Plan Overview

A Data Management Plan created using DMPonline.be

Title: Alex' FWO DMP

Creator: Alexandra Pančíková

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Project abstract:

Parkinson's disease (PD) is the second most common neurodegenerative disease worldwide. Genome-wide association studies have identified 90 SNPs associated with PD. However, at most of these, the relevant genes remain to be discovered and lead SNPs are unlikely to be causal, instead tagging cryptic structural variation. Major developments in omics technologies enable us to simultaneously detect gene expression and chromatin accessibility in thousands of single cells. Likewise, long-read sequencing now offers increased accuracy, yield and read lengths, allowing us to identify any type of variant, anywhere in the genome and query epigenetic states with single molecule base pair resolution.

Here, I propose to leverage long-read whole-genome and single-cell multiomic data we have recently generated for brains of 190 donors to shed light on the role of structural variation in PD. To this end, I will develop computational frameworks to construct personalized diploid genomes of all donors and use them to assess allele-specific signals in the patient-matched single-cell data. In addition, DNA methylation patterns on the long single molecule reads will further prioritise and confirm risk variants. Whereas traditional approaches can profile allele-specific effects at SNPs only, our setup is uniquely powered to reveal all types of functional variation. In conclusion, this project will uncover gene regulatory variation in the brain and the causal variants contributing to PD pathology.

ID: 211907

Start date: 01-11-2024

End date: 31-10-2028

Last modified: 09-01-2025

Questionnaire

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

I will use in-house generated whole-genome long-read sequencing data and single-cell multi-omics data. I will generate code, algorithms and scripts.

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)

1. Responsible people: Alexandra Pancikova, Jonas Demeulemeester (Principal Investigator), Stein Aerts (Principal Investgator)
2. Storage capacity/repository

DURING the research

- the data and all the code will be stored on KU Leuven drives (OneDrive for business, J-drive, ManGo) and Flemish Supercomputer Center (VSC) servers with mirroring, with restricted access to assure the data safety

AFTER the research

- All the code will be well-documented and released on GitHub and Zenodo with proper versioning
- Raw sequencing data will be archived under restricted access in European Genome Phenome Archive (EGA) and/or Aligning Science Across Parkinson's (ASAP) platform
- Derived single-cell data and deidentified aggregated variant calls will be shared in GEO/ArrayExpress as read counts and vcf files, respectively.
- All other files will be archived on KU Leuven servers for 5 years

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)

Not applicable.

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)

The data used in this study is generated from post mortem human brain tissue, and we thus apply specific security measures as required. The data is only accessible to the researchers working on the project and multi-factor authentication is required to access the data.
Samples were obtained as part of the ASAP consortium.
Ethical committee approval was obtained - S64966.

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

Not applicable.

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
WGS	Long-read whole genome sequencing data (Oxford Nanopore Technologies) of post mortem human brain samples from donor cohort	• Generate new data	Digital	Experimental	Raw • NA - fast5 or pod5 Processed • NA- CRAM or BAM files • NA- VCF files • metadata: .txt, .csv	>50TB	
snRNA-seq data	single-nucleus RNA sequencing (10X Genomics snRNA-seq and multiome; ParseBio snRNA-seq) of post mortem human brain samples from donor cohort	• Generate new data	Digital	Experimental	Raw • NA- binary call format (.bcl) Processed • NA: FASTQ files (.fastq and zipped as .gz) • metadata .txt, .csv	<50TB	

snATAC-seq data	single-nucleus ATAC sequencing (10X Genomics multiome and snATAC-seq) of post mortem human brain samples from donor cohort	• Generate new data	Digital	Experimental	Raw • NA- binary call format (.bcl) Processed • NA: FASTQ files (.fastq) • metadata .txt, .csv	>50TB	
NanoWGS	Computational pipeline to process WGS data	• Generate new data (based on existing software)	Digital	Derived/Aggregated	• NA- Nextflow files (.nf, .config) • metadata - .txt, .csv	<1Gb	
SnRNA-seq processing and analysis pipeline	Collection of	• Generate new data (based on existing software)	Digital	Derived/Aggregated	• NA - analysis scripts and notebooks (.py, .ipynb,.R, .sh) • textual data (.txt, .csv)	<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

Post mortem human brain samples were obtained as part of the ASAP consortium.
Ethical approval was obtained at the beginning of the project - S64966.

