DATA MANAGEMENT PLAN (DMP)

Endoplasmic reticulum and mitochondria: common drivers of human β cell and neuronal fate

Form Number: D-2022-1717

1. General Information	
Name applicant :	*Pierre Vanderhaeghen*
FWO Project Number & Title:	PANDAROME Endoplasmic reticulum and mitochondria: common drivers of human β cell and neuronal fate
Affiliation:	KU Leuven VIB – KU Leuven Center for Brain and Disease Research *Stem Cell and Developmental Neurobiology Lab*

2. Data description	
Will you generate/collect new data and/or make use of existing data?	Generate new data

VIB-KU LEUVEN CENTER FOR BRAIN & DISEASE RESEARCH

Describe the origin, type and format of the data (per dataset) and its (estimated) volume	The research and technical staff will generate, collect, process, analyze and store the data listed below, as detailed in the project description. The following datasets will be generated: **1. Experimental data** **Digital images** Microscopy pictures, gel scans, graphs, illustrations, figures **Electrophysiology data** Patch clamp recordings from neuronal cells **Omics data** Sequencing **Vectors** Bacterial vectors, viral vectors **Cell lines** Bacteria strains, pluripotent cell lines, primary cell lines **Genetically modified organisms** Living mice and related documentation (ethical approval documents, husbandry and welfare data) **2. Derived and compiled data** **Research documentation** Research documentation generated by the research and technical staff or collected from online sources and from collaborators, including ethical approval documents, laboratory notes, protocols, animal husbandry data. **Manuscripts** **Algorithms and scripts** **3. Canonical data** **Nucleic acid sequences following RNAseq experiments** These datasets represent an important source of information for the laboratory of the PI (including future staff), for scientists, journalists and higher education teachers working in the field of neurocience and developmental biology, but also for non-profit organizations and industries active in these fields. Data will be stored in the following formats: - Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Mark-up Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; - Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb);

	- Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif; - Digital images in vector formats: scalable vector graphics (.svg), encapsulated postscript (.eps), Scalable Vector Graphics (.svg), Adobe Illustrator (.ai); - Flow cytometry data: Flow Cytometry Standard (.fcs); - Electrophysiology data: Applied Biosystems Sequence Tracer Sequence Trace (.abf); - Nucleotide and protein sequences: raw sequence data trace (.ab1), text-based format (.fasta/.fa) and accompanying QUAL file (.qual), Genbank format (.gb/.gbk); - Next generation sequencing raw data: binary base call format (.bcl), .fastq(.gz) - Sequence alignment data: (.sam), .bam - Read/UMI count data: .tsv(.gz), Matrix Market format (.mtx), .loom, .rds(.gz) - Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C.

3. Ethical and legal issues	
Will you use personal data? If so, shortly describe the kind of personal data you will use AND add the reference to your file in your host institution's privacy register.	□ No
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 - Human cell lines: human embryonic stem cell lines obtained from international non for profit organizations (WICELL). Ethical committee approval S61642 by EC Research UZ/KU Leuven. - Genetically modified organisms: animals are housed in facilities of the Laboratory Animal Center of KU Leuven, which applies Standard Operation Procedures concerning housing, feeding, health monitoring to assure consistent care in accordance with European and national regulations and guidelines. Animal administrative, husbandry and animal welfare data are sensitive data and are stored in the LAIS database according to security procedure of KU Leuven. Ethical aproval KU Leuven Ethical committee for animal experimentation ECD P030/2018.
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 We do not exclude that the proposed work could result in research data with potential for tech transfer and valorization. VIB has a policy to actively monitor research data for such potential. If there is substantial potential, the invention will be thoroughly assessed, and in a number of cases the invention will be IP protected (mostly patent protection or copyright protection). As such the IP protection does not withhold the research data from being made public. In the case a decision is taken to file a patent application it will be planned so that publications need not be delayed.

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 No third-party agreement restricts dissemination or exploitation of the data from this project. In particular, existing agreements between VIB and KU Leuven do not restrict publication of data.

