DATA MANAGEMENT PLAN (DMP)

FWO PROJECT: **11N3122N**

Thi Xuan Ai Pham, KU Leuven

Laboratory for Epigenetic Reprogramming

Stem Cell Institute, KU Leuven

Principle investigator: Vincent Pasque

NOVEL STEM CELL-BASED HUMAN EMBRYO MODELS AND SINGLE-CELL TECHNOLOGIES TO STUDY EARLY EMBRYONIC DEVELOPMENT

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

KU LEUVEN

Describe the origin, type and format of the data (per dataset) and its (estimated) volume	I will generate, collect, process, analyze and store the data listed below, as detailed in the project description. (not relevant to this section) The following datasets will be generated: **1. General** The major aim of the project is the analysis of single-cell epigenomic and single-cell transcriptomic data. This data will be acquired on in vitro models, including cells obtained by reprogramming somatic cells into induced pluripotent stem cells or cells derived from induced pluripotent stem cells and/or embryonic stem cell lines. **1. Experimental data** **Dataset 1.1. – Digital images** Microscopy pictures, gel scans, graphs, illustrations, figures. **Dataset 1.2. – Cytometry data** Flow Cytometry and fluorescence-activated cell sorting (FACS) data **Dataset 1.3. – Omics data** Genomics, transcriptomics, epigenomics data. **2. Derived and compiled data** **Dataset 2.1 – Research documentation** Research documentation generated by myself and collaborators or collected from online sources and from collaborators, including ethical approval documents, laboratory notes, protocols. **Dataset 2.2 – Manuscripts** The results will be published as BioRxiv preprints and articles in peer reviewed journals. **Dataset 2.3 – Algorithms and scripts** Algorithms and scripts to investigate the single-cell data and to integrate different omics layers (transcriptome/epigenome) will be designed. **Dataset 2.4 – Software** The combined set of scripts, algorithms, visualization tools & computer programs. **3. Canonical data** **Dataset 3.1 – Nucleic acid sequences** These datasets are the backbone of my PhD project, but after publishing will also represent an important source of information for others studying early human embryonic development. Data will be stored in the following formats:

	- Text files: Plain text data (Unicode, .txt), MS Word (.doc/.docx), Adobe Portable Document Format (.pdf), LaTex (.tex) format; - Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tsv), delimited text (.txt), MS Excel (.xls/.xlsx); - 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); - Nucleotide sequences: raw sequence data trace (.ab1), text-based format (.fasta(.gz)/.fa(.gz)) and accompanying QUAL file (.qual), Genbank format (.gb/.gbk); - Next generation sequencing raw data: binary base call format (.bcl) - Sequence alignment data: (.sam), .bam - Coverage data: .bed, .bg, .bedGraph, .bw, .bigwig - Structural variations data: .vcf(.gz), .bcf - Read/UMI count data: .tsv(.gz), Matrix Market format (.mtx), .loom, .rds(.gz) - Nucleic acid samples resulting from (single-cell) nucleic acid amplification, or sequence library preparations will be stored in labeled tubes or SBS plates in -20C freezers. We have electronic laboratory databases that will keep the physical storage address of these samples. Raw as well as processed data will be submitted to a public repository in the aforementioned described standard formats, to enable sharing and long-term validity of the data.

2. 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.	x No □ Yes
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).	□ No X Yes - Ethical approval S64962: Characterization and functional assessment of the regulatory landscape of embryonic lineages using blastoids as a model. - Ethical approval (S66375) to work on human embryos with Dr. Laurent David (in Nantes) - Ethical approval to work on human embryos with Dr. Hilde Van Devele (in Brussels) (S66184): "Chromatin regulators perturbations in human embryos (CREPE). Human embryo inhibitor (PRC2) treatment"

Does your work possibly result in research data with potential for tech transfer and valorization? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?	☐ No X Yes I do not exclude that the proposed work could result in research data with potential for tech transfer and valorization. The KU Leuven 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?	X No ☐ Yes No third-party agreement restricts dissemination or exploitation of the data from this project.

