Long-read multiomics reveals the evolution and heterogeneity of genomically complex sarcomas DPIA

DPIA

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

- Not applicable

Long-read multiomics reveals the evolution and heterogeneity of genomically complex sarcomas 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)

In this project, we will process sequencing data from primary patient samples. Raw ONT Nanopore signal data in POD5 for will be basecalled to yield sequences in binary unmapped BAM format. PacBio and Illumina sequences will be in uBAM and FASTQ formats. Derived data such as variant calls, personal genomes, (binned) read counts, genomic intervals will generally be contained in flat text files with the corresponding extensions (VCF, FA, TXT, BED).
Code will be written in bash, R, Python, Nextflow as flat text files. Containerized tools and software will be created via Dockerfiles and images (SIF). Manuscripts and presentations will be prepared (DOCX, PPTX, PDF).

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)

To accommodate large volumes of data, we have a multi-tiered storage and archiving system. Raw sequencing data is automatically backed up to encrypted cloud storage within Belgium. An automatic policy pushes data into colder storage over time to limit costs. Data which is actively worked on is kept on a 200TB volume at the Flemish Supercomputing Center. Large raw data requiring occasional access is stored on KU Leuven ManGO/IRODS/VSC Tier-1 to limit retrieval costs. All of these implement redundant storage to prevent data loss. Sequencing data used in publications will be deposited at EGA under controlled access for permanent archiving.

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)

This project involves processing of genetic data. All storage locations (VSC, ManGO/IRODS, Cloud) enforce controlled access via two-factor authentication. At EGA, data is encrypted and a data access committee (VIB/UZ Leuven/UCL) will be responsible for granting access.

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

Data derived from sequencing reads can still be identifiable. We will therefore treat all sequencing-derived data as potentially identifiable by default.

Long-read multiomics reveals the evolution and heterogeneity of genomically complex sarcomas GDPR

GDPR

Have you registered personal data processing activities for this project?

- Yes

Long-read multiomics reveals the evolution and heterogeneity of genomically complex sarcomas
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
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Fiber-Seq	Long-read (PacBio) whole-genome Fiber-Seq performed on matched normal and sarcoma tissue samples from patients (50 + 100)	Generate new data	Digital	Observational	raw: • .bam processed: • flat text files (.txt, .csv, .tsv, .vcf) and b/gzipped versions • metadata (.html, .pdf)	>50TB	
Single-cell multiome	Long-read (Nanopore) single-cell multiome data from 25 selected sarcoma cases	Generate new data	Digital	Observational	raw: • pod5 • .bam processed: • flat text files (.txt, .csv, .tsv, .vcf) and b/gzipped versions • metadata (.html, .pdf)	>50TB	

Spatial transcriptomics	Long-read Nova-ST untargeted spatial transcriptomics data from 3 selected sarcoma cases	Generate new data	Digital	Observational	raw: • pod5 • .bam • .tiff processed: • flat text files (.txt, .csv, .tsv, .vcf) and b/gzipped versions • metadata (html, pdf)	<50TB	
Code & tools	Scripts and code written for data analysis	Generate new data	Digital	Software	Code: • .R, .py, .txt, .config, .yml • Dockerfiles, Snakefiles • Singularity containers .sif	<100GB	
Publications	Text and figures	Generate new data	Digital	Other	.docx, .pdf, .tiff, .svg, .ai, .jpg	<100GB	

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:

The project is based on the new data generated in the host lab.

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.

- Yes, human subject data

Tissue samples were obtained from UZ Leuven. Ethical approval is obtained (S66276)

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.

- Yes

Human genetic data (Fiber-Seq, Single-cell multiome & spatial transcriptomics)

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.

- Yes

MDTAs are in place with collaborators at UZ Leuven and University College London (UCL). These restrict the use and dissemination of the data.

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.

- No

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 sequencing data generation, experiments and protocols are saved in the lab notebook (physical or electronic). Detailed protocols will be published on protocols.io and described in publications. All metadata and pertinent information from the sample preparation and subsequent sequencing is described in lab notebooks. Metadata generated by the sequencing machines is preserved.
Any and all used and generated code, scripts and software will be documented and versioned. Jupyter Notebooks that allow for easy documentation of the code will also be used. Additionally, we use electronic lab notebooks to preserve any pertinent information about the analyses.
The code is saved on VSC (Flemish Supercomputer) and backed up to GitHub. The scripts, algorithms and any software will be described in manuscripts and publically released and versioned on GitHub and Zenodo.
Raw whole genome sequencing data is named as:

- YYYYMMDD_PROJECT_NAME
- YYYYMMDD: Sequencing date
- PROJECT: 3 character project code (in this project, the code is SRC)

NAME: descriptive sample name
All datasets will be accompanied by metadata that is stored in electronic lab notebooks and in samplesheets. We use controlled vocabularies or ontologies when applicable to provide unambiguous meaning, for example:

- Gene Onotology: molecular function, cellular component, and biological role
- ENSEMBL or NBCI identifiers: gene identity
- HUGO Gene Nomenclature Committee: names and symbol of human genes

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.