4. Documentation and metadata	
What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?	Metadata will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in the electronic laboratory notebook (E-notebook) that refer to specific datasets. All datasets will be accompanied by a README.txt file containing all the associated metadata (see more details below).

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.	□ No X Yes While specific data types might require particular metadata, as a general rule the metadata will be based on a generalized metadata schema such as Dublin Core or DataCite, including the following elements: • Title: free text • Creator: Last name, first name, organization • Date and time reference • Subject: Choice of keywords and classifications • Description: Text explaining the content of the data set and other contextual information needed for the correct interpretation of the data, the software(s) (including version number) used to produce and to read the data, the purpose of the experiment, etc. • Format: Details of the file format, • Resource Type: data set, image, audio, etc. • Identifier: DOI (when applicable) • Access rights: closed access, embargoed access, restricted access, open access. For specific datasets, additional metadata will be associated with the data file as appropriate such as experimental procedures to generate transcriptomic data. The final dataset will be accompanied by this information under the form of a README.txt document. This file will be located in the top level directory of the dataset and will also list the contents of the other files and outline the file-naming convention used (see section 7 below). This will allow the data to be understood by other members of the laboratory and add contextual value to the dataset for future reuse.

5. Data storage & backup during the FWO project	
Where will the data be stored?	Digital files will be stored on KU Leuven servers, except for private data that will be stored on KU Leuven secure server (digital vault). - Omics data: omics data generated during the project will either be stored on KU Leuven servers or on The Flemish Supercomputer Centre (VSC), initially in the staging area and later in the archive area. - Vectors: As a general rule at least two independently obtained clones will be preserved for each vector, both under the form of purified DNA (in -20°C freezer) and as a bacteria glycerol stock (-80°C). All published vectors and the associated sequences will be sent to the non-profit plasmid repository Addgene, which will take care of vector storage and shipping upon request. - Cell lines: Newly created human pluripotent cell lines will be deposited in the hPSCreg database. Other cell lines will be stored locally in the laboratory. - Genetically modified organisms: Mice will be maintained in facilities of the Laboratory Animal Center of KU Leuven, which applies Standard Operation Procedures concerning housing, feeding, health monitoring to assure consistent care in accordance with European and national regulations and guidelines. All animals will be registered in the Leuven Animal Information System (LAIS) database, along with corresponding genotyping information, ethical approval documents and animal provider receipts. Drosophila lines will be stored in a dedicated room and managed using a specific database for storage of the corresponding information (including genotype, origin, number of vials and date of transfer, crossing schemes) and vial tracking via unique QR codes. Other biological and chemical samples: storage at 4°C and/or as frozen samples in cryovials as appropriate. - Algorithms, scripts and softwares: All the relevant algorithms, scripts and software code driving the project will be stored in a private online git repository from the GitHub account of the department (https://github.com/vibcbd). - Nucleic acid and protein sequences: All nucleic acid and protein sequences generated during the project will be stored on KU Leuven servers. Upon publication, all sequences supporting a manuscript will be made publicly available via repositories such as the GenBank database or the European Nucleotide Archive (nucleotide sequences from primers / new genes / new genomes), NCBI Gene Expression Omnibus (microarray data / RNA-seq data / CHIPseq data), the Protein Database (for protein sequences), the EBI European Genome-phenome Archive (EGA)

	for personally identifiable (epi)genome and transcriptome sequences.
How will the data be backed up?	KU Leuven drives are backed-up according to the following scheme: - data stored on the "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 "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. - data stored on the digital vault is backed up using snapshot technology, where all incremental changes in respect of the previous version are kept online. As standard, 10% of the requested storage is reserved for backups using the following backup regime: an hourly backup (at 8 a.m., 12 p.m., 4 p.m. and 8 p.m.), the last 6 of which are kept; a daily backup (every day) at midnight, the last 6 of which are kept; and a weekly backup (every week) at midnight between Saturday and Sunday, the last 2 of which are kept. - All omics data stored on the Flemish Supercomputer Centre (VSC) will be transferred on a weekly basis to the archive area which 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 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 eseries storage systems, and a CTDB samba cluster in the front-end. - the "J-drive" is based on a cluster of NetApp FAS8040 controllers with an Ontap 9.1P9 operating system.
What are the expected costs for data storage and backup during the project? How will these costs be covered?	The costs of digital data storage are as follows: 173,78€/TB/Year for the "L-drive" and 519€/TB/Year for the "J-drive".

