1. General Project Information	
Name Grant Holder & ORCID	**Alejandro Sifrim (0000-0001-8247-4020)**
Contributor name(s) (+ ORCID) & roles	**Katy Vandereyken (0000-0002-4477-5866) – Copromotor** **David Wouters (0000-0002-8000-8023) – Junior researcher**
Project number [1] & title	3M220688 - An AI-driven study of mRNA subcellular localization using highly-multiplexed super-resolution in situ transcriptomics
Funder(s) GrantID [2]	G005923N
Affiliation(s)	☑ KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: ROR identifier KU Leuven: 05f950310

[1] "Project number" refers to the institutional project number. This question is optional. Applicants can only provide one project number.

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

Please provide a short project description	The localization of mRNA molecules within a cell plays a crucial role in many fundamental biological processes such as cell migration, polarization, and differentiation. However, this post-transcriptional phenomenon has been understudied due to technological limitations, where only few genes could be assayed. Recently, novel technologies have been proposed for highly-multiplexed, subcellular-resolution in situ assaying of transcripts. Here we propose the application of such cutting-edge technologies on well-described biological models (fruit fly, intestinal enterocyte polarization, axonic and dendritic growth in the brain, human and mouse embryo development) to perform a large-scale study of RNA localization patterns, their molecular actors and functional consequences. To achieve this, we propose the development of novel computational analysis strategies for the automated characterization of spatial expression patterns using deep convolutional autoencoder neural networks. This will allow us to describe known and novel genes which rely on specific localization to perform their function, providing deeper insights into the molecular biology of the studied models.

2. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data [3].

				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
MERFISH_1	D. Melanogaster planar cell polarity (MERFISH)	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☒ Images ☐ Sound ☐ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	TIFF, CSV, binary data, flat text files	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☒ > 5 TB ☐ NA	
MERFISH_2	Mammalian gut enterocyte apical-basal polarization (MERFISH)	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☒ Images ☐ Sound ☐ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	TIFF, CSV, binary data, flat text files	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☒ > 5 TB ☐ NA	
MERFISH_3	Neuron and oligodendrocytes in human induced	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☒ Images ☐ Sound ☐ Numerical	TIFF, CSV, binary data, flat text files	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB	

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

	pluripotent stem cell derived motor neurons and post-mortem human brains (MERFISH)			☐ Textual ☐ Model ☐ Software ☐ Other:		☒ > 5 TB ☐ NA	
MERFISH_4	Human and murine early embryo development (MERFISH)	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☒ Images ☐ Sound ☐ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	TIFF, CSV, binary data, flat text files	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☒ > 5 TB ☐ NA	
scRNAseq_1	D. Melanogaster planar cell polarity (scRNAseq)	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	FASTQ, CSV, binary data, flat text files	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
scRNAseq_2	Mammalian gut enterocyte apical-basal polarization (scRNAseq)	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	FASTQ, CSV, binary data, flat text files	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	

scRNAseq_3	Neuron and oligodendrocytes in human induced pluripotent stem cell derived motor neurons and post-mortem human brains (scRNAseq)	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	FASTQ, CSV, binary data, flat text files	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
scRNAseq_4	Human and murine early embryo development (scRNAseq)	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	FASTQ, CSV, binary data, flat text files	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	

GUIDANCE:
The data description forms the basis of your entire DMP, so make sure it is detailed and complete. It includes digital and physical data and encompasses the whole spectrum ranging from raw data to processed and analysed data including analysis scripts and code. Physical data are all materials that need proper management because they are valuable, difficult to replace and/or ethical issues are associated. Materials that are not considered data in an RDM context include your own manuscripts, theses and presentations; documentation is an integral part of your datasets and should described under documentation/metadata.
RDM Guidance on data

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.	/
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, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.	☒ Yes, human subject data; provide SMEC or EC approval number: G088621N, S65125. ☐ Yes, animal data; provide ECD reference number: ☐ Yes, dual use; provide approval number: ☐ No Additional information:
Will you process personal data[4]? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).	☒ Yes (provide PRET G-number or EC S-number below) ☐ No Additional information: S65125, G088621N
Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.	☐ Yes ☒ No If yes, please comment:
Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements, research collaboration agreements)? If so, please explain to what data they relate and what restrictions are in place.	☐ Yes ☒ No If yes, please explain:

[4] See Glossary Flemish Standard Data Management Plan

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain to what data they relate and which restrictions will be asserted.	☐ Yes ☒ No If yes, please explain:

3. Documentation and Metadata	
Clearly describe what approach will be followed to capture the accompanying information necessary to keep **data understandable and usable**, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded). *RDM guidance on documentation and metadata*.	Experimental metadata (going from sample metadata, to experimental/imaging parameters) will be meticulously deposited and added into the ManGO active research data repository (as well as kept in JSON files in the primary folder structure) and linked to the individual data files, intermediate and final processed files. All data processing will be tracked using Jupyter notebooks for reproducibility of processing/analysis. We will use common best practices to annotate datasets with generally used sample and gene ontologies. Trained statistical/AI models will also be tracked be tracked through the metadata and linked to their underlying training data and their respective Jupyter notebooks.

Will a metadata standard be used to make it easier to **find and reuse the data**? If so, please specify which metadata standard will be used. If not, please specify which metadata will be created to make the data easier to find and reuse. *REPOSITORIES COULD ASK TO DELIVER METADATA IN A CERTAIN FORMAT, WITH SPECIFIED ONTOLOGIES AND VOCABULARIES, I.E. STANDARD LISTS WITH UNIQUE IDENTIFIERS.*	☐ Yes ☒ No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: If no, please specify (where appropriate per dataset or data type) which metadata will be created: There is currently no metadata standard for spatial transcriptomics experiments. We will keep track of experimental metadata (both at the sample level as well as experimentally) using commonly used gene identifiers (ENSEMBL IDs). If a standard emerges during the lifetime of the project we will adhere to that.

4. Data Storage & Back-up during the Research Project	
Where will the data be stored? *Consult the interactive KU Leuven storage guide to find the most suitable storage solution for your data.*	☐ Shared network drive (J-drive) ☐ Personal network drive (I-drive) ☐ OneDrive (KU Leuven) ☐ Sharepoint online ☐ Sharepoint on-premis ☐ Large Volume Storage ☐ Digital Vault ☒ Other: MaNGO + VSC staging/archiving volumes
How will the data be backed up? *WHAT STORAGE AND BACKUP PROCEDURES WILL BE IN PLACE TO PREVENT DATA LOSS?*	☒ Standard back-up provided by KU Leuven ICTS for my storage solution ☐ Personal back-ups I make (specify) ☐ Other (specify)

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.	☒ Yes ☐ No We've consulted with RDM and there is sufficient capacity to store the data for the lifetime of the project + 5 years. If no, please specify:
How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *CLEARLY DESCRIBE THE MEASURES (IN TERMS OF PHYSICAL SECURITY, NETWORK SECURITY, AND SECURITY OF COMPUTER SYSTEMS AND FILES) THAT WILL BE TAKEN TO ENSURE THAT STORED AND TRANSFERRED DATA ARE SAFE.* *Guidance on security for research data*	The data will only be available to authorized personnel throught the ManGO user identification system. On compute servers the data will be stored in volumes with managed user permissions, making it only available for authorized persons.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	We have budgeted for 200TB of data storage at 30 Euro/TB/year for the lifetime of the project (as part of the FWO funding). Long-term post-project storage will be covered by complementary project funding (VLIR RELANCE infrastructure grant).

