SpaceTimeOmics: combining high-throughput spatial omics with AI to model gene regulation in space and time

A Data Management Plan created using DMPonline.be

Creators: Lars Borm, 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.

Data Manager: Lars Borm

Grant number / URL: BIO1-G044124N

ID: 204765

Start date: 01-01-2024

End date: 31-12-2027

Project abstract:

Tracking and understanding cellular differentiation during the development of a multicellular organism at single-cell resolution is a long-standing challenge in biology. While single-cell multi-omics, combined with genomic sequence analysis, provides a powerful means to infer gene regulatory networks for a specific tissue, or for a specific time point (i.e., a snapshot), such approaches are limited in throughput and lack the power to reconstruct dynamic cell lineages. To address this, we will develop and apply high-throughput spatial transcriptomics technologies that allow tracking cellular lineages during development. We will implement a new version of the EEL method (Borm et al., Nat Biotech 2022), which provides large-scale, low-cost, and single-cell resolution measurements of gene expression for 500-1000 genes simultaneously. We will apply this approach to generate a comprehensive spatial atlas of Drosophila melanogaster, from the egg to the adult, covering each developmental state, of all cell types in the animal. By integrating this 3D transcriptome atlas with pre-existing single-cell multi-omics data, we will reconstruct spatially-controlled trajectories for each cell type. By exploiting the integrated chromatin accessibility data and the underlying genome sequence, we will develop deep learning models that predict chromatin accessibility and gene expression from the genome sequence, across time and space.

Last modified: 30-05-2024

SpaceTimeOmics: combining high-throughput spatial omics with AI to model gene regulation in space and time DPIA

DPIA

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

Question not answered.

GDPR

Have you registered personal data processing activities for this project?

Question not answered.

SpaceTimeOmics: combining high-throughput spatial omics with AI to model gene regulation in space and time Application DMP

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.

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
Aim 1.4 Image analysis code	Code to analyse the raw imaging data.	reused	Digital	Software	.py	< 100MB	
Aim 2.1 Raw imaging data of multiplexed smFISH	Raw imaing data of multiplexed smFISH. Per sample there will be a image for each decoding cycle and color channel. 132 experiments are planned. Which will have 13 cycles in 3 color channels: 33 files per experiment.	New	Digital	*Experimental*	Original files will be Nikon .nd2 files which will be compressed to ZARR with metadata	>50TB Expected 260TB after compression.	
Aim 2.1 Imaging metadata	Imaging specific metadata linked to the images mentioned above.	New	Digital	Experimental	.yaml and/or incorporated into ZARR	<100MB	
Aim 2.1 Log files	Log files of wetlab experiment	New	Digital	Experimental	.log	<100MB	
Aim 2.1 RNA locations	Extracted locations (XYZ) and gene identity of all detected RNA molecules	New	Digital	Derived and compiled data	.parquet	<5TB	
Aim 2.1 Cell by Gene expression matrix with metadata.	Tabular data with gene expression for each cells. Also contains metadata: Location, cell type, annotation, sample, age.	New	Digital	Derived and compiled data	.h3ad	<5TB	
Aim 2.2 Cellular lineages	Connections between cells as sparse matrix.	New	Digital	Derived and compiled data	Sparse matrix format	<100MB	
Aim 2.3-4 Gene expression maps	Images of gene expression for others to browse on website.	New	Digital	Derived and compiled data	.png, .http	< 100GB	
Aim 3 Data integration	Integration with excisting single cell RNA and ATAC sequencing data (own and public)	reused	Digital	Derived and compiled data			
Aim 3.1 eGRN inference	SCENIC+ inference of gene regulatory networks	New	Digital	Derived and compiled data			
Aim 3.2 Enhancer activity over time	Convolutional Neural Network to identify enhancers driving differentiation in time	New	Digital	Derived and compiled data			
Aim 3.3 DeepSCENIC-ST	Extend DeepSCENIC with time and space.	New	Digital	Derived and compiled data	.py	<100MB	

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

Aim	Dataset/tool	Source
Aim 1	pysmFISH	https://github.com/linnarsson-lab/pysmFISH_auto/tree/master/pysmFISH
Aim 3	Drosophila single cell RNA- and ATAC-seq data.	Published: FCA: DOI 10.1126/science.abk2432 Brain: DOI 10.1038/s41586-021-04262-z Eye antennal disk: DOI 10.15252/msb.20209438 Also unpublished datasets, generated in-house.
Aim 3.1	SCENIC+	DOI 10.1038/s41592-023-01938-4
Aim 3.2-3	DeepSCENIC	Unpublished, generated in house.