- Yes

All metadata and pertinent information from the sample preparation, sequencing and computational analysis is described in (electronic or physical) lab notebooks saved on KUL OneDrive.
Metadata will be saved in excel sheets or .csv files on KUL OneDrive. These formats allow for structured metadata and are easily

read by machines.
For sequencing data stored on ManGO, the platform allows for easy metadata manipulation and management.
For whole genome sequecing data, the standardised metadata from the sequencing instrument is kept together with sequencing data (in html and json format).
The data repositories where the data will be deposited for sharing (for example EGA) also require standard metadata schema which will be followed

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

Where will the data be stored?

DIGITAL DATA

- Code and scripts will be stored on VSC (Flemish supercomputer) and GitHub
- Human sequencing data will be stored on the VSC and the ManGO platform
- Any other relevant data (papers, abstracts, figures) will be stored on KUL OneDrive for business

PHYSICAL SAMPLES

- DNA and tissue samples will be stored at -80deg at at UZ Leuven

How will the data be backed up?

Data stored on ManGO: Snapshots are made at regular intervals (hourly, daily and monthly) in case data needs to be recovered. The data itself is synchronized on two separate hardware storage systems, each 6 PB large, located at Leuven and at Heverlee (ICTS). The data is protected against calamities at either site by synchronizing it in real-time at hardware level.
Raw sequencing data coming off the sequencers is automatically archived on the Leuven Genomics Core Cloud.
Data stored on KU Leuven OneDrive for Business is backed up.

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

KU Leuven servers, VSC and Genomics Core Cloud offer sufficient storage for active data (ManGO, OneDrive, VSC) generated during this project as well as long-term archiving.

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

Access to the OneDrive, VSC, ManGO and Genomics Core Cloud is protected by KU/UZ Leuven user ID and password and after multi-factor authentification. The user rights only grant access to their own data, or data that was shared with them. Data in these drives are mirrored in the second ICTS datacenter for business continuity and disaster recovery so that a copy of the data can be recovered within an hour.

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

Our costs for digital data storage are:

- 35€/TB/Year for the ManGO platform
- 20€/TB/Year VSC Tier-2
- VSC Tier-1 data – free (project-based)
- KUL OneDrive for business - free (max 2TB/lab)
- 20€/TB/Year Genomics Core Cloud archive (retrieval & download costs apply)

Data storage and backup costs are covered by the general lab budget

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...).

The relevant data will be preserved at the university servers for a minimum of 10 years according to the KUL RDM policy.

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

Data and metadata generated during this project will be deposited in EGA under restricted access, where the data will receive unique and persistent identifiers.
Code, models and associated metadata will be shared on GitHub and Zenodo with version control
Other research data will be archived on KU Leuven servers as described above.
Raw sequencing data coming off the sequencers is archived on the Leuven Genomics Core Cloud.

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

Our costs for digital data archiving are:

- 35€/TB/Year for the ManGO platform
- 20€/TB/Year VSC Tier-2
- KUL OneDrive for business - free (max 2TB/lab)
- 20€/TB/Year Genomics Core Cloud archive (retrieval & download costs apply)

Archiving costs are covered by the general lab budget

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
- Yes, in a restricted access repository (after approval, institutional access only, …)

Data and metadata generated during this project will be deposited in EGA under restricted access, where the data will receive unique and persistent identifiers
Code, models and associated metadata will be saved on GitHub and Zenodo with proper versioning
To ensure data findability, links and references to this data will be included in the data availability statements of the associated publication(s)

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

Access to restricted access data, such as human sequencing data, is governed by the Data Access Committees of KU Leuven/UZ Leuven or VIB.

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.

- Yes, Privacy aspects
- Yes, Ethical aspects

Human sequencing data are considered sensitive personal data, and are are only made available on restricted access repositories such as the EGA. Access to these datasets is under control of a Data Access Committee.

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

Datasets and metadata generated from sequencing human samples will be deposited under restricted access on EGA, where they will be assigned a unique and persistent identifier.
Code, models and associated metadata will be saved on Github and Zenodo with proper versioning
Protocols will be deposited on protocols.io

When will the data be made available?

All research output (data, metadata, code) will be made openly accessible at the latest at the time of the publication

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

DATA
Data is typically available under a Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, a Creative Commons Attribution (CC-BY), or an ODC Public Domain Dedication and Licence, with a material transfer agreement when applicable.
CODE
Software and code usually are available under a GNU General Public License or an Academic Non-commercial Software License.

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

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

Zenodo and GitHub deposition is free. Submission to EGA is free. Other data management costs are accounted for and will be covered by the laboratory budget.

6. Responsibilities

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

Robert A. Forsyth, Dr. Laurens Lambrechts and Joris Vande Velde in the lab working on the project will manage data documentation and metadata.

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

Robert A. Forsyth, Dr. Laurens Lambrechts and Joris Vande Velde in the lab working on the project will manage data storage and backup.

Who will manage data preservation and sharing?

During the project: as mentioned above, Robert A. Forsyth, Dr. Laurens Lambrechts and Joris Vande Velde. After the project, the principal investigator (Jonas Demeuelemeester) will guarantee data preservation and data sharing according to KUL/VIB RDM policy

Who will update and implement this DMP?

Jonas Demeulemeester