DMP title

Project Name My plan (FWO DMP) - DMP title

Grant Title 11K6222N

Principal Investigator / Researcher Bernard Thienpont

Project Data Contact Prof. Bernard Thienpont

Description Cellular heterogeneity is integral to biology, impacting development and physiology as well as diseases such as cancer. The genetic control over deterministic phenotype changes has been studied in detail: we know how transcription factors can induce differentiation and how mutations can drive oncogenic transformation. However, for accommodating such changes, cells need to control the plasticity of their phenotype through processes that are far less understood. Advancements in single cell-omics enable us to measure cell-intrinsic stochastic heterogeneity. I hypothesize that this aspect of transcriptional heterogeneity is epigenetically regulated, and controls acquisition of therapy resistance. To test this hypothesis, I am combining CRISPR and epigenetic inhibitor screens with single cell RNA and ATAC-seq. I will measure how altering epigenetic enzyme activity affects stochastic and deterministic cellular heterogeneity. Next, the role of epigenetic drivers of heterogeneity in therapy response will be validated by assessing how they affect drug tolerance of cancer cells, and whether reducing heterogeneity curbs resistance acquisition. Lastly, I will quantify the influence of defined â€œsignatures of heterogeneityâ€ on therapy outcome in patients. I will be focusing specifically lung adenocarcinoma cells, because they notoriously have a heterogenous phenotypic composition. My ultimate aim is to improve patient outcome by understanding the mechanisms that drive tumor heterogeneity. This Data Management Plan describes how data will be generated, organized, stored, and backed up throughout the duration of the project and after its completion.

Institution KU Leuven

1. General Information

Name applicant

Paulien Van Minsel

FWO Project Number & Title

11K6222N Epigenetic regulation of transcriptional heterogeneity controls cancer therapy resistance acquisition.

Affiliation

- KU Leuven

KU Leuven Department for Human Genetics

2. Data description

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

- Generate new data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Type of data	Format	Volume	How created

Raw NGS sequencing data: Single-cell transcriptomics and DNA data	.fastq	5 TB	Obtained through processing of cells from CROPseq screen, CRISPR screens, small molecule inhibitor screen, patient biopsies, tumours in PDTX mice. Data generated using Illumina NovaSeq and NextSeq 2000 machines.
Data analysis, including genome-alignment data, statistical analysis, count matrix and gene/cell index.	.fa, .R, .RData, .png, .pdf, .txt, .xlsx, .rds, .tsv, .html, .ipynb, .bam	15 TB	Analysis of the generated NGS sequencing data. From processing and quality control to heterogeneity analysis. Scripts produced in python, R and Bash. Jupyter notebook.
Measurements in mice with patient-derived tumor xenografts (PDTX), such as tumour growth delay curves. Microscope images, stainings.	.xlsx, .tiff, .jpg	5 GB	Experimental measurements performed in mice.
Analysis of data measurements obtained in PDTX mice.	.R, RData, .rds, .pdf, .png, .xlsx	5 GB	Analysis of data generated in mice expreriments. Such as statistical analysis perfomed in R.
Western blot images and analysed data	.tiff .xls	100 MB	Tif images of chemiluminiscent signal using ImageQuant istrument. Analysis using ImageJ.
PCR gel images	.tiff .jpg	100 MB	DNA or PCR products run on agarose gels containing SYBRSafe reagent. Gels visualised using UVP Benchtop UV Transilluminator.
DNA plasmids	tubes of liquid containing DNA plasmids	/	Plasmids produced during the course of the project from existing DNA plasmids provided by commercial suppliers

