FWO DMP Template - Flemish Standard Data Management Plan

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO's e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

The DMP template used by the Research Foundation Flanders (FWO) corresponds with the Flemish Standard Data Management Plan. This Flemish Standard DMP was developed by the Flemish Research Data Network (FRDN) Task Force DMP which comprises representatives of all Flemish funders and research institutions. This is a standardized DMP template based on the previous FWO template that contains the core requirements for data management planning. To increase understanding and facilitate completion of the DMP, a standardized **glossary** of definitions and abbreviations is available via the following link.

1. General Project Information	
Name Grant Holder & ORCID	**Bradley Balaton 0000-0002-5130-7868**
Contributor name(s) (+ ORCID) & roles	**Vincent Pasque (0000-0002-5129-0146) Principal Investigator**
Project number[1] & title	Modelling human X-chromosome inactivation using conversion of naïve human embryonic stem cells to the trophoblast lineage
Funder(s) GrantID[2]	**1263323N**
Affiliation(s)	x KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: Provide ROR[3] identifier when possible:

[1] "Project number" refers to the institutional project number. This question is optional since not every institution has an internal project number different from the GrantID. Applicants can only provide one project number.

[2] Funder(s) GrantID refers to the number of the DMP at the funder(s), here one can specify multiple GrantIDs if multiple funding sources were used.

[3] Research Organization Registry Community. https://ror.org/

Please provide a short project description	Mammals have evolved multiple mechanisms to ensure correct gene dosage of X-linked genes. These include inactivation of one of the two X chromosomes in females, a process that is critical for female development and linked to sex-specific susceptibilities to multiple diseases. Despite its importance in health and disease, very little is known about X chromosome inactivation in humans in part because most of the work so far has been carried out in rodents. Here I propose to model in vitro X chromosome inactivation in humans to identify key features of the process and to reveal essential factors and mechanisms. I will use a system recently established in the Pasque lab in which X chromosome inactivation is induced during differentiation of human naive pluripotent stem cells into trophoblast stem cells. This system will enable me to investigate the transcriptional and chromatin dynamics of human X chromosome inactivation and the factors involved in the process. I will use a combination of approaches including bulk and single-cell RNA-seq, degron cell lines, chromatin profiling and microscopy to comprehensively determine the temporal regulation of gene expression and chromatin dynamics during X chromosome inactivation in humans. Altogether, I aim to reveal conserved and species-specific differences in the molecular regulation of dosage compensation in mammals.

2. 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[4].

				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
Blastoid single cell sequencing data	We will use SMART-seq 3 to sequence blastoids with and without mutations to determine how the mutation affects X-linked gene expression	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☒ Experimental ☐ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ NA	Sequencing data	<1 TB	
Blastoid RNA-FISH data	We will use confocal microscopy and RNA-FISH to examine expression of X-linked genes in	Generate new data	Digital	Experimental	Image data	<1 GB	

[4] Add rows for each dataset you want to describe.

	blastoids						
Other blastoid and embryo datasets	We will download and analyze blastoid scRNA-seq datasets that have already been published	Reuse existing data	Digital	Aggregated data	Sequencing data, although I may be able to use already processed tabular data instead	<5 TB	

GUIDANCE:

DATA CAN BE DIGITAL OR PHYSICAL (FOR EXAMPLE BIOBANK, BIOLOGICAL SAMPLES, …). DATA TYPE: DATA ARE OFTEN GROUPED BY TYPE (OBSERVATIONAL, EXPERIMENTAL ETC.), FORMAT AND/OR COLLECTION/GENERATION METHOD.

EXAMPLES OF DATA TYPES: OBSERVATIONAL (E.G. SURVEY RESULTS, SENSOR READINGS, SENSORY OBSERVATIONS); EXPERIMENTAL (E.G. MICROSCOPY, SPECTROSCOPY, CHROMATOGRAMS, GENE SEQUENCES); COMPILED/AGGREGATED DATA[5] (E.G. TEXT & DATA MINING, DERIVED VARIABLES, 3D MODELLING); SIMULATION DATA (E.G. CLIMATE MODELS); SOFTWARE, ETC.

EXAMPLES OF DATA FORMATS: TABULAR DATA (.POR,. SPSS, STRUCTURED TEXT OR MARK-UP FILE XML, .TAB, .CSV), TEXTUAL DATA (.RTF, .XML, .TXT), GEOSPATIAL DATA (.DWG,. GML, ..), IMAGE DATA, AUDIO DATA, VIDEO DATA, DOCUMENTATION & COMPUTATIONAL SCRIPT.

DIGITAL DATA VOLUME: PLEASE ESTIMATE THE UPPER LIMIT OF THE VOLUME OF THE DATA PER DATASET OR DATA TYPE.

PHYSICAL VOLUME: PLEASE ESTIMATE THE PHYSICAL VOLUME OF THE RESEARCH MATERIALS (FOR EXAMPLE THE NUMBER OF RELEVANT BIOLOGICAL SAMPLES THAT NEED TO BE STORED AND PRESERVED DURING THE PROJECT AND/OR AFTER).

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.	GEO: **GSE179040 and other similar datasets**

[5] These data are generated by combining multiple existing datasets.