External datasets/tools

Aim	Dataset/tool	Source
Aim 3	Tangram	DOI: 10.1038/s41592-021-01264-7
Aim3	Drosophila single cell RNA- and ATAC-seq data.	embryo and larva: DOI 10.1101/2024.02.06.577903 Embryo: https://doi.org/10.1038/nature25981

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.

- No

The project is on Drosophila melanogaster.

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.

- No

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

Aim1
Developing the wet-lab chemistry to do multiplexed smFISH detection in Drosophila may have patentable parts. However, this space is very crowded so this would only be applicable if there is a major advance over existing protocols/products.

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

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.

- No

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).

All experiments will get a unique experiment identification number. This ID number will be linked to all instances of the data. Either in the file name or in the folder name.

The ID has the format of <Initials>EXPYYYYMMDD_<topic>

Like: LBEXP20240417_Mouse_brain_section_3

The data generation is automated by a home-build robot. This machine performs detailed logging of each automated step of the experiment. Furthermore, any manual procedures will be written to the same log file by the user. The log file will also contain all associated metadata of the sample (age, sex, strain, sectioning method etc.). The completed log file will be uploaded to the ELN, and will travel with the raw and processed data. Any downstream processing will be added to the ELN and to log files with the raw/processed data.

The robot also makes an experiment sumary that is easier to read than the log file (but all information is also in the log file). This file has the name: <Experimental ID>_config.yaml and is in the yaml format.

This file contains:

```
Age: #Age of sample
Barcode: #True if barcoded
Barcode_length: #Barcode length
Chamber_EXP: #Flow cells
Chemistry: #Name of experimental method
Codebooks: #Gene decoding codebooks
  Codebook_Atto425: None
  Codebook_Cy3: codebookHG2_20210508.parquet
  Codebook_Cy5: gene_hGBM20201124.parquet
  Codebook_Cy7: None
  Codebook_DAPI: None
  Codebook_Europium: None
  Codebook_FITC: None
  Codebook_TxRed: None
Description: #One sentence description of experiment
EXP_name: #Experiment ID
Experiment_type: #Type of experiment
Heatshock_temperature: #Temperature
Hyb_time_1_A: #Time
Hyb_time_1_B: None
Hyb_time_1_C: None
Hyb_time_2_A: None
Hyb_time_2_B: None
Hyb_time_2_C: None
Hybmix_volume: #Volume
Imaging_temperature: #Temperature
Machine: #Microscope ID
Multicolor_barcode: #True if multicolor barcode
Operator: #Name of operator
Orientation: #Orientation of sample
Overlapping_percentage: #Percentage of overlap
Pipeline: #Image analysis pipeline
Position: #Positionin sample
Probes_FASTA: #Fasta files with probe sequences
  Probes_FASTA_Atto425: None
  Probes_FASTA_Cy3: HG2.fasta
  Probes_FASTA_Cy5: HG.fasta
  Probes_FASTA_Cy7: None
  Probes_FASTA_DAPI: None
  Probes_FASTA_Europium: None
  Probes_FASTA_FITC: None
  Probes_FASTA_TxRed: None
Program: #Program
```

Protocols_io: #Link to protocol
Readout_temperature: #Temperature
RegionImaged: #Name of anatomical region
Sample: #Sample ID
SectionID: #Section ID
Species: #Species
Staining_temperature: #Temperature
Start_date: #Date YYYYMMDD
StitchingChannel: #Channel name
Stitching_type: #Stitching method
Strain: #Sample strain
Stripping_temperature: #Temperature
Target_cycles: #Nubmer of cycles
Tissue: #Tissue type
roi: #field of view numbers.

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: ZARR, TIFF, JPEG 2000, PNG, GIF
• Tabular data: CSV, TXT, H5
• Text: XML, PDF/A, HTML, JSON, TXT, RTF, YAML
• 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 log file that contains all metadata and experimental procedures to generate the data. Also see the information above.
Specific cases:
For final versions of the cell by gene matrix, we will adhere to the metadata standard of https://cellxgene.cziscience.com/.
Unfortunately, there are no metadata standards for the raw data of the field of spatially resolved transcriptomics. However, there are new projects such as SpatialData (https://github.com/scverse/spatialdata) which have the potential to become the standard. We will closely monitor these and incorporate them when applicable and mature.

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

Where will the data be stored?