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

We obtained some personal data for human samples from the collaborating biobanks (Banner Sun Health, USA; Edinburgh, UK; QSBB London, UK; PUK London, UK) including disease status, age, sex and ethnicity. All data was anonymised. We generate sequencing data that contains genetic information.

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

Our agreement with Aligning Science Across Parkinson's (ASAP) obliges open science approaches when disseminating any data. In addition, we have material transfer agreements with multiple biobanks (QSBB, Banner, Edinbrugh, PUK) that may restrict the dissemination of data within certain contexts.

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

Not applicable

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). The detailed protocols will be published on protocols.io and described in the publication. 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, I use electronic lab notebooks to preseve any pertinent information about analysis.

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.

All datasets will be accompanied by metadata that is stored in electronic lab notebooks and in samplesheet.

Sequencing data have a standardised naming to allow for easy tracking and reuse of the data.

Raw single nuclei sequencing data is named as:

- SEQUENCER_NAME_YYYYMMDD/PROJECT__CUAL__NAME_* where:
- SEQUENCER_NAME: e.g. NovaSeq6000, NextSeq2000
- YYYYMMDD: Sequencing date
- PROJECT: 3 character project code (in this project, the code is ASA)
- CUAL: 6 character Globally Unique, Correctable, and Human-Friendly Sample Identifier for Comparative Omics Studies (generated with: https://github.com/johnchase/cual-id: cual-id)
- NAME: descriptive sample name

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 ASA)

- NAME: descriptive sample name

In the lab, 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 shared drives or KUL OneDrive.

Metadata will be saved in excel sheets or .csv files on KUL OneDrive or shared drives. 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 ManGO platform
- any other relevant data will be stored on KU Leuven provided J-drive or KUL OneDrive for business

PHYSICAL SAMPLES

- post mortem brain tissue samples will be stored in freezer locally in the laboratory

How will the data be backed up?

- Data stored on ManGO platform: 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.
- Data stored on the KUL L-drive is backed up daily using snapshot technology, where all incremental changes in respect of the previous version are kept online; the last 14 backups are kept.
- Data stored on the KUL J-drive is backed up hourly, daily (every day at midnight) and weekly (at midnight between Saturday and Sunday); in each case the last 6 backups are kept.

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 and VSC offer sufficient storage for active data (J/L-Drive, ManGO, OneDrive) generated during this project.

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

Access to the KUL drives (J-Drive), OneDrive, VSC and ManGO servers are accessible only with a KU 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?

- The costs of digital data storage offered by KUL is:
  - 569,2€/5TB/Year for the L-drive
  - 519€/TB/Year for the J-drive
  - 35€/TB/Year for the ManGO platform
  - KUL OneDrive for business - free
- Data storage and backup costs are covered from 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 on 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, GEO/ArrayExpress, and/or ASAP platform 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 proper versioning
- Other research data will be archived on KU Leuven servers as described above.

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

- The costs of digital data storage offered by KUL is:
  - 569,2€/5TB/Year for the L-drive
  - 519€/TB/Year for the J-drive
  - 35€/TB/Year for the ManGO platform
  - KUL OneDrive for business - free
- Data storage and backup costs are covered from 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 and/or ASAP platform under restricted access, where the data will receive unique and persistent identifiers
- GEO/ArrayExpress will be used for data that does not require restricted access (ie aggregated population variant calls)
- 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.

The access to the restricted access dataset, such as post mortem human sequencing datasets, 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

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 human sequencing will be deposited under restricted access on the ASAP platform and on EGA, where they will be assigned a unique and persistent identifier. Data that does not require restricted access will be shared on GEO/ArrayExpress
- 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 CommonsAttribution (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

Not applicable

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

Zenodo and GitHub deposition is free. Submission to EGA, ASAP platform and GEO 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?

Alexandra Pančíková and colleagues from the lab working on the same data - Olga Sigalova, Gert Hulselmans, Koen Theunis, Julie De Man

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

Alexandra Pančíková and colleagues from the lab working on the same data - Olga Sigalova, Gert Hulselmans, Koen Theunis, Julie De Man

Who will manage data preservation and sharing?

During the project - Alexandra Pančíková and the aforementioned colleagues from the lab. Sharing will be supported by KUL-based ASAP project manager - Sara Salama. After the project, the principal investigators (Jonas Demeuelemeester and Stein Aerts) will guarantee data preservation and data sharing according to KUL RDM policy

Who will update and implement this DMP?

Alexandra Pančíková