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?	Both the “L-drive” and “J-drive” servers are accessible only by laboratory members, and 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. Access to the digital vault is possible only through using a KU Leuven user-id and password, and user rights only grant access to the data in their own vault. Sensitive data transfer will be performed according to the best practices for “Copying data to the secure environment” defined by KU Leuven. The operating system of the vault is maintained on a monthly basis, including the application of upgrades and security patches. The server in the vault is managed by ICTS, and only ICTS personnel (bound by the ICT code of conduct for staff) have administrator/root rights. A security service monitors the technical installations continuously, even outside working hours.

6. Data preservation after the end of the FWO project FWO expects that data generated during the project are retained for a period of minimally 5 years after the end of the project, in as far as legal and contractual agreements allow.	
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, ...).	The minimum preservation term of 5 years after the end of the project will be applied to all datasets.
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. For all other datasets, long term storage will be ensured as follows: - Digital datasets: files will be stored on the "L-drive". - Omics data: datasets will be stored on the "L-drive" or, for larger datasets, on the Vlaams Supercomputer Centrum. - Vectors: As a general rule at least two independently obtained clones will be preserved for each vector, both under the form of purified DNA (in -20°C freezer) and as a bacteria glycerol stock (-80°C). - Cell lines: cell lines will be stored locally in the laboratory (-80°C). - Other biological and chemical samples: storage at 4°C and/or as frozen samples in cryovials as appropriate.

What are the expected costs for data preservation during these 5 years? How will the costs be covered?	The total estimated cost of data storage during the 5 years after the end of the is to be 2000€ and will be covered by VIB dotation. This estimation is based on the different costs described in table 5.

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?	Participants to the present project are committed to publish research results 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).
Where/how will the data be made available for reuse?	X In an Open Access repository X Upon request by mail X Other (specify): open-access publications in peer-reviewed journals, including supplemental information As a general rule, datasets will be made openly accessible via existing platforms that support FAIR data sharing (www.fairsharing.org). Sharing policies for specific research outputs are detailed below: - Omics datasets will be deposited in open access repositories such as the PRIDE Archive for proteomics data, the EMBL-EBI platform for genomics and epigenomics data, or the NCBI Gene Expression Omnibus (GEO).

	- Vectors: Upon publication, all vectors supporting a manuscript will be made publicly available via the non-profit plasmid repository Addgene, along with the corresponding DNA sequences. Addgene in turns performs quality control on the DNA, curates the plasmids online with all relevant information (maps, sequences), and for a minimal cost (typically $65) ships the vectors upon simple request and signature of a material transfer agreement. The MTA will be prepared before depositing the vectors with the help of our organization's Tech Transfer office. For transfer between nonprofit or academic institutions, Addgene typically uses the Uniform Biological Material Transfer Agreement (https://www.addgene.org/terms/1047/). All non-published vectors and the associated documentation will be shared by the PI upon request and after signature of a material transfer agreement, at no cost except the cost of shipment. - Cell lines: All human pluripotent cell lines supporting publications will be registered in hPSCreg, the European human embryonic stem cell registry supported by the European Commission (https://hpscreg.eu/). Information about the deposited lines (including donor information, derivation method, availability and characterization) will also be made accessible. Registration of cell lines in hPSCreg will provide visibility, confirm ethical procurement and facilitate comparison with other hPSC lines. The PI will remain the distributor of the pluripotent cell lines. All other cell lines supporting publications will be deposited in the American Type Culture Collection (ATCC) database (https://www.atcc.org/), which is a private, non-profit biological resource center. This will provide a secure back-up for this material. Investigators can purchase cell lines from the ATCC database upon signature of a material transfer agreement (https://www.lgcstandards-atcc.org/~/media/PDFs/MTA_2.ashx) and, in some cases, of a Limited Use/Label License (e.g. for CRISPR products or iPSC materials) and/or a Customer Acceptance of Responsibility (for potentially highly pathogenic materials). Information about the cell lines (including organism, cell type, tissue, biosafety level and disease if applicable) will also be made accessible. - Genetically modified organisms: All genetically modified organisms used in publications will be made available to researchers upon request at the time of publication. - Other digital datasets that support publications (including image, video or audio files, electrophysiology data, cytometry data, spectroscopy data and simulation data) will be made publicly available via an open research data platform such as Mendeley Data or Zenodo.

