DMP - FWO - Fonds Wetenschappelijk Onderzoek - G0I2722N (EOS ID: 40007513)

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	Cédric Blanpain - cedric.blanpain@ulb.ac.be
FWO Project Number & Title	G0I2722N (EOS ID: 40007513) Cell-cell cOmmuNicaTion As a driver of Cancer cell state identiTy – Decoding the impact of cell-cell communications on the identity of tumor states in skin cancers
Affiliation	VIB-KU Leuven Center for Brain & Disease Research

Responsible: Stein Aerts, Jean-Christophe Marine

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 Fresh/frozen/fixed tissue samples from research animals (mice) Experimental data Digital images Microscopy pictures, gel scans, graphs, illustrations, figures. Digital images in vector formats: scalable vector graphics (.svg), encapsulated postscript (.eps), Scalable Vector Graphics (.svg), Adobe Illustrator (.ai) (1-15Mb).` Acquired images & processed images of tumor sections after (m)IHC, (m)IF and (m)FISH staining experiments will be saved in different formats depending on the microscope that will be used for the image acquisition and software processing. (.tiff, .jpeg, .czi. or .nd2-files). The size of the files ranges from 1GB-10GB for single plane images of from 10GB-50GB for multiple plane images. Cytometry data Flow Cytometry and fluorescence-activated cell sorting (FACS) data. Format: Flow Cytometry Standard (.fcs); (10-50 MB) The data will be processed using FlowJo software and exported as .tiff image files or .xlsx-tables (<5MB). FACS-sorted cells will be collected to extract RNA or proteins. Cellsorting and chip-derived data Cell sorting experiments on chip can be documented based on raw data files or pre-processed data. Raw data files contain the full gaussian profiles of the detected cells. Preprocessed data record the experimental settings and area and height of every detected

	peak similar to benchmark cytometers. Format: .tdms for raw (100-300MB) and pre-processed (10-50 MB). Pre-processed can also be converted to .fcs for further processing in standard flow cytometry software In house developped matlab scripts are used to further analyze raw data traces if needed, resulting in matlab figures (.fig 10-50 MB) and datasets (.mat 100-800MB) Omics data Single-cell RNA-seq, single-cell ATAC-seq, single-cell multi-omics, and spatial comics data. Formats: - Next generation sequencing raw data: binary base call format (.bcl), .fastq(.gz) Average size of these files ranges from 500MB per cell for single cell sequencing up to 3GB for bulk sequencing and 20GB for 10X sequencing. - Read/UMI count data: .tsv(.gz), Matrix Market format (.mtx), .loom, .rds(.gz), h5ad - Sequence alignment data: (.sam), .bam (10-20 GB), .cram - Coverage data: .bed, .bg, .bedGraph, .bw, .bigwig - Structural variations data: .vcf(.gz), .bcf - Nucleic acid samples resulting from (single-cell) nucleic acid amplification, or sequence library preparations will be stored in labeled tubes or SBS plates in -20C freezers. We have electronic laboratory databases that will keep the physical storage

	address of these samples. Genetically modified organisms and PDX lines All animals will be registered in the Leuven Animal Information System (LAIS) database or ticket@lab, along with corresponding genotyping information, ethical approval documents and animal provider receipts. Treatment and follow-up of the animals tumor development and general health status will be detailed in an excel-file (<5MB). Simulation data In silico analysis of data enables to simulate real observational data, and will be used to determine specificity and sensitivity of developed algorithms. Data format: Quantitative tabular data Derived and compiled data Research documentation 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/clinical 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. Manuscripts The results will be published as BioRxiv preprints and articles in peer reviewed journals. Algorithms and scripts

	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). Software Software will be made available via GitHub Canonical data Nucleic acid sequences - Nucleotide sequences: raw sequence data trace (.ab1), text-based format (.fasta/.fa) and accompanying QUAL file (.qual), Genbank format (.gb/.gbk); - FASTQ formatted files of the omics experiments will be made available via NCBI/Gene Expression Omnibus 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 genomics 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 Refer to ethical committee approval: - for work on human biopsy data, cell lines and iPS cells: S63799, S62275, S63316, S56777 - for work with laboratory animals: Ethical Committee Animal Experimentation ECD 097/2020, 125/2021 and 147/2021 -Human cell lines: either commercial cell lines or cell lines from our collaborator Ghanem Ghanam at Bordet Institute, under MTA. -Genetically modified organisms: animals are housed 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. Animal administrative, husbandry and animal welfare data are sensitive data and are stored in the LAIS database or ticket@lab according to security procedure of KU Leuven.