3. Documentation and metadata	
What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?	Documentation will consist of notes in the electronic laboratory notebook (E-notebook) that refer to specific datasets. Those notes will describe the biological samples used, experimental setup and protocols used, sequences generated, the links to the specific computer location as well as the names of the respective datasets. I will also maintain a metadata sheet with the connection between lab samples and files on our data storage, so that data files, lab samples, and experimental notes remain properly linked. Algorithms, scripts and software usage will be documented, e.g. using Jupyter Notebooks. When scripts, algorithms and software tools are finalized, they will be additionally described in manuscripts and on GitHub (see www.github.com/pasquelab for our previous datasets and scripts).

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 Sequencing data types require particular metadata, such as data submitted to EGA, GEO, SRA, ArrayExpress, or ENA. Local data that is not (yet) submitted to these resources will be based on 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. When depositing data in a local or public repository, 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.

4. Data storage & backup during the FWO project	
Where will the data be stored?	Digital files will be stored on KU Leuven servers. - 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 + archive area and later only in the archive area (archive is mirrored). - 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. - Algorithms, scripts and software: All the relevant algorithms, scripts and software code driving the project will be stored in private online git repositories of the PI (e.g., .https://github.com/pasquelab). As soon as the manuscript is publicly available, the repository will be changed to a public repository. - 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. - Other data files (Digital images, cytometry data, etc..) will be stored on local KU Leuven servers and PI computers.
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. - All omics data stored on the Flemish Supercomputer Centre (VSC) will be transferred on a monthly basis to the archive area which is mirrored.

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.	□ No X 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 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)
What are the expected costs for data storage and backup during the project? How will these costs be covered?	The total estimated cost of data storage during the project is ˜9,000 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.
Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?	- 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. - The VSC storage is only accessible to VSC accounts, and specifically our volume is only accessible to group members. - No personal data will be stored on the VSC nor local servers.

5. Data preservation after the end of the FWO project KU Leuven 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 or preprint deposition. For all other datasets, long term storage will be ensured as follows: - Large sequencing/omics data: will be stored on VSC archive - Small digital files: files will be stored on the "L-drive". - Developed algorithms and software will be stored on VSC archive and/or L-drive, as well on public repositories such as Github.com. - Third-party software and algorithms that are used are referenced by their version numbers (e.g., in our Jupyter notebooks), and are installed as modules on the VSC and/or containers (Docker, Singularity) on the VSC, to ensure reproducibility.

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 2,500 EUR. This estimation is based on 70EUR/Tb/year. The storage after the project is much smaller because during the project a large working space is needed, and post-publication data will be made accessible via open access platforms. The costs for this data preservation will be upfront paid for from this FWO project

6. 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)?	X No □ Yes
Which data will be made available after the end of the project?	The PI in the present project is 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 In a restricted access repository □ 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 EMBL-EBI platform for genomics and epigenomics data, or the NCBI Gene Expression Omnibus (GEO), the EBI ArrayExpress databases for functional genomics data or the EBI European Genome-phenome Archive (EGA) for personally identifiable genetic and phenotypic data. Data at EGA will be collected from individuals whose consent agreements authorise data release only for specific research use to bona fide researchers. - Other digital datasets that support publications (including image, cytometry data, and simulation data) will be made publicly available via an open research data platform such as Mendeley Data or Zenodo. - 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 (http://www.kuleuven.be/pasquelab) 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 added to our institutional repository, Lirias 2.0, based on the authors name and ORCID ID (the metadata will be added, not the full manuscripts). - Algorithms, scripts and software: All the relevant algorithms, scripts and software code driving the project will be made available to public repositories such as www.github.com/pasquelab - Nucleic acid and protein sequences: All nucleic acid and protein sequences generated during the project will 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. - 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?	□ Immediately after the end of the project X Upon publication of the research results □ After an embargo period. 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 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. A CC-BY license will be opted for when possible. 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. For KU Leuven data submitted to the EBI European Genome-phenome Archive (EGA), which operates under controlled access, the data access/submission requests will be received by the operational committee and processed in consultation with the PIs produced data. The advisory committee will provide general guidance in terms of policies and will be referred to in handling controversial cases.
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 (internally generated operational procedures (https://en.wikipedia.org/wiki/Standard_operating_procedure)) 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.

7. 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?	I will ensure data storage and back up, with support from ICTS, HPC, and gbiomed-IT staff.
Who will be responsible for ensuring data preservation and sharing?	The PI is responsible for data preservation and sharing, with support from ICTS, HPC, and gbiomed-IT staff.
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.