DMP - FWO - Fonds Wetenschappelijk Onderzoek - 1273822N

The participants to the present project understand the value of FAIR Data Management and Open Access to Scientific Publications and Research Data. They are fully committed to abiding by the related FWO policies.

The present data management plan (DMP) describes the specific outputs of this project and how they will be made available to the community. All participants will be informed of updates in the DMP, and any new participant will receive training to ensure compliance with the consortium's data management conventions.

1. General Information	
Name applicant	Stein Aerts - stein.aerts@kuleuven.vib.be
FWO Project Number & Title	1273822N Deciphering the evolution of gene regulatory networks and enhancer architectures in neuronal cell types using single-cell deep learning and comparative genomics
Affiliation	VIB-KU Leuven Center for Brain & Disease Research

Responsible: Nikolai Hecker - nikolai.hecker@kuleuven.be

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

Describe the origin, type and format of the data (per dataset) and its (estimated) volume	Observational data Tissue samples - Biological and chemical samples: frozen samples -80°C (temporarily until used in experiments) - purpose: frozen brain samples from animal cadavers, chicken and mouse, will be used for single cell omics and spatial transcriptomics experiments. Samples have previously been obtained in the lab and will be generated in the course of the project Experimental data Digital images Microscopy pictures, digital images in raster formats - H&E and DAPI stainings, maximum intensity projections related to spatial transcriptomics (.tiff, .jpg, .png), ~50GB - Fluorescence images from MPRA/enhancer essays for candidate enhancer validation (.tiff, .png), ~50GB Graphs, illustrations, figures, digital images in vector formats - schematics, visualization of data and results (.svg, .eps, .ai), ~5GB Omics data Single cell transcriptomics and chromatin accessibility: chicken telencephalon and mouse brain multiome (scRNAseq+scATACseq) -Next generation sequencing raw data: binary base call format (.bcl), .fastq (.gz) -Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb) -Sequence alignment data: (.sam), .bam -Coverage data: .bed, .bg, .bedGraph, .bw, .bigwig -Read/UMI count data: .tsv(.gz), Matrix Market format (.mtx), .loom, .rds(.gz), Hierarchical Data Format (.h5, .h5ad)

	-purpose: identification/comparative analysis of cell brains, gene regulatory networks and enhancers -expected size: ~500G-2TB Spatial transcriptomics: chicken telencephalon and mouse brain -Next generation sequencing raw data: binary base call format (.bcl), .fastq (.gz) -Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb) -Sequence alignment data: (.sam), .bam -Coverage data: .bed, .bg, .bedGraph, .bw, .bigwig -Read/UMI count data: .tsv(.gz), Matrix Market format (.mtx), .loom, .rds(.gz), Hierarchical Data Format (.h5, .h5ad) -purpose: used for localization of identified cell types, spatial mapping of single cell omics data -expected size: ~100GB-1TB Simulation data 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, including laboratory notes and protocols. -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;

	Manuscripts
	The results will be published as BioRxiv preprints and articles in peer reviewed journals -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); Algorithms and scripts Scripts and algorithms for analyzing and to integrate the single-cell omics, spatial transcriptomics and genome; for comparative genomics analysis; for training, applying, and evaluating deep learning learning models - Notebooks in Jupyter (.ipynb) and R (.Rmd) - Standalone scripts and modules in Python (.py), (Perl (.pl, .pm), R (.R)) - Nextflow pipelines (.nf) - Shell scripts in BASH (.sh) Software The combined set of scripts, algorithms, visualization tools & computer programs.

	Canonical data Nucleic acid sequences Human motor cortex SNAREseq v2, single nucleus multiome data -source: NeMo archive, https://assets.nemoarchive.org/dat-ek5dbmu -publication: Bakken el al. (2021). Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature, 598(7879), 111-119. -copyright: CC BY 4.0 -SnapTools fragment file format (.snap) -Read/UMI count data: .tsv(.gz), Matrix Market format (.mtx), .loom, .rds(.gz), Hierarchical Data Format (.h5, .h5ad) -Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); - purpose: comparative analysis of cell types, gene regulatory networks -size: ~1TB Mouse brain single nucleus ATACseq -source: NeMo archive, https://assets.nemoarchive.org/dat-wywv153 -publication: Li et al. (2021). An atlas of gene regulatory elements in adult mouse cerebrum. Nature, 598(7879), 129-136. -copyright: CC BY 4.0 -Read/UMI count data: .tsv(.gz), Matrix Market format (.mtx), .loom, .rds(.gz), Hierarchical Data Format (.h5, .h5ad) -Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); -size: ~20TB - purpose: comparative analysis of cell types, gene regulatory networks Mouse brain single nucleus RNAseq - source: ArrayExpress, E-MTAB-11115, https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-11115/ -publication: Kleshchevnikov et al. (2022). Cell2location maps fine-grained cell types in spatial

	transcriptomics. Nature biotechnology, https://doi.org/10.1038.
	-Read/UMI count data: .tsv(.gz), Matrix Market format (.mtx), .loom, .rds(.gz), Hierarchical Data Format (.h5, .h5ad)
	-Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access
	-copyright: public domain, see: https://www.ebi.ac.uk/arrayexpress/help/data_availability.html
	- size: ~10GB
	- purpose: comparative analysis of cell types, gene regulatory networks
	Genome alignment chain files
	- source: UCSC GenomeBrowser
	-- mouse genome based alignments: https://hgdownload.soe.ucsc.edu/goldenPath/mm10/
	-- chicken genome based alignments: https://hgdownload.soe.ucsc.edu/goldenPath/galGal6/
	-- human genome based alignments: https://hgdownload.soe.ucsc.edu/goldenPath/hg38/
	-publication: Lee et al. (2022). The UCSC Genome Browser database: 2022 update. Nucleic acids research, 50(D1), D1115-D1122.
	-copyright: public domain, see: https://genome.ucsc.edu/license/
	- purpose: comparative analysis of gene regulatory networks and enhancers
	- size: ~50G
	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 single-cell and spatial transcriptomics. But also for non-profit organizations and industries active in the field of genetics.

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 Tissue samples for experiments will be obtained from animal cadavers by trained staff in the lab, approved by the Ethical Committee for Animal Experimentation (ECD) : - chicken brain: M001-2021 - mouse brain: P007/2021
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?	No
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

4. 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) or written 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. We 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. 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 (see www.github.com/aertslab for our previous scripts and tools).

| 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. | 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

Metadata will include 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.
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. Details of data processing and analysis will be kept in Jupyter Notebooks or R Notebooks.

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. 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. - 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). Over the next years, archive at VSC will be discontinued. We are using a KU Leuven NAS at Gasthuisberg (synology) combined with cloud storage for archiving. - 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 (https://github.com/aertslab). 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, 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 weekly basis to the archive area be transferred on a weekly basis to the archive area which is mirrored, or (when archive is discontinued) to the gbiomed NAS and cloud storage.

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 controlers with an Ontap 9.1P9 operating system. - the Staging and Archive on VSC are also sufficiently scalable (petabyte scale); on the gbiomed NAS synology, the lab has 460TB storage, it is scalable. Cloud storage is scalable.
What are the expected costs for data storage and backup during the project? How will these costs be covered?	Data storage for the project is covered by the Aerts lab's general costs. In the case that additional storage is required, costs can be covered by the bench fee of this project. Estimated storage costs during the project are 3132€ based on following estimated: -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". Electricity costs for the -80° freezers present in the labs are included in general lab costs. Data storage and backup costs are included in general lab costs.
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. - The VSC storage is only accessible to VSC accounts, and specifically our volume is only accessible to group members - No personal data is processed or generated within in this project

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. All datasets will be stored on the university's central servers with automatic back-up procedures for at least 5 years, conform the KU Leuven RDM policy. The costs (€156 per TB per year for "Large volume-storage") will be covered by by the lab's general storage costs but can be supplemented by the bench fee of this project if necessary .
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". -Tissue samples: Tissues will be stored locally in the laboratory. -Omics data: datasets will be stored on the "L-drive" or, for larger datasets, on the Vlaams Supercomputer Centrum. - Following publication, the results associated with each study will also be deposited in the Dryad repository, where they will be preserved indefinitely.
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 5 years after the end of the project is 5190€. This estimation is based on the following costs: -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 storage and backup costs are included in general lab costs.

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). 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 data repository if requested by the journal and following deposit advice given by the journal.

Where/how will the data be made available for reuse?	In an Open Access repository, Other 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: -Open-access publications in peer-reviewed journals, including supplemental information. 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 data repository if requested by the journal and following deposit advice given by the journal. Depending on the journal, accessibility restrictions may apply. Proper links to datasets will be provided in the corresponding publications. - Patient data: Upon publication, all anonymized patient details supporting a manuscript will be made publicly available as supplemental information. -Omics datasets will be deposited in open access repositories such as the PRIDE Archive for proteomics data, the Panorama platform for targeted proteomics data, the EMBL-EBI platform for genomics and epigenomics data, the LIPID MAPS Lipidomics Gateway for lipidomics data, the Metabolomics Workbench Data Repository for metabolomics data, or the NCBI Gene Expression Omnibus (GEO) or the EBI ArrayExpress databases for functional genomics data. -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 ? EUR) 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. -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. -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 (www.aertslab.org) 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 (e.g. www.github.com/aertslab.). -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?	Upon publication of the research results
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. A budget for publication costs is included in the bench fee for this project.

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 (ELN) and from Raf De Coster for the KU Leuven drives.
Who will be responsible for ensuring data preservation and sharing?	The PI of the lab 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 bears the end responsibility for updating & implementing this DMP?	The PI of the lab is ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.