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?	Yes We do not exclude 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 , VIB and imec in accordance with the framework agreement of the institutes. VIB has a policy to actively monitor research data for such potential. If there is substantialpotential, 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. Inthe case a decision is taken to file a patent application it will be planned so that publications need not be delayed. The use of mouse models will be subjected to the terms described in their respective MTAs. Specific examples: -Potential biomarkers for tumor cell states will be claimed if this opportunity arises. -New bioinformatics software can be licensed - new fluidic chip concepts will be claimed
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?	The data will be generated following standardized protocols. Clear and detailed descriptions of these protocols will be stored in our lab protocol database, and published along with the results. . 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. Cryotubes of biological samples (bacterial and yeast strains) stored at -80°C will be labelled with a reference number that links to an entry in or strain database. All datasets will be accompanied by a README.txt, .docx or .xlsx file containing all the associatedmetadata (see more details below). The data will be generated following standardized protocols. Clear and detailed descriptions of these protocols will be stored in our lab protocol database, and published along with the results.

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, including the following elements: • Title: free text • Creator: Last name, first name, organization • Date and time reference • Subject: Choice of keywords and classifications • Description: Text explaining the content of the data set and other contextual information needed for the correct interpretation of the data, the software(s) (including version number) used to produce and to read the data, the purpose of the experiment, etc. • Format: Details of the file format, • Resource Type: data set, image, audio, etc. • Identifier: DOI (when applicable) • Access rights: closed access, embargoed access, restricted access, open access. When depositing data in a local or public repository, the final dataset will be accompanied by this information under the form of a README.txt, .docx or .xlsx .document. This file will be located in the top level directory of the dataset and will also list the contents of the other files and outline the file-naming convention used (see section 7 below). This will allow the data to be understood by other members of the laboratory and addcontextual value to the dataset for future reuse

	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. Give details as needed for the project. Specific examples (adjust as required): - SOPs for biological data generation are kept on a dedicated KU Leuven shared drive. A central excel file is stored on that same drive, detailing for examples: (1) sample ID; (2) SOP with which data generation was performed; (3) abnormalities or deviations from SOP in data generation; (4) experimental QC values (e.g. DNA concentrations); (5) location of the source sample in the freezer. - For bioinformatics processing, a data analysis log will be kept that details: (1) sequencing run ID; (2) the bioinformatics SOPs/scripts that were applied; (3) location of source files; (4) abnormalities or deviations. All original data generated on chip will be kept on a dedicated Imec shared drive. A central onenote file is stored that notes the details for every experiment (1) sample (2) SOPs used (3) location of raw and processed datafiles 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. For following type of experiments, the below international metadata standards will be used: qPCR: MIQE compliant RDML files will be exported from qBase+ (https://www.ncbi.nlm.nih.gov/pubmed/19246619) FACS & Flow Cytometry: MiFlowCyt compliant metadata files will be exported from Fortessa & Sony and ARIA Fusion (https://www.ncbi.nlm.nih.gov/pubmed/18752282) Image metadata will be exported as OME-XML files using Fiji and QuPath software

5. Data storage & backup during the FWO project

Where will the data be stored?	Digital files will be stored on KU Leuven servers, except for private data that will be stored on KU Leuven secure server (digital vault). - Tissue samples: Tissues will be stored locally in the laboratory. Cryoblocks of tumor tissue will be stored at -20°C, paraffin blocks of tumor tissue will be stored at room temperature. All human tissue samples will be registered with a Belgian biobank, in compliance with the Belgian lawon human body material (dd 19-12-2008). - 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 area and later in the archive area or on a gbiomed NAS storage. - Vectors: As a general rule at least two independently obtained clones will be preserved for each vector, both under the form of purified DNA (in -20°C freezer) and as a bacteria glycerol stock (-80°C). All published vectors and the associated sequences will be sent to the non-profit plasmid repository Addgene, which will take care of vector storage and shipping upon request. - 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 and yeast strains will be stored in a -80°C freezer in the labs. Costs are covered by general lab expenses. - 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 or ticket@lab, along with corresponding genotyping information, ethical approval documents and animal provider receipts. Drosophila lines will be stored in a dedicated room and managed using a specific database for storage of the corresponding information (including genotype, origin, number of vials and date of transfer, crossing schemes) and vial tracking via unique QR codes. Other biological and chemical samples: storage at 4°C and/or as frozen samples in cryovials as appropriate.

