DMP in the context of FWO Postdoctoral Fellow – senior LIQUIDEVO: Unravelling breast cancer heterogeneity, progression and treatment resistance in the context of a rapid post-mortem institutional tissue donation program using phylogenetic reconstruction algorithms and an Open Science approach.

Project Name: DMP LIQUIDEVO
Principal Investigator / Researcher: Christine Desmedt
Institution: KU Leuven

1. GENERAL INFORMATION

Name of the project lead (PI)
DMP LIQUIDEVO
François Richard

Internal Funds Project number & title
1297322N LIQUIDEVO: Unravelling breast cancer heterogeneity, progression and treatment resistance in the context of a rapid post-mortem institutional tissue donation program using phylogenetic reconstruction algorithms and an Open Science approach.

2. DATA DESCRIPTION

2.1. Will you generate/collect new data and/or make use of existing data?

We will generate and collect new data (prospective data and samples collection and make use of existing data (retrospective data and samples collection).

The project will collect pseudonymized clinical data collected after the patient consented to the study. Clinical, histopathological and treatment data will be entered in the eCRF that we have designed in REDcap. Also sample metadata (SPREC based) will be collected in LabCollector, a database that we have customized for the needs of LIQUIDEVO.

Our research will generate:
- Genomic data: sequencing data (.fastqc and bam files);
- Scripts (R, bash) to analyse the data;
- A web platform allowing browsing in the results;
- Code Ocean capsules and framagit pages to reproduce and access the code, respectively;
- Manuscripts and publications.

In terms of data storage, data will be stored on UZ Leuven drives (eCRF) and KU Leuven drives (all other data) in a GDPR compliant manner and with regular backups provided by the IT services for at least 10 years after the end of LIQUIDEVO.

With regard to manuscript publication: in line with GDPR and KU Leuven Open Science policy, no personnel data will be made publicly available through public repositories such as GitHub and Code Ocean. A version of the manuscript will be uploaded to Lirias. Personal data (e.g. raw sequencing data) will be deposited on public repositories such as the EGA, under restricted access.

2.2. What data will you collect, generate or reuse? Describe the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a numbered list or table and per objective of the project.

Work package	Data type	N° of cases	Data Source	Data content	Data Format	Volume (Go)
WP1	Primary and secondary use, clinical and sample data collection	15	Patient file	Demographic, pathological, clinical, treatment, outcome data and blood results; Samples data	eCRF (RedCap), R files, Lab collector, Excel files	0.5
WP2	Primary use, WGS	125	HiSeq 4000	genomics data	fastq and .bam files	3750
WP4	Primary use, TGS	270	HiSeq 4000	genomics data	fastq and .bam files	405
WP6	Primary use; website source code	1	LTBCR - bioinformatics	R code	R and text files	1

3. ETHICAL AND LEGAL ISSUES

3.1. Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.

Yes, we will use personal data. Data, irrespective of the WPs, include demographic (age, gender), pathological, clinical, treatment and outcome data as well as sequencing data on tissue and liquid samples including whole genome as well as targeted genome sequencing.

Please note, GDPR does not apply to deceased patients.

The compliance monitoring form is the following: E-2021-2462.

3.2. Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s).

LIQUIDEVO is part of the UPTIDER project which was approved by the EC UZ/KU Leuven: S64410.

3.3. Does your research possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

There is potential tech transfer/valorization in the putative assays/biomarkers we measure or we aim to discover. We are working with the Leuven Research and Development department (LRD) and the legal advisors from UZ Leuven. They are involved in all the Material Transfer Agreements and Data Transfer Agreements (MTA/DTAs) we have set-up in the context of this project.

3.4. Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions regarding reuse and sharing are in place?

We are working with the Leuven Research and Development department (LRD) and the legal advisors from UZ Leuven. They are involved in all the Material Transfer Agreements and Data Transfer Agreements (MTA/DTAs) we have set-up in the context of this project. There is no specific data restriction.

4. DOCUMENTATION AND METADATA

4.1. What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?

All collected clinical data are stored in an eCRF (REDCap) in which the documentation is automatically provided through the auto generated dictionary and code book where all the requested data are described. All collected data belonging to the sample collection are stored in an online lab management system called Lab Collector based on SPREC requirements where all the requested data are described.

The lab-specific generic data management plan will guide researchers through data collection and processing workflows, ensuring efficient safe storage, keeping track of data use and associated processing related to each experimental step, and enabling data query filtering of the collected data.

Regarding omics data, bioinformatics pipeline are shared on a common git repository and heavily described so that they can be reproduced. When possible docker-like structure container will be used to easily share and perform bioinformatics pipeline. This container allows to enclose code and software so that they can be reused with the exact same versions and parameters.

Third-party software and algorithms that are used are referenced by their version numbers in our method section and are installed as modules on the VSC and/or containers (Docker, Singularity) on the VSC, to ensure reproducibility.

At the publication level, a companion code capsule hosted by code ocean will be build. It will allow reproducing the figures of the paper, browsing code and raw data. In case of restricted access on the data, a synthetic dataset mimicking the characteristic of the real data will be made available instead (see section 7.2).

4.2. Will a metadata standard be used? If so, describe in detail which standard will be used. If not, state in detail which metadata will be created to make the data easy/easier to find and reuse.

Metadata will follow Data Cite's recommendations (Fenner et al., 2019).

5. DATA STORAGE AND BACKUP DURING THE PROJECT

5.1. Where will the data be stored?

The minimum preservation term of 10 years after the end of the project will be applied to all datasets.
Biological samples obtained under research agreement will be kept according to the EC licenses and agreements. In effect, consent is obtained to store the samples for the specific research purposes stipulated in the informed consent. Putative remnant clinical/patient samples are stored in the UZ Leuven biobank. Hard copies (eg. the Informed Consent forms and paper lab notebooks) are kept in locked cabinets in the lab of the PI concerned.

The data will be stored on the KU Leuven servers. All systems but Lab Collector run on a secured and backed up server of KULeuven (managed by ICT of the Biomedical Sciences Group). These systems also

provide a logging system so no data can ever be erased, making that everything will be traceable and stored long-term (well beyond the common 5-year requirement). Lab Collector data base is externally and professionally managed by the company “AgileBio”.

Developed algorithms and software will be stored on the KU Leuven servers, as well on public repositories such as framagit.com and codeocean.com.

Regarding the web platform, it will be hosted on a virtual server that will be created by the IT department of the KU Leuven (500 euros / year, covered by laboratory funds). Aggregated data will be transferred from the L-drive to the virtual server at the time of publication.

5.2. How will the data be backed up?

The hosting KUL server is automatically backed up using KUL services, multiple times per day. Concerning LabCollector, backups are automatically performed twice a day.

5.3. 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.

There is sufficient storage and back-up capacity on all KU Leuven servers:
- the “L-drive” is an easily scalable system, built from General Parallel File System (GPFS) cluster with NetApp series storage systems, and a CTDB samba cluster in the front-end.
- the Staging and Archive on VSC are also sufficiently scalable (petabyte scale).

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

The total estimated cost of data storage and backup during the project is 1100 per year. This estimation is based on the following costs:

- 500 euros for storage on the L-drive of 4200 Go at 0.111 EUR/Go/year.
- 600 for storage on the VSC cluster of 1000 Go at 0.6 EUR/Go/year.

Budget for compute and data storage is budgeted via the laboratory funds.

5.5. Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Our eCRF, our lab management system (Lab Collector), and the network drive dedicated to the team (L-drive) are password access-protected by users, with person-based decision on rights to access and modify data. Moreover, this is the same for the VSC secured storage which is only accessible to VSC accounts, and specifically our volume will only be accessible to group members.

6. DATA PRESERVATION AFTER THE END OF THE PROJECT

6.1. Which data will be retained for the expected 10 year period after the end of the project? If only a selection of the data can/will be preserved, clearly state why this is the case (legal or contractual restrictions, physical preservation issues, ...).

The minimum preservation term of 10 years after the end of the project will be applied to all datasets described above.

6.2. Where will these data be archived (= stored for the long term)?

As a general rule, datasets will be made openly accessible, whenever possible via existing platforms that support FAIR data sharing (www.fairsharing.org), at the latest at the time of publication or preprint deposition.

For all other datasets, long term storage will be ensured as follows:

- Large sequencing/omics data: will be stored on “L-drive”.
- Small digital files: files will be stored on the “L-drive”.
- Developed algorithms and software will be stored on L-drive, as well on public repositories such as framagit.com and codeocean.com.
- Clinical and sample data will be stored in our lab management tools (eCRF and Lab Collector).

6.3. What are the expected costs for data preservation during these 10 years? How will the costs be covered?

The total estimated cost of data storage during the 10 years after the end of the project is 5000 euros. This estimation is based on the total given in Table 1 and the cost of storage on the L-Drive (0,111/Go/year). The cost will be covered by the laboratory budget.

7. DATA SHARING AND RE-USE

7.1. Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions or because of IP potential)?

There is no specific restriction.
Personal data will only be published after de-identification and identifiers will not be published.
Please note, GDPR does not apply to deceased patients.

7.2. Which data will be made available after the end of the project?

We are committed to publish research results (concerning all datasets) to communicate them to peers and to a wide audience. All research outputs supporting publications will be made openly accessible. Depending on their nature, some data may be made available prior to publication, either on an individual basis to interested researchers and/or potential new collaborators, or publicly via repositories (e.g. negative data). MTA and or DTA have been set-up in this sense and will be set-up if needed.

As part of the open access plan, data will be put available for external users through open source pathways. In that case, these data will be made available after appropriate IP protection.

Upon publication, all anonymized patient details supporting a manuscript will be made publicly available as supplemental information. Personal data will only be published/shared after de-identification and identifiers will not be published/shared.

Omics datasets will be deposited in open access repositories such the NCBI Gene Expression Omnibus (GEO) or The European Genome-phenome Archive (EGA). All the relevant algorithms, scripts and software code driving the project will be stored in a private online git repository of the laboratory. As soon as the manuscript is publicly available, the repository will be changed to a public repository. The platform Code Capsule will be used upon publication to increase reproducibility of the research and open both data and code supporting the publication. The code required to generate the figures of the paper will be available online. Publicly available data will be directly included in the capsule allowing to easily reproduce the paper online. In case part of the data falls under restricted access, a synthetic version of it will be made available, allowing the user to run the code. Proper data will be made available with a granted MTA or DTA.

7.3. Where/how will the data be made available for reuse?

In an Open Access repository.
Upon publication, all anonymized patient details supporting a manuscript will be made publicly available as supplemental information.

Omics datasets will be deposited in open access repositories such the NCBI Gene Expression Omnibus (GEO) or The European Genome-phenome Archive (EGA). All the relevant algorithms, scripts and software code driving the project will be stored in a private online git repository of the laboratory. As soon as the manuscript is publicly available, the repository will be changed to a public repository.

A web platform hosted on a virtual server at the KU Leuven will allow browsing the results of the publication, showing only aggregated data.

7.4. When will the data be made available?

Upon publication. However, depending on their nature, some data may be made available prior to publication, either on an individual basis to interested researchers and/or potential new collaborators, or

publicly via repositories (e.g. negative data). MTA and or DTA have been set-up in this sense and will be set-up if needed.

7.5. Who will be able to access the data and under what conditions?

Whenever possible, datasets and the appropriate metadata will be made publicly available through repositories that support FAIR data sharing.
Moreover, as mentioned above MTA and or DTA have been set-up and will be set-up if needed.

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

It is the intention to minimize data management costs by implementing standard procedures e.g. for metadata collection and file storage and organization from the start of the project, and by using free-to-use data repositories and dissemination facilities whenever possible. Data management costs will be covered by the laboratory budget.

8. RESPONSIBILITIES

8.1. Who will be responsible for the data documentation & metadata?

Metadata will be documented by the senior post-doc, research manager, PhD students and technical staff at the time of data collection and analysis.

8.2. Who will be responsible for data storage & back up during the project?

The senior post-doc, research manager and technical staff will ensure data storage and back up, with support from ICTS, gbiomed-IT staff, and UZ-IT staff.

8.3. Who will be responsible for ensuring data preservation and sharing?

The PI is responsible for data preservation and sharing, with support from the team, ICTS, gbiomed-IT staff, and UZ-IT staff.

8.4. Who bears the end responsibility for updating & implementing this DMP?

The PI is ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.