Lab books	Paper notes	5 paper lab books	Detailed written notes with date for experiments carried out concerning this project. All practices documented following good laboratory practices.
Experimental/computational protocols related to experiments performed during this project and their analysis	.docx, .xlsx, .txt	<1 GB	Detailed standard operating procedures as well as details of data collection, processing and analysis.
qRT-PCR data	.xls	100 MB	qRT-PCR raw data, collected using Quant Studio, as well as analysis done in Excel.
Sequencing libraries	tubes of liquid containing libraries for Illumina Sequencing.	/	Libraries obtained through processing of the cancer cells used in this project, final libraries generated according to standard operating procedures from Illumina and 10x Genomics.
Processed data; labmeeting presentations, experiment overviews, reports, posters, graphical figures and manuscript	.doc, .ppt, .xlsx	<10 GB	Data processed for presentation or publication.
Crypreserved stocks of cell lines to be kept in liquid nitrogen cell storage tanks.	Cryovials containing frozen liquid cell lines.	/	Stocks of cell lines generated during the course of the experiment. Produced from original commercially available cell lines through transfection or lentiviral transduction.
Microscopy images of cell cultures	.jpeg, .tiff	100 MB	Images from Cell Culture at relevant time points such as seeding/treatment. Obtained through Incucyte, Light microscope and Luna FL Cell counter

Numerical data from cell culture	.xlsx	50 MB	Cell counts from Cell Culture at relevant timepoints such as seeding, treatment, collection, before 10x processing... . Obtained through Luna FL Cell counter and Incucyte.

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- Yes

The final workpackage in the last stages of the project involves the use of personal data of patients with locally-advanced or metastatic non-small-cell lung cancer. The UZ Leuven KWS system will serve as the source for the clinical information, and electronic case report forms will be used for collection of these coded data. All obtained research data and clinical information will be added in a coded manner to a REDCap database, located on a KU Leuven hosted and secured server. A trained and clinically involved person will add all research information to the database, especially designed for this research. Only coded information will be extracted and used for the analyses by researchers directly involved in the project. All records identifying the patient will be kept confidential. Access to clinical charts will be done under supervision of Dr. Wauters.

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)

- Yes

Animal Experiments will be performed as part of this project for the last work package in the final stages of this project, all animal experiments will first be sent for approval by the Ethical committee for Animal Experimentation (ECD) at KU Leuven.

Does your work 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?

- Yes

It is possible intellectual property might arise from this work. In this case this will be managed based on the current guidelines by the KU Leuven.

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 are in place?

- No

4. Documentation and metadata

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

Physical paper lab books will be kept by the applicant (Paulien Van Minsel) with clear and dated protocols noting experiment design, conditions and detailed step by step oversight of data acquisition, according to good laboratory practices. In the notes instruction will be provided where to find extra provided digital data, such as excel files and pictures with file names and dates. Detailed protocols will also be provided (.docx). Lab books will at all times be kept in lockable file cabinets, accessible upon reasonable request. List of available physical data, such as cell lines, plasmids, DNA oligo's (primers), ... will be kept on Dropbox and regularly updated. Information concerning processing and analysis of the data will be organized in dated folders, labeled properly according to type of experiment, conditions analyzed and source. Files will be stored on locally kept hard disks, KU Leuven High Performance Computing (HPC- part of the Flemish Supercomputer Center) and finally on OneDrive (managed by the KU Leuven IT department) to enable the possibility of sharing.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

Folders will be organized in a logical hierachy with the broader research topics first, then sub-projects and subsequently within these we will store the data results and analysis in structured subfolders. Each file name will consist of the date and ordered experiment name to allow for easy retrieval of all necessary information to ensure reusability of the data. This will be maintained for both the personal KU Leuven managed PCs, the HPC as well as the external SSDs for long term storage.

5. Data storage and backup during the FWO project

Where will the data be stored?

Paper lab notebooks are kept in lockable cabinets in the lab. Digital data files are stored on local KU Leuven PCs and/or KU Leuven High Performance Computing (HPC), which is part of the Flemish Supercomputer Centre (VSC). Next Generation Sequencing data are stored on HPC archive storage, which has snapshots that can be recovered. Processed data (alignment files and other intermediated files) are stored on HPC staging storage, which is suitable for large storage capacity and actively computed. Analysis results including scripts, figures and manuscripts are transferred to local KU Leuven managed personal PCs and transferred to SSDs for long-term preservation and back-up. Non-digital or non-written data (cell lines, plasmids, libraries...) will be stored in appropriate conditions -80°C, -20°C freezers or liquid nitrogen to allow for long-term storage.

How is backup of the data provided?

