Raising the voice of silent mutations in cancer

A Data Management Plan created using DMPonline.be

Creator: Grecia Marron Linares

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Grant number / URL: PDMt1/23/021

ID: 202353

Start date: 01-10-2023

End date: 30-09-2024

Project abstract:

In the past 15 years, next-generation sequencing has drastically increased our capacity to identify mutations in cancer samples. Whereas synonymous mutations represent 23% of the somatic variants detected in the protein coding region of cancer genomes, these mutations are typically ignored and most studies only focus on mutations that cause amino acid changes, assuming that synonymous (or silent) mutations are meaningless random events. Nevertheless, several experimentally tested synonymous mutations have been ascribed a causative role in various diseases. Thus, I hypothesize that the role of synonymous mutations in cancer is currently underestimated and that some of these mutations can significantly contribute to cancer formation and cancer drug sensitivity. I will use a compendium of bioinformatic tools to delineate the landscape of silent driver mutations in cancer. Mutational effects on functional DNA or RNA elements will be used to rank mutations according to their likeliness to be cancer drivers. I will select the 3 top candidate driver synonymous mutations and will test them experimentally in CRISPR-Cas9 engineered cell line models for effects on RNA and protein expression of the mutant gene and on cancer cell behavior and drug sensitivity. This project will thus contribute to a better understanding and awareness of pathogenic synonymous mutations in cancer.

Last modified: 15-11-2023

Raising the voice of silent mutations in cancer

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.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		Indicate: ***N**(ew data) or **E**(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
MSC3	experimental	E	D	N, S	BAM, SRA,TSV	>5TB	
PublicCancerData	experimental	E	D	N, T	FASTA, FASQ, .cvs, .pdf	<5TB	
LabData	experimental	N	D	N	.sav, .csv, .pzf, .xls, .nd2	<100GB	
ImageData	experimental notes	N	D	A,I	.gif, .jpg, .png, .tiff	<1TB	
TextData	experimental	N	D, P	T	.doc, .csv	<100GB	
Manuscripts	publications	N	D, P	T	.doc, .pdf	<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 public data originate from the MC3 project (Ellrott et al., Cell Systems, 2018), and from published studies on hematological cancers (Liu et al., Nature Genetics, 2017 and Reddy et al., Cell, 2017).

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, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- Yes, human subject data (Provide SMEC or EC approval number below)

Ethics number approval S66987 (Approval experiments in Mel-ST cell line)

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- No

Genomics datasets from human cancer patients will be analyzed in this project. We do not have access to the identity of these patients as pseudonymized patient IDs have been given before the data became available to us. These pseudonymized patient IDs will be used at all times during this project and when sharing the data in publications

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 or 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.

- No

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keepdata 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).

Exome and transcriptome data obtained from both, the MSC3 and relevant cancer publications, consist of BAM files, SRA files transformed into fastq files and MAF files. Methylation and protein array data consist of TSV files.

Spreadsheets of numerical data will be generated by a variety of instruments and associated software packages. These spreadhseets will contain the results of cell and molecular biology experiments, including western blot protein analysis, RT-qPCR, flow cytometry data, Incucyte data, next generation sequencing data (.sav, .csv, .pzf, .xls, .prism). Furthermore, picture files and graphical representations of data will be generated (.gif, .jpg, .png, .tiff, .nd2).

Notes will be taken in both hard copy notebooks and electronic documents (.doc, .csv)
Manuscripts describing obtained research results will be generated (.doc, .pdf)

Will a metadata standard be used to make it easier tofind and reuse the data?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- No

No real metadata standard will be used. Each researcher working on this project will document data, procedures and experiments. While each researcher will annotate his/her data according to their own standards, the data that will be made publically available will be annotated according to the standards of the utilized public databases (GEO, EGA, etc.) or of the peer reviewed journals where the data are published. This will make it easily findable and reusable.
A spreadsheet database of all cel lines and samples generated for this project and other projects in the lab will be maintained on the KU Leuven central storage drive.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Shared network drive (J-drive)
- OneDrive (KU Leuven)
- Large Volume Storage

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

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

Each researcher working on this project will document data, procedures and experiments in folders on the KU Leuven central. This document was generated by DMPonline (http://dmponline.dcc.ac.uk) 4 of 7 storage drives, organized per subproject any type of experiment. A spreadhseet database of all cel lines and samples generated in this project will be maintained on the KU Leuven central storage drive to which anyone that does not belong to the lab can have access.