	- Antibodies, synthetic and recombinant compounds: samples will be stored as appropriate in the laboratory. Within availability, they will be shared with interested researchers upon request. - Research documentation: All protocols used to generate published data will be described in the corresponding manuscript(s), and the related documentation will be included as supplementary information. These data and all other documents (daily logs, raw data) deposited in the E-Notebook are accessible to the PI and the research staff, and will be made available upon request. - Manuscripts: All scientific publications will be shared openly. Manuscripts submitted for publication will be deposited in a pre-print server such as bioRxiv, arXiv, Nature Precedings or ASAPbio). At the time of publication, research results will be summarized on the PI's website (add website address) and post-print pdf versions of publications will be made available there if allowed by copyright agreements, possibly after an embargo as determined by the publisher. Before the end of the embargo or in cases where sharing the post-print is not allowed due to copyright agreements, a pre-print version of the manuscript will be made available. Publications will also be automatically listed in our institutional repository, Lirias 2.0, based on the authors name and ORCID ID. - Algorithms, scripts and softwares: As soon as a manuscript is publicly available, the online private git repository containing the corresponding algorithms, scripts and software code will be changed to a public repository. - Nucleic acid and protein sequences: Upon publication, all sequences supporting a manuscript will be made publicly available via repositories such as the GenBank database or the European Nucleotide Archive (nucleotide sequences from primers / new genes / new genomes), NCBI Gene Expression Omnibus (microarray data / RNA-seq data / CHIPseq data), the Protein Database (for protein sequences). - Data that do not support publication will be either deposited in an open access repository or made available upon request by email.
When will the data be made available?	□ X Upon publication of the research results □ *Briefly comment.* As a general rule all research outputs will be made openly accessible at the latest at the time of publication. No embargo will be foreseen unless imposed e.g. by pending publications, potential IP requirements – note that patent application filing will be planned so that publications need not be delayed - or ongoing projects requiring confidential data. In those cases, datasets will be made publicly available as soon as the embargo date is reached.

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. As detailed above, metadata will contain sufficient information to support data interpretation and reuse, and will be conform to community norms. These repositories clearly describe their conditions of use (typically 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). Interested parties will thereby be allowed to access data directly, and they will give credit to the authors for the data used by citing the corresponding DOI. For data shared directly by the PI, a material transfer agreement (and a non-disclosure agreement if applicable) will be concluded with the beneficiaries in order to clearly describe the types of reuse that are permitted.
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	
Who will be responsible for the data documentation & metadata?	Metadata will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in the electronic laboratory notebook (E-notebook) that refer to specific datasets.
Who will be responsible for data storage & back up during the project?	The research and technical staff will ensure data storage and back up, with support from René Custers and Alexander Botzki for the electronic laboratory notebook and from Raf De Coster for the KU Leuven drives.

Who will be responsible for ensuring data preservation and sharing?	The PI is responsible for data preservation and sharing, with support from the research and technical staff involved in the project, from René Custers and Alexander Botzki for the electronic laboratory notebook and from Raf De Coster for the KU Leuven drives.
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.