	- Algorithms, scripts and softwares: All the relevant algorithms, scripts and software

	code driving the project will be stored in a private online git repository from the GitHub account of the department (https://github.com/vibcbd). - 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. All chip derived data is stored in the project account on the internal imec server Isilon, an on-prem file storage system that offers internal file shares and back-up. In the very rare case that storage space is insufficient, a cloud-based, easily scalable, storage system will be used. Archiving happens by compiling the data on different tier levels of the Isilon file servers. This is initiated by the archiving Storage Team at imec, which searches for the best archiving solution where the financial and sustainability parameters are taken into account.

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. - 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. - All omics data stored on the Flemish Supercomputer Centre (VSC) will be transferred on a weekly basis to the archive area. Incremental backups are done daily from one 20 TB QNAP NAS to a second 20 TB QNAP NAS. Since archive will be discontinued, we are backing up data from VSC on a NAS volume, and by the time archive will be out of operation, we will either have a mirror of our NAS volume, or a backup in cloud storage. The back-up of imec's Isilon file servers is ensured through snapshot technology. The back-up of the cloud-based storage system is managed by Microsoft Azure Backup service.

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 sequencing data is maintained on VSC storage and backed up on a gbiomed-hosted NAS volume with sufficient RAID and hot spare configuration. Imec’s Isilon file servers has been designed to allow sufficient storage space. In very rare case that storage space is insufficient, a cloud-based, easily scalable, storage system will be used (no extra costs).

What are the expected costs for data storage and backup during the project? How will these costs be covered?	The total estimated cost of data storage during the project is 20,000 euro. 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". - The costs for digital chip data at imec is estimated around 50€/TB per month for both on-premise as cloud storage and back-up, and is considered as general lab costs covered by imec. - the Staging on VSC has been renewed in 2021 with 3 Pb of space; and the gbiomed NAS is scaleable and will start with 150 Tb space. Cloud storage is scalable. -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 specimens 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. 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. 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. All private data will be rendered anonymous before processing outside the digital vault. Only the PI will be granted access to the server to deposit private data. The PI will be the only responsible for linking patient information, survey data and/or tissue samples, and will strictly respect confidentiality. All de-identified data will be exported from the database by the PI, and stored on KU Leuven servers from where it can be accessed by the research and technical staff from the laboratory. Data on VSC is only accessible by VSC username, password, and MFA, and permissions on data folders are set for specific groups in case of GDPR. For security, the imec environment requires that all logins to imec environments are protected by a strong password (minimal 14 characters) combined with multi factor user-authentication. The internal imec information security service monitors continuously for potential security breaches.

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 the Aerts and Marine labs. Datasets collected in the context of clinical research, which fall under the scope of the Belgian Law of 7 May 2004, will be archived for 25 years, in agreement with UZ Leuven policy and the European Regulation 536/2014 on clinical trials of medicinal products for human use. At imec, all chip data will be stored on the same Isilon file servers for 5 years.

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, the gbiomed NAS volumes, and cloud storage. -Vectors: As a general rule at least two independently obtained clones will be preserved for each vector, both under the form of purified DNA (in -20°C freezer) and as a bacteria glycerol stock (-80°C). -Cell lines: human cell lines will be stored in the UZ Leuven Biobank (-80°C). Human pluripotent stem cell lines generated during this project will be deposited in hPSCreg. Animal cell lines will be stored in liquid nitrogen cryostorage of the laboratory. -Genetically modified organisms: Drosophila lines will be housed locally. All other lines that are no actively used for experiments will be cryopreserved. -Other biological and chemical samples: storage at 4°C and/or as frozen samples in cryovials as appropriate. - Following publication, the results associated with each study will also be deposited in the Dryad repository, where they will be preserved indefinitely. For the chip data, at imec, after the standard retention period of 5 years, data will be subjected to evaluation by the archiving Storage Team. This will include weighing the potential value versus the costs of keeping it available versus the sustainability of the archiving solutions. Decisions will be made by the data owners, in close collaboration with the Storage team.
What are the expected costs for data preservation during these 5 years? How will the costs be covered?	It is the intention to minimize data management costs by implementing standard procedures (internally generated operational procedures (https://en.wikipedia.org/wiki/Standard_operating_procedure)) e.g. for metadata collection and file storage and organization from the start of the project, and by using free-to-use data repositories and dissemination facilities whenever possible. Data management costs will be covered by the laboratory budget.

7. 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 3 rd 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. Depending on the journal, accessibility restrictions may apply. Biological material will be distributed to other parties if requested
Where/how will the data be made available for reuse?	In an Open Access repository, In a restricted access repository
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 has been requested in 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. For imec, the project manager ensures correct data storage and back up.
Who will be responsible for ensuring 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. For the chip data at imec, the departmental managers will be responsible to ensure data preservation and sharing after project closure..
Who bears the end responsibility for updating & implementing this DMP?	The PIs are ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.