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

Cost of €51,90 / 100GB / year for storage on the KU Leuven J-drive. Moreover, the cost to storage 5Tb / year on the KU Leuven L-drive is €569,2. We anticipate to need a budget of maximum 700 euro for data storage and back-up during the PDMt-1 mandate. The costs will be covered from money that the lab has saved over the years. Liquid nitrogen storage capacity costs €50 / year for one entire sample column (13 boxes of 81 samples). We expect to need half a column for sample storage related to this project. These costs will be covered by the grant. We do not expect any costs related to -20°C and -80°C storage given the investments that our lab made in the past five years.

Data Preservation after the end of the Research Project

Which data will be retained for 10 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...).

- All data will be preserved for 10 years according to KU Leuven RDM policy

The data will be stored on the central servers of my host institute KU Leuven (with automatic back-up procedures) for at least 10 years, conform the KU Leuven research data management policy.

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

- KU Leuven RDR

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

Cost of €51,90 / 100GB / year for storage on the KU Leuven J-drive. Moreover, the cost to storage 5Tb / year on the KU Leuven L-drive is €569,2. We anticipate to need a budget of maximum 6000 euro for data storage and back-up during 10 years. The costs will be covered from money that the lab has saved over the years.
Liquid nitrogen storage capacity costs €50 / year for one entire sample column (13 boxes of 81 samples). We expect to need half a column for sample storage related to this project. These costs will be covered by the grant. We do not expect any costs related to -20°C and -80°C storage given the investments that our lab made in the past five years.

Data Sharing and Reuse

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

- Yes, as open data
- Other (specify below)

All methodologies that are developed, results that are obtained and reagents that are generated will be made available to the scientific community. Several of the public datasets that we start from (MC3 data and other) are under controlled access and cannot be shared by us because of contractual restrictions in the data access agreement we signed.
Results from analyses obtained with these datasets can and will be shared.

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

Both our J-drive and L-drive are organized in a way that only former lab members can have access to the data storage there.

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 per dataset or data type where appropriate.

- Yes, intellectual property rights

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- KU Leuven RDR (Research Data Repository)
- Other data repository (specify below)

In data repositories
In peer reviewed journals

Results from all aims will be published in peer-reviewed scientific journals. Open access will be ensured via Europe PMC and our institutional repository Lirias.
Cell lines and DNA contructs generated will be available within the framework of a scientific collaboration.
Next generation sequencing datasets (e.g. RNA-seq) generated will be deposited in an appropriate data repository such as GEO or EGA.

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.

- CC-BY 4.0 (data)
- Data Transfer Agreement (restricted data)

Depending of the type of data, we will choose the most appropriate license

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- Yes, a PID will be added upon deposit in a data repository

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

We have calculated open access costs for scientific publications in the project budget.
After the ending of this project, money that has been saved by the host lab over the years will be used to cover the costs to publish work originating from this project that could not be published within the duration of the project.

Responsibilities

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

The PI and postdoc Grecia Marron will be responsible for managing the data documentation and the metadata during the project

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

The PI and postdoc Grecia Marron will be responsible for managing data storage and backup during the research project

Who will manage data preservation and sharing?

The PI and postdoc Grecia Marron will be responsible for managing data preservation and sharing

Who will update and implement this DMP?

The PI and postdoc Grecia Marron will be responsible for implementing the DMP.
They will update the DMP anytime conditions change. A mid-term review will be
accompanied by a detailed DMP and a final reviewed DMP will be sent along with the
final report.