5. Data Preservation after the end of the Research Project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...). *Guidance on data preservation*	☒ All data will be preserved for 10 years according to KU Leuven RDM policy ☐ All data will be preserved for 25 years according to CTC recommendations for clinical trials with medicinal products for human use and for clinical experiments on humans ☐ Certain data cannot be kept for 10 years (explain)
Where will these data be archived (stored and curated for the long-term)? *Dedicated data repositories are often the best place to preserve your data. Data not suitable for preservation in a repository can be stored using a KU Leuven storage solution, consult the interactive KU Leuven storage guide.*	☒ KU Leuven RDR (for publication data) ☐ Large Volume Storage (longterm for large volumes) ☐ Shared network drive (J-drive) ☒ Other (specifiy): We have budgeted long-term large-scale cold archiving storage as part of a VLIR RELANCE infrastructure grant. We're currently working out potential hardware solutions with KUL RDM ICTS. Sequencing data will be deposited to EGA/GEO data repositories.
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	We have budgeted long-term large-scale cold archiving storage (10 Euro/TB/year) as part of a VLIR RELANCE infrastructure grant.

6. Data Sharing and Reuse

Will the data (or part of the data) be made available for reuse after/during the project? Please explain per dataset or data type which data will be made available. *NOTE THAT 'AVAILABLE' DOES NOT NECESSARILY MEAN THAT THE DATA SET BECOMES OPENLY AVAILABLE, CONDITIONS FOR ACCESS AND USE MAY APPLY. AVAILABILITY IN THIS QUESTION THUS ENTAILS BOTH OPEN & RESTRICTED ACCESS. FOR MORE INFORMATION: HTTPS://WIKI.SURFNET.NL/DISPLAY/STANDARDS/INFO-EU-REPO/#INFOEUREPO-ACCESSRIGHTS*	☒ Yes, as open data ☐ Yes, as embargoed data (temporary restriction) ☒ Yes, as restricted data (upon approval, or institutional access only) ☐ No (closed access) ☐ Other, please specify: Non-human data will be made openly available. Human data will be deposited in either public data repositories under a controlled data access policy with a data access committee evaluating data access requests.
If access is restricted, please specify who will be able to access the data and under what conditions.	Access to human data will be granted by the data access committee to bonafide researchers affiliated with recognized research institutions.
Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain per dataset or data type where appropriate.	☒ Yes, privacy aspects ☐ Yes, intellectual property rights ☒ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ☐ No If yes, please specify: Data access to human data will restricted according to the specified clauses in the informed consent forms for the different studies. Specifically sequencing data will be mostly restricted given the identifiability of the subjects. Microscopy imaging data will be made publicly available.
Where will the data be made available? If already known, please provide a repository per dataset or data type.	☒ KU Leuven RDR ☒ Other data repository (specify) EGA/GEO (scRNAseq) ☐ Other (specify)

When will the data be made available?	☒ Upon publication of research results ☐ Specific date (specify) ☐ Other (specify)
Which data usage licenses are you going to provide? If none, please explain why. *A DATA USAGE LICENSE INDICATES WHETHER THE DATA CAN BE REUSED OR NOT AND UNDER WHAT CONDITIONS. IF NO LICENCE IS GRANTED, THE DATA ARE IN A GREY ZONE AND CANNOT BE LEGALLY REUSED. DO NOTE THAT YOU MAY ONLY RELEASE DATA UNDER A LICENCE CHOSEN BY YOURSELF IF IT DOES NOT ALREADY FALL UNDER ANOTHER LICENCE THAT MIGHT PROHIBIT THAT.* *Check the RDR guidance on licences for data and software sources code or consult the License selector tool to help you choose.*	☐ CC-BY 4.0 (data) ☒ Data Transfer Agreement (restricted data) ☐ MIT licence (code) ☒ GNU GPL-3.0 (code) ☐ Other (specify)
Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *INDICATE WHETHER YOU INTEND TO ADD A PERSISTENT AND UNIQUE IDENTIFIER IN ORDER TO IDENTIFY AND RETRIEVE THE DATA.*	☒ Yes, a PID will be added upon deposit in a data repository ☐ My dataset already has a PID ☐ No
What are the expected costs for data sharing? How will these costs be covered?	We don't expect additional costs for data sharing

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	David Wouters

Who will manage data storage and backup during the research project?	David Wouters
Who will manage data preservation and sharing?	Alejandro Sifrim
Who will update and implement this DMP?	Alejandro Sifrim