Digital data

- Primary processing of the images will be on the VIB datacore which has secure storage. When processed, the files will be written to the ManGO platform of KU Leuven.
- The ManGO platform for storage and management of large volumes of active research data will be used for medium-term storage. This platform allows secure storage, manual and automated metadata coupling, data workflows, and file sharing.
- KU Leuven further offers fast ("J-drive) and slower ("L-drive") server storage that allows reading/writing/modification of non-confidential, confidential, and strictly confidential data.
- 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.
- Alternative cold storage will be explored provided by commercial

Note on the large data size of the raw images:

The field of spatial transcriptomics is relatively young and many standards for raw-data storage are still under development. Furthermore, there are currently no public repositories to share this kind of data. Everyone in the field would like this but, the large data size (tens to hundreds of terrabytes) poses a serious chalange for storage and sharing. Even though the storage capacity of the current infrastructure available to us, is sufficient, this project has an exceptionally large demand for storage. Therefore we will look for ways to reduce the storage load through compression, and explore alternative ways of storage of raw data, such as tape storage. Furthermore, there is also a discussion in the field if all the raw data needs to be stored. These methods generate very similar imaging to Illumina DNA sequencing, and Illumina machines do not save the raw images. In fact, the company 10X already discards all the raw data on their Xenium platform, and only saves the derived RNA localizations. Nevertheless, we will aim to store all the raw data because we think this is important for transparency and data sharing. But wil will also monitor the developments in the field and actively contribute to the discussion on raw data storage.

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, VIB datacore or ManGO platform.

Physical samples

- Fly tissue samples: Tissues will be stored locally in the laboratory.
- Fly line stocks are preserved as a minimum of two separate cultures, each maintained at 18°C on a 4-to-5-week generation cycle.

How will the data be backed up?

VIB Datacore drives are backed-up according to the following scheme

- Data stored on the VIB Datacore is stored with internal redundancy and duplicated on an independent server. 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 2 PB large. The data is protected against calamities at either site by synchronizing it in real-time at hardware level.

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

Unprocessed smFISH images are expected to be 4 TB per experiment and will be stored and processed via the VIB Datacore. These images will be directly compressed and they are expected to be 2 TB and are stored on the KU Leuven ManGO platform). Around 130 experiments are planned. Corresponding to 520 TB raw data and around 260 TB after compression.

Both the VIB Datacore and ManGO platform are equipped to handle and store this volume of data during this project.

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 with two-factor authentication. This only provides access to their own data, or data that was shared to them.
- Access to the VIB Datacore servers is possible only through using a VIB user-id and password with two-factor authentication. This only provides access to their own data, or data that was shared to them.

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.
- Large datasets (>100 GB) are stored on the ManGO platform. Total storage requirement for this project on the ManGO platform is expected to be 260 TB, costing an estimated 9,100€/Year.
- Data storage and backup costs are included in general lab costs.

However, if we find a cheaper data storage alternative we might put the raw data there after the derivative files have been generated for the biological analysis. An option would be to put the data on a "cold storage" platform provided by Google or Amazon, that has the lowest cost and is suitable for data that is not often needed. Google's lowest cost storage is 14 €/TB/Year.

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.

This project will generate a very large raw dataset expected to be in the range of 250TB after compression. The largest portion of the dataset are the raw images. To save space these will be converted to the ZARR format and compressed with a lossless compression algorithm so that all raw data is saved.

Furthermore, the metadata, RNA localization and cell by gene tables files will be stored.

Any temporary derivatives of the raw data that are not the final result, will be deleted. For example, to detect objects in the images the background of the images first needs to be filtered out. These filtered images will be deleted to save space. However the stored raw data and the analysis code will ensure that these can be regenerated if needed.

Raw data of failed experiments, if any, will be deleted to save space.

Final machine learning models will be stored.

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.
We will look into cheaper storage alternative such as tape storage or other cold storage alternatives with has an estimated cost of 7 euro/TB/year.

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.

- Other, please specify:
- Yes, in an Open Access repository

For the raw image data there are currently no open repositories that can host this kind and size of data. Therefore, the raw image data will be available upon request and directly transfered.
The derived cell by gene matrices with their metadata will be made available in Open Access Repositories.

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

Not applicable.

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.

- No

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

Datasets for which databases are available will be deposited there:

- The cell by gene matrices will be shared on https://cellxgene.cziscience.com/ and potentially also on SCope.aertslab.org.
- Computational workflows, models, and metadata will be stored on platforms such as Github, Kipoi, and Zenodo with proper versioning.
- There are currently no repositories for raw spatial imaging data that can handle the large volume of raw images. Therefore, the raw images will be available upon request. Potentially, a minimal example dataset will be shared on https://idr.openmicroscopy.org/ or a similar database. If the situation changes in the future, we will upload the data to available databases.

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?

Lars Borm and Stein Aerts are 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?

Lars Borm and Stein Aerts are ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.