Plan Name / My plan (IDN DMP)

ID: COLUMBO: Engineered CRISPR toolbox for revolutionizing single-cell genome imaging in embryogenesis and oncology

Grant number*:* IDN/21/006

PI/Researcher: Jeroen Lammertyn, Dragana Spasic, Thierry Voet, Vitor Pinheiro, Alejandro Sifrim

PI/Researcher ID:

Plan data contact: Dragana Spasic

Description: Growing evidence has revealed that numerous cellular processes in healthy and diