINE RIFT

Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

During the research project, I will primarily focus on the reuse of genomic data (short paired-end data, <10TB). This process involves utilizing the WICE cluster at the Flemish Supercomputing Center (VSC) for the computational aspects of my work. The genomic data can be complemented by datasets publicly accessible through the European Nucleotide Archive (ENA). Additionally, I will source reference genomes from the NCBI Genome database to facilitate the assembly of raw data in a reference-guided manner.

To promote transparency and reproducibility in my research, the scripts (.slurm, .py, .R and .sh; <1KB) developed for data processing will be shared publicly on my personal GitHub page (https://github.com/GeertsManon/). Finally, I am committed to publishing my findings in open-access journals whenever possible to enhance the visibility and impact of my research.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

Short-term data preservation is ensured through a personal VSC account (vsc34774, <5TB). For long-term preservation, we will secure both raw and essential intermediate data (BAM, VCF) for 10 years via staging facilities arranged by Prof. Dr. Hugo Gante, located at /staging/leuven/stg_00100/ (30TB). The storage capacity can be easily upgraded after contacting the VSC. I am planning to deposit unpublished raw data to the ENA and annotated genomes to GenBank. All scripts will be deposited on my GitHub (https://github.com/GeertsManon/). Our findings will be published in open-access journals whenever possible. Data deposition in an online database is anticipated within 5 years post-research, adhering to preservation commitments.

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

NA

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

NA

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

NA

A GENOMIC LOOK INTO THE EVOLUTION AND BIOGEOGRAPHY OF THE ICHTHYOFAUNA OF THE ALBERTINE RIFT

FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
genomic data	short paired-end sequences	Generate new data Reuse existing data	Digital	experimental	*.fastq.gz*	<10TB	
genomic data	assembled short paired-end sequences	Generate new data	Digital	experimental	*.bam* *.fasta*	<10TB	
scripts	novel and/or existing tools/scripts for the processing of genomic data	Generate new data Reuse existing data	digital	software	.py .slurm .sh .r	<1KB	
metadata	description of sequenced samples	new data	digital	metadata	.tab	<1MB	
reference genomes	annotated reference genomes from NCBI Genome or NCBI Nucleotide	reused	digital	experimental	.fasta	<1GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

https://www.ncbi.nlm.nih.gov/datasets/
https://www.ebi.ac.uk/ena/browser/home
https://www.ncbi.nlm.nih.gov/nucleotide/

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

My research involves genetic/genomic resources that are captured by the EU Regulation related to the Nagoya Protocol.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

For each sample, a detailed tabulated text file will be generated and stored alongside the raw data in the staging storage. This documentation will encompass critical metadata such as the location of the sample collection, the date of collection, the method used for DNA extraction, specifics of the sequencing process, the total number of reads, etc. Additionally, all scripts written for processing the raw data will be stored at the same locality in the staging storage, ensuring they are easily accessible alongside the corresponding datasets. To further aid in reproducibility and transparency, these scripts will also be archived on my personal GitHub. On this platform, I will provide detailed explanations on the usage of each script, including the necessary steps to replicate my data processing workflow.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

3. Data storage & back-up during the research project

Where will the data be stored?

For long-term preservation, we will secure both raw and essential intermediate data (BAM, VCF) for 10 years via staging facilities arranged by Prof. Dr. Hugo Gante, located at /staging/leuven/stg_00100/ (<30TB).

Short-term data preservation is ensured through a personal VSC account (vsc34774, <5TB). I will also use standard backup provided by KU Leuven ICTS.

How will the data be backed up?

The data will be backed up and secured by the VSC. An additional data backup strategy involves storing the raw data on a dedicated hard drive. This hard drive serves as a physical backup, ensuring a copy of the raw sequencing data is securely preserved offline.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

The staging storage currently offers up to 30 TB of data capacity. Should the project's data storage needs exceed this limit at any point, we can promptly request additional capacity from the VSC, ensuring seamless data management throughout and after the project.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

To ensure data security on the VSC, access is currently restricted to myself and authorized group members with VSC accounts. Despite their access, explicit permissions (rwx) to my data are not granted to them at this stage, safeguarding against unauthorized access or modifications. This setup utilizes VSC's robust access control mechanisms, ensuring data integrity and security.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The anticipated costs for data storage and backup are currently estimated at under 100 euros per year for a capacity of 30TB. These expenses will be met through allocated research project funds or institutional support, ensuring continuous data availability and integrity throughout the research period.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

We will secure both raw and essential intermediate data (BAM, VCF) as well scripts for 10 years via staging facilities arranged by Prof. Dr. Hugo Gante, located at /staging/leuven/stg_00100/ (<30TB).

Where will these data be archived (stored and curated for the long-term)?

This preservation will be achieved by utilizing the staging capabilities of the cluster, with the designated storage directory being /staging/leuven/stg_00100/. Additionally, unpublished raw data will be deposited in the ENA upon drafting of a manuscript, annotated genomes will be made accessible via GenBank and scripts will be made available on my personal Github.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The anticipated costs for data storage and backup are currently estimated at under 100 euros per year for a capacity of 30TB. These expenses will be met through allocated research project funds or institutional support, ensuring continuous data availability and integrity throughout the research period.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

If access is restricted, please specify who will be able to access the data and under what conditions.

NA

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

ENA for raw data and GenBank for annotated sequences

When will the data be made available?

upon publication of research results

Which data usage licenses are you going to provide? If none, please explain why.

No, factual data (gene sequences) cannot be licensed

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

https://orcid.org/0000-0002-5709-3004

What are the expected costs for data sharing? How will these costs be covered?

ENA: No limit for data volume. No costs.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Manon Geerts

Who will manage data storage and backup during the research project?

Manon Geerts

Who will manage data preservation and sharing?

Hugo Gante

Who will update and implement this DMP?

Manon Geerts