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, please describe these issues further and refer to specific datasets or data types when appropriate.	☒ Yes, human subject data ☐ Yes, animal data ☐ Yes, dual use ☐ No If yes, please describe: We will generate data for human cell lines. These are commercially available cell lines and we have ethical approval to generate and make publicly available this data.
Will you process personal data[6]? If so, briefly describe the kind of personal data you will use. Please refer to specific datasets or data types when appropriate. If available, add the reference to your file in your host institution's privacy register.	☐ Yes ☒ No If yes: - Short description of the kind of personal data that will be used: - Privacy Registry Reference:
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.	☐ Yes ☒ No If yes, please comment:
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 to what data they relate and what restrictions are in place.	☐ Yes ☒ No If yes, please explain:

[6] See Glossary Flemish Standard Data Management Plan

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 to what data they relate and which restrictions will be asserted.	☐ Yes ☒ No If yes, please explain:

3. 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).	**I am using electronic lab notebooks to keep track of wet lab details and jupyter notebook to keep track of code used for analysis. Code will be made available on github.** **I will keep track of annotations made during my analyses and include these with the processed data.**
Will a metadata standard be used to make it easier to **find 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. *Repositories could ask to deliver metadata in a certain format, with specified ontologies and vocabularies, i.e. standard lists with unique identifiers.*	☒ Yes ☐ No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: GEO requires metadata under the MINSEQE standard when submitting sequencing data. Additionally we will make it clear which cells are annotated with each sex and cell type as we believe this is important for reanalysis by others. If no, please specify (where appropriate per dataset or data type) which metadata will be created:

4. Data Storage & Back-up during the Research Project	
Where will the data be stored?	**Digital files will be stored on KU Leuven servers or on the Flemish Supercomputer Centre (VSC).** **Upon publication, sequencing data will be submitted to GEO and code submitted to github.** **All data will also be archived on KU Leuven servers or the VSC.**
How will the data be backed up? *WHAT STORAGE AND BACKUP PROCEDURES WILL BE IN PLACE TO PREVENT DATA LOSS? DESCRIBE THE LOCATIONS, STORAGE MEDIA AND PROCEDURES THAT WILL BE USED FOR STORING AND BACKING UP DIGITAL AND NON-DIGITAL DATA DURING RESEARCH.*[7] *REFER TO INSTITUTION-SPECIFIC POLICIES REGARDING BACKUP PROCEDURES WHEN APPROPRIATE.*	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. - All omics data stored on the Flemish Supercomputer Centre (VSC) will be transferred on a monthly basis to the archive area which is mirrored

[7] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

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 ☐ No If yes, please specify concisely: 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 “J-drive” is based on a cluster of NetApp FAS8040 controllers with an Ontap 9.1P9 operating system. - the Staging and Archive on VSC are also sufficiently scalable (petabyte scale) If no, please specify:
How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *Clearly describe the measures (in terms of physical security, network security, and security of computer systems and files) that will be taken to ensure that stored and transferred data are safe.* [7]	Both the VSC and KU Leuven servers require 2 factor authentication to access. Additionally, the work computers used to access these files are stored in a locked cabinet in a locked office.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?	The total estimated cost of data storage during the project is ~1,000 EUR. This estimation is based on the following costs: - The costs of digital data storage are as follows: 128,39€/TB/Year for the "L-drive" and 519EUR/TB/Year for the "J-drive". - The cost of VSC archive is 70 EUR/TB/Year, and staging 130EUR/TB/Year. - We expect costs to drop slightly during the coming four years. Additional budget for compute and data storage is budgeted for in ongoing projects, and will be costed in complementary project applications.

5. 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...).	**All sequencing and imaging data will be preserved locally for at least five years after publication. Those submitted to public repositories should be available indefinitely.**

Where will these data be archived (stored and curated for the long-term)?	**Sequencing data will be archived on the VSC and on GEO.** **Imaging data will be archived on the KU Leuven servers**
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	**As mentioned above, archival costs are in the range of 70-130 euros per TB per year. These costs will be covered through grants and ongoing projects.**

6. 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. *Note that 'available' does not necessarily mean that the data set becomes openly available, conditions for access and use may apply. Availability in this question thus entails both open & restricted access. For more information: https://wiki.surfnet.nl/display/standards/info-eu-repo/#infoeurepo-AccessRights*	☒ Yes, in an Open Access repository ☐ Yes, in a restricted access repository (after approval, institutional access only, …) ☐ No (closed access) ☐ Other, please specify:
If access is restricted, please specify who will be able to access the data and under what conditions.	
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, privacy aspects ☐ Yes, intellectual property rights ☐ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ☒ No If yes, please specify:
Where will the data be made available? If already known, please provide a repository per dataset or data type.	**The data will be published in the Gene Expression Omnibus (GEO)**

When will the data be made available? *This could be a specific date (dd/mm/yyyy) or an indication such as 'upon publication of research results'.*	**Data will be made available upon publication**
Which data usage licenses are you going to provide? If none, please explain why. *A data usage license indicates whether the data can be reused or not and under what conditions. If no licence is granted, the data are in a grey zone and cannot be legally reused. Do note that you may only release data under a licence chosen by yourself if it does not already fall under another licence that might prohibit that.* *Example Answer: E.g. "Data from the project that can be shared will be made available under a Creative Commons Attribution license (CC-BY 4.0), so that users have to give credit to the original data creators."* [8]	*Data from the project that can be shared will be made available under a Creative Commons Attribution license (CC-BY 4.0), so that users have to give credit to the original data creators*
Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *Indicate whether you intend to add a persistent and unique identifier in order to identify and retrieve the data.*	☒ Yes ☐ No If yes:
What are the expected costs for data sharing? How will these costs be covered?	**GEO is funded by the NIH and is free to host our data.**

[8] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	**The award holder (Bradley Balaton) will manage data documentation and metadata during the research project**
Who will manage data storage and backup during the research project?	**The award holder will manage data storage and back.**
Who will manage data preservation and sharing?	**The award holder will ensure that the data is preserved and publicly available at the end of the project. The principle investigator (Vincent Pasque) will handle data preservation and sharing after the project has ended.**
Who will update and implement this DMP?	**The award holder will update and implement this DMP.**