Besides regularly provided automated backups by ICTS (of HPC archive storage), the data stored on personal PCs will be weekly backed up on the external SSD hard-disks (up to 1-2 TB storage capacity). Whenever a project has been completed or published, all raw data, necessary intermediate results, scripts, figures, and menuscripts will be moved from HPC staging to archive storage, which will be automated with backups by ICTS.

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

Yes. Our current HPC storage has 21 TB on staging and 12TB archive. Extensions of this volumne can be asked for at ICTS at all times (for an additional cost). Storgae price will be charged annually on actual usage.Extra external hard disks will be bought, as appropriate. Same for liquid

nitrogen containers or freezers, but sufficient capacity is currently available.

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

HPC staging storage cost: 20 euro/year/1TB, archive storage cost 70 euro/year/1TB (with free snapshots). Dropbox: free. External hard disks: max. 500€ (=3 disks of 1-2 TB). These costs will be covered by the budget of the project lead (Prof. Bernard Thienpont).

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

All files and data on HPC are under strict permission management. The first author of the project is the moderator of the project group. Only members registered on VSC and added by the moderator into project group can access the storage space. Additionally, all the permissions of folders, files and data can be changed by the creator and the moderator of the group, which makes only the project group members can access project related data. After recieveing NGS data, raw data will be immediately updated to HPC archive storage and protected by read only permission. All VSC accounts are strictly protected by Public-key cryptography.

All KU Leuven PCs and SSDs are protected by passwords. Paper lab notebooks are kept in lockable cabinets.

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

All generated data will be preserved, raw and processed, for at least 5 years.

Where will the data be archived (= stored for the longer term)?

Long term storage will be provided by the HPC archive storage and external SSD hard disks. Plasmids, cell lines, ... will be stored at appropriate storage conditions at -20° and -80°C freezers or liquid nitrogen storage costs.

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

HPC staging storage cost: 20 euro/year/1TB, archive storage cost 70 euro/year/1TB (with free snapshots). External hard disks: max. 500€ (=3 disks of 1-2 TB). Lab books in lockable cabinets: free. These costs will be covered by the budget of the project lead (prof. Bernard Thienpont).

7. Data sharing and reuse

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)?

- No

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

Main research findings will be made available through publication of journal articles in established, peer-reviewed (non-predatory) academic journals. Raw and processed sequencing data (.xfastq: .txt: .bam) will be transferred to a public repository (EBI ArrayExpress).

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

- In an Open Access repository
- Upon request by mail

Upon manuscript submission, raw and processed sequencing data (.fastq; .txt; .bam) will be transferred to a public repository (EBI ArrayExpress). From the project onset, data documentation will be tailored for deposition in public repositories, with spreadsheet headers corresponding to fields required by these repositories. All methods used will be documented in sufficient detail to allow for
independent replication.

When will the data be made available?

- Upon publication of the research results

Relevant raw data will be transferred to a public repository upon publication of the research results in established, peer-reviewed academic journals.

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

Relevant raw data to publications in journal articles will be available in a public repository upon publication.

Sharing of data, other than those publically available, will be assessed on a case-by-case basis by the project lead (Prof. Bernard Thienpont) upon reasonable request.

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

Costs for data sharing will be discussed with collaborators on a case-by-case basis. Publication costs (Open Access) will be covered by consumables budget.

8. Responsibilities

Who will be responsible for data documentation & metadata?

The applicant (Paulien Van Minsel), in-lab bio-informatic collaborators (Qian Yu, PhD student) and the project lead (Prof. Bernard Thienpont) will share resonsibility for data documentation and metadata generation/preservation.

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

The applicant (Paulien Van Minsel) and will be primarily responsible for collecting/generating data, and for correct documentation. In-lab collaborator (Qian Yu, PhD student) has shared responsibility for the mainting and correctly documenting of the NGS raw data and its analysis for the duration of the project. The KU Leuven ICTS will be responsible for maintenance of the HPC staging and archive storage.

Who will be responsible for ensuring data preservation and reuse ?

The applicant (Paulien Van Minsel) and the project lead (Prof. Bernard Thienpont) will share the responsibility for ensuring data preservation and reuse.

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

The applicant (Paulien Van Minsel) will bear the responsibility of updating and implimenting this DMP for the duration of the project. The project lead will bear the responsibility of maintenance and storage of the data when the projects is finished.