Enhancer-AI: AI-driven modelling and design of cell type specific enhancers for gene therapy

A Data Management Plan created using DMPonline.be

Creators: Patrick Vandormael, n.n. n.n., n.n. n.n.

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: n.n. n.n.

Project Administrator: Patrick Vandormael

Grant number / URL: S005024N

ID: 204304

Start date: 01-10-2023

End date: 31-10-2027

Project abstract:

Gene therapy offers the promise of an efficient, single-dose therapeutic solution for many incurable diseases. In reality, despite large efforts of the scientific community and a strong industrial interest, bringing gene therapy to the market has proven challenging. The major hurdles faced by gene therapy come in the form of its safety and efficacy with off- target effects caused by low specificity or inappropriate transgene expression levels. The use of well-designed synthetic regulatory regions, called enhancers, could provide a solution to reach cell type-specificity and high levels of transgene expression. This SBO project proposes to develop new computational and experimental tools, and combine them in a pipeline to identify enhancers in complex tissues. This pipeline will exploit recent advances in single-cell multi-omics, gene regulatory network (GRN) inference, and deep learning. We will use this pipeline to design enhancers that are specifically active in three chosen brain cell types that are of high relevance for gene therapy. Practically, we will: 1) use mouse and human brain samples to generate a single-cell multi- omic atlas; 2) select unique regulatory regions to each cell type and train enhancer models; and 3) design and validate synthetic enhancers in vivo, using massively parallel reporter assays. As a clinically relevant case study, we will focus on microglia enhancers in the context of Alzheimer's disease. In parallel, we will use the AI-based GRN and enhancer models to interpret and prioritize regulatory variation in whole genome sequences, in order to improve diagnosis and prediction of risk and progression for neurodegenerative disease. Our project will yield licensable enhancers, software tools, and diagnostic AI models with direct industrial applicability, and will demonstrate our ability to interpret and generate enhancers specific to any cell type, which can find application in a wide range of diseases, even beyond the scope of our project.

Last modified: 05-03-2024

Enhancer-AI: AI-driven modelling and design of cell type specific enhancers for gene therapy
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

Questionnaire

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

Question not answered.

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)

Question not answered.

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)

Question not answered.

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)

Question not answered.

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

Question not answered.

Enhancer-AI: AI-driven modelling and design of cell type specific enhancers for gene therapy
FWO DMP (Flemish Standard DMP)

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
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Task 1.1: SCENIC+	Computational workflow for determining eGRNs from multi-ome data	reused (own)	digital	Derived and compiled data			
Task 1.1: deepSCENIC	Model to predict eGRNs using VAE and CNN	new	digital	Derived and compiled data			
WP 1 & 2: existing scRNA-seq & scATAC-seq datasets	publicly available datasets from mouse and human brain	reused (public)	digital	Experimental measurement, quantitative	- textual data: FASTQ file (.fastq, zipped as .gz) -metadata: textual data (.rtf, .xml, .txt)		
WP 1 & 2: existing scRNA-seq & scATAC-seq datasets	previously generated datasets from mouse and human brain	reused (own)	digital	Experimental measurement, quantitative	- textual data: FASTQ file (.fastq, zipped as .gz) -metadata: textual data (.rtf, .xml, .txt)		

WP 1 & 2: scRNA-seq	Single-cell transcriptional profiling of mouse, macaque, and human postmortem brain	new	digital	Experimental measurement, quantitative	- Raw: binary base call format (.bcl) - textual data: FASTQ file (.fastq, zipped as .gz) -metadata: textual data (.rtf, .xml, .txt)		
WP 1 & 2: scATAC-seq	Profiling genome-wide chromatin accessibility at single-cell resolution of mouse, macaque, and human post-mortem brain	new	digital	Experimental measurement, quantitative	- Raw: binary base call format (.bcl) - textual data: FASTQ file (.fastq, zipped as .gz) -metadata: textual data (.rtf, .xml, .txt)		
Task 1.2: DeepTopic	Model for training and evaluation of deep learning models from output of SCENIC+/deepSCENIC and cisTopic	new	digital	Derived and compiled data			
Task 1.2: DeepExplainer	Tool for explaining model predictions including nucleotide dependencies	new	digital	Derived and compiled data			
Task 1.3: GAN	Generative Adversarial Network for the generation of fully synthetic enhancers	new	digital	Derived and compiled data			
Task 1.4: MPRA analysis pipeline	single pipeline combining MPRA analysis steps using NextFlow	new	digital	Derived and compiled data			
Task 1.5: regulatory polygenic risk score algorithm	Model to predict PD disease risk based on eGRNs and enhancer models	new	digital	Derived and compiled data			
Task 1.5: regulatory polygenic risk score data	whole genome sequence, and brain single-cell multi-ome data of PD patients and controls	reused (own)	digital	Experimental measurement, quantitative	- textual data: FASTQ file (.fastq, zipped as .gz) -metadata: textual data (.rtf, .xml, .txt)		
Task 2.1-2.4, 3.2-3.3: MPRA libraries	Lenti-MPRA and AAV-MPRA plasmid libraries	new	physical	Biological and chemical samples: samples stored at -80°C	N/A		~200 µl per sample

Task 2.2 & 3.2: excitatory neuron MPRA	bulk and single-cell MPRA (sequencing) of cortical excitatory neuron-specific enhancers	new	digital	Experimental measurement, quantitative	- Raw: binary base call format (.bcl) - textual data: FASTQ file (.fastq, zipped as .gz) -metadata: textual data (.rtf, .xml, .txt)		
Task 2.3 & 3.2: Dopaminergic neuron MPRA	bulk and single-cell MPRA (sequencing) of dopaminergic neuron-specific enhancers	new	digital	Experimental measurement, quantitative	- Raw: binary base call format (.bcl) - textual data: FASTQ file (.fastq, zipped as .gz) -metadata: textual data (.rtf, .xml, .txt)		
Task 2.4 & 3.2: Microglial MPRA	bulk and single-cell MPRA (sequencing) of microglia-specific enhancers in human and mouse cell line	new	digital	Experimental measurement, quantitative	- Raw: binary base call format (.bcl) - textual data: FASTQ file (.fastq, zipped as .gz) -metadata: textual data (.rtf, .xml, .txt)		
Task 3.3: spatial MPRA	MRPA (sequencing) to determine enhancer activity with spatial resolution in the mouse brain	new	digital	Experimental measurement, quantitative	- Microscope data: raw images (.nd2) - metadata: textual data (.rtf, .xml, .txt)		
Task 4.1-4.2: snRNA-Seq	testing effect of enhancer-driven shRNA on other brain cell types	new	digital	Experimental measurement, quantitative	- Raw: binary base call format (.bcl) - textual data: FASTQ file (.fastq, zipped as .gz) -metadata: textual data (.rtf, .xml, .txt)		

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:

Own datasets/tools

WP/Task	dataset/tool	source
Task 1.1	SCENIC+	https://doi.org/10.1038/s41592-023-01938-4
WP1 & 2	mouse multi-ome (scRNA-Seq & scATAC-Seq)	https://doi.org/10.7554/eLife.73971
WP1 & 2	human multi-ome (scRNA-Seq & scATAC-Seq)	unpublished, generated in-house
Task 1.5	human whole genome and single-cell multi-ome of PD and control	unpublished, generated in-house through ASAP consortium

External datasets/tools

WP/Task	dataset/tool	source
WP1 & 2	mouse multi-ome (scRNA-Seq & scATAC-Seq)	publicly available datasets such as those described in our previous studies (De Rop et al, 2022 eLife)
WP1 & 2	human multi-ome (scRNA-Seq & scATAC-Seq)	publicly available datasets such as those described in our previous studies (Bravo González-Blas et al, 2023 Nature Methods)
WP1 & 2	human microglia (scRNA-seq & scATAC-seq)	https://doi.org/10.1016/j.cell.2023.08.037

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
- Yes, animal data

Our study involves the creation and re-use of data derived from both human bodily materials (HBM) and laboratory animal experiments. We are committed to ensuring all ethical considerations are addressed.

Human Bodily Materials and Data

Data creation:

- This study will generate new sequencing data from human samples. We will seek approval from Ethics Committee Research UZ / KU Leuven for the use of HBM for this purpose to ensure ethical compliance. Some of these efforts are already ongoing, such as the analysis of samples from the ASAP consortium, and have already obtained ethical approval.
- We will also be re-using sequencing data that was previously generated by consortium members. We will ensure that this use falls under the previous approvals granted by the Ethics Committee.
- For the *ex vivo* study of human organotypic brain slices, we will be using patient brain material that is being collected in an existing study. The JDW lab already has approval for the use of these materials.

Data re-use:

We will be re-using sequencing data generated by others. In doing so, we will adhere to the Data Transfer Agreement (DTA) and the Data Access Committee (DAC) guidelines of the data provider.

Laboratory Animal Experiments

We will be conducting experiments on mice and will seek approval from the Ethical Committee for Animal Experimentation before their initiation. For ex vivo experiments involving Rhesus macaques, we already have obtained approval before the start of this project. We will seek additional approval for any in vivo experiments involving Rhesus macaques that will be conducted during the project. These approvals have not been requested yet, as the design of their protocol will depend on the outcomes of preceding work packages.

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

The types of personal data that will be processed or generated in this study are: age, gender, health data, and (epi)genetic data. The personal data that will be provided from collaborating clinicians will always be pseudonymized or anonymised.

For personal and sensitive data, we will abide by the Belgian law on the protection of individuals with regard to the processing of personal data (30th July 2018) and the General Data Protection Regulation 2016/679. The Privacy Team of KU Leuven will be notified before the start of the data processing activities via the PRET questionnaire (upon registering the study with the Clinical Trial Center and requesting ethical approval) and the Data Steward therefore:
- has designated the categories of persons who have access to the sensitive data, with a precise description of their capacity in relation to the processing of these data;
- keeps the list of the designated categories of persons at the disposal of the competent supervisory authority (Data Protection Authority);
- has ensured that the designated persons are obliged by a legal or statutory obligation, or by an equivalent contractual provision, to observe the

confidential nature of the data concerned.

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

We expect that the proposed work could result in research data with potential for tech transfer and valorization. Ownership of the data generated belongs to KU Leuven and VIB in accordance with the framework agreement of both institutes. 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 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.

- No

No third-party agreement restricts dissemination or exploitation of the data or strains generated from this project. Existing agreements between VIB and KU Leuven do not restrict publication of data.

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.

- Yes

The consortium partners will execute a Cooperation Agreement for this SBO project. The parties have agreed that the terms of the Cooperation Agreement will be largely governed by Article 4 of the model Cooperation Agreement which is available on the FWO website under "Explanatory Document on the Collaboration Agreement Strategic Basic Research". A refinement of the IPR agreements will be possible when setting up the final Cooperation Agreement, but the overall principles will not be altered. As mentioned in the project proposal, VIB will take the lead in valorization of the SBO project deliverables.

At this moment in time, no agreements related to the EnhancerAI platform have been established between the consortium partners and members of the Advisory Committee. Of note, the sequences of the novel synthetic enhancers will not be shared with the members of the Advisory Committee.

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).

Data will be accompanied by **documentation** containing all contextual and descriptive features of the research data, which allow to understand and (re)use the data. This includes data collection methods, protocols, and code explanation. Documentation is stored at the study- and the data-level, providing data provenance from the original source data to specific datasets linked to publications. Data will be generated following standardized **protocols**. Clear and detailed descriptions of these protocols will be stored in our lab protocol database and electronic laboratory notebook (E-notebook) and published along with the results, eg. on protocols.io (https://www.protocols.io/workspaces/aertslab/publications). **Algorithms, scripts and software** usage will be documented,e.g. using Jupyter Notebooks. Internally, we use git.aertlab.org to save and version the scripts. When scripts,algorithms and software tools are finalized, they will be additionally described in manuscripts and on GitHub (seewww.github.com/aertslab for our previous scripts and tools). **Metadata** will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in the E-notebook and/or in hard copy lab notebooks that refer to specific datasets.

All datasets will be accompanied by metadata that is stored in our electronic lab notebook and in our central samplesheet. We have scripts that process the metadata, for example to obtain all fastq files of a certain project.

Digital files will be **named** following a standard procedure, so that all the name of all files in a given dataset will be in the same format.

Raw data is named as:
SEQUENCER_NAME_YYYYMMDD/PROJECT__CUAL__NAME_*:
- SEQUENCER_NAME: e.g. NovaSeq6000, NextSeq2000
- YYYYMMDD: Sequencing date
- PROJECT: 3 character project code
- 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

To allow long term access and use of research data will be stored or converted to **open file formats** as much as possible.

- Containers: TAR, ZIP
- databases: XML, CSV, JSON
- Statistics: DTA, POR, SAS, SAV
- Images: TIFF, JPEG 2000, PNG, GIF
- Tabular data: CSV, TXT
- Text: XML, PDF/A, HTML, JSON, TXT, RTF
- Sequencing data: FASTA, FASTQ

We use **controlled vocabularies** or ontologies when applicable to provide unambiguous meaning, for example:

- Gene Onotology: molecular function, cellular component, and biological role of RNA seq
- ENSEMBL or NBCI identifiers: gene identity
- HUGO Gene Nomenclature Committee: names and symbol of human genes
- Mouse Genome Informatics: names and symbol of mouse genes
- FlyBase: names and symbol of Drosophila genes
- Chicken Gene Nomenclature Committee: names and symbol of chicken genes
- UniProt protein accessions: protein identity

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

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) and/or in hard copy lab notebooks that refer to specific datasets. All datasets will be accompanied by a README.txt file containing all the associated metadata, which will include the following elements:

- Title: free text
- Creator: Last name, first name, organization
- Date and time reference
- Subject: Choice of keywords and classifications
- Structure: internal structure of the dataset, or the meaning of abbreviations (not necessary when it is clear from the in-file documentation).
- 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. Additionally, we will closely monitor MIBBI (Minimum Information for Biological and Biomedical Investigations) for metadata standards more specific to our data type.

For specific datasets, additional metadata will be associated with the data file as appropriate.

3. Data storage & back-up during the research project

Where will the data be stored?

Digital data

- Primary storage for **active digital files** will be on KU Leuven servers. KU Leuven offers fast ("J-drive) and slower ("L-drive") storage that allows reading/writing/modification of non-confidential, confidential, and strictly confidential data.

- KU Leuven further offers the ManGO platform for storage and management of **large volumes of active research data**. This platform allows secure storage, manual and automated metadata coupling, data workflows, and file sharing.
- Data that is no longer active, can be **archived** on the KU Leuven K-drive, which allows reading of non-confidential, confidential, and strictly confidential data.
- **Personal data** of human subjects will be stored on a dedicated KU Leuven secure server (Digital vault).

- **Algorithms, scripts and software**: All the relevant algorithms, scripts and software code will be stored on the lab GitHub account (https://github.com/aertslab).
- **Omics** data: omics data generated during the project will be stored on KU Leuven servers or on the ManGO platform. 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), or the EBI European Genome-phenome Archive (for human (epi)genome and transcriptome sequences).

Physical samples

- Tissue samples: Tissues will be stored locally in the laboratory. All human tissue samples will be registered with a Belgian biobank, in compliance with the Belgian law on human body material (dd 19-12-2008).
- 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 bacterial 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 cell lines will be stored locally in the laboratory in liquid nitrogen storage and will be deposited in the UZ Leuven-KU Leuven Biobank. Other human cell lines will be stored locally in liquid nitrogen cryostorage of the laboratory when actively used for experiments. Animal cell lines will be stored in liquid nitrogen cryostorage of the laboratory.
- Bacterial strains will be stored in a -80°C freezer.
- 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.

How will the data be backed up?

KU Leuven drives are backed-up according to the following scheme:

- data stored in manGO: 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 “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.

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 offer sufficient storage for active data (J/L-drive, ManGO) and archived data (K-drive). Required data-storage volumes can be easily scaled up.

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

The buildings on our campus are restricted by badge system so only employees are allowed in and visitors are allowed under supervision after registration.

Access to the “L-drive”, “J-drive”, and ManGO servers is possible only through using a KU Leuven user-id and password, and user rights only grant access to their own data, or data that was shared to 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.

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. Only the PI and medical team members will be granted access to the server to deposit private data. The PI and medical team members will be the only responsible for linking patient information and/or samples, and will strictly respect confidentiality.

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 are as follows: 569,2€/5TB/Year for the “L-drive”, 519€/TB/Year for the “J-drive”, and 35€/TB/Year for the ManGO platform. Data storage and backup costs are included in general lab costs.

-Maintaining a mouse colony alive costs about 1,200 euro per year (for 6 cages), excluding the costs of genotyping. When no experiment is planned with a particular mouse strain, and in compliance with the 3R’s rule (https://www.nc3rs.org.uk), cryopreservation will thus be used to safeguard the strain, prevent genetic drift, loss of transgene and potential infections or breeding problems. Cryopreservation of sperm/embryos costs about 500 to 700 euro per genotype, plus a minimal annual storage fee (25 euro per strain for 250 to 500 embryos). Frozen specimen are kept in two separate liquid nitrogen tanks at two different sites on campus. When necessary, the costs of revitalization from cryopreserved sperm/embryos are about 1,100/600 euro.

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...).

According to KU Leuven RDM policy, relevant research data will be preserved on the university's servers for a minimum of 10 years. Such data include data that are at the basis of a publication, that can only be generated or collected once, that are generated as a result of a substantial financial or personal effort, or are likely to be reused within the research unit or in wider contexts.

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

As a general rule all research outputs (data, documentation, and metadata) related to publications will be made openly accessible, whenever possible via existing platforms that support FAIR data sharing (www.fairsharing.org). We aim at communicating our results in top journals that require full disclosure upon publication of all included data, either in the main text, in supplementary material or in a separate data repository.
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 are as follows: 569,2€/5TB/Year for the "K-drive" and the “L-drive”, 519€/TB/Year for the “J-drive”, and 35€/TB/Year for the ManGO platform. Data storage and backup costs are included in general lab costs.

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, …)

- Upon publication, datasets and metadata generated from **animal omics** will be stored in public repositories such as Zenodo or the NCBI Gene Expression Omnibus, where they will receive a unique and persistent identifier.
- Datasets and metadata generated from **human omics** will be deposited under restricted access on the European Genome Phenome Archive (EGA), where they will be assigned a unique and persistent identifier.
- **Computational workflows, models**, and metadata will be stored on platforms such as Github, Kipoi, and Zenodo with proper versioning.
- To ensure data findability, links and references these datasets, workflows and modes will be included in the data availability statements of the associated publication.

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

Access to restricted access dataset (such as human omics datasets) is governed by the Data Access Committees of KULeuven/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, Intellectual Property Rights
- Yes, Privacy aspects

- Human omics data are considered sensitive personal data, and are are only made available on restricted access repositories such as the European Genome Phenome Archive (EGA). Access to these datasets is under control of a Data Access Committee.
- The researchers involved and the IP team of the VIB TechTransfer Office shall make the necessary arrangements in order to maintain an embargo on the public access of research data, at least until the essential steps in securing intellectual property (e.g. the filing of a patent application) have been taken. 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.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

- Upon publication, datasets and metadata generated from animal omics will be stored in public repositories such as Zenodo or the NCBI Gene Expression Omnibus, where they will receive a unique and persistent identifier.
- Datasets and metadata generated from human omics will be deposited under restricted access on the European Genome Phenome Archive (EGA), where they will be assigned a unique and persistent identifier.
- Computational workflows, models, and metadata will be stored on platforms such as Github, Kipoi, and Zenodo with proper versioning.
- protocols will be deposited on protocols.io.

When will the data be made available?

All research outputs (data, documentation, code, and associated metadata) 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.

Which data usage licenses are you going to provide? If none, please explain why.

Data is typically available 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. 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

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.

6. Responsibilities

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

The researchers who generate the data are responsible for managing data, documentation, and metadata.

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

The researchers who generate the data are responsible for storage and backup, with support from René Custers and Alexander Botzki for the electronic laboratory notebook (ELN) and from Raf De Coster for the KU Leuven drives.

Who will manage 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 (ELN) and from Raf De Coster for the KU Leuven drives.

Who will update and implement this DMP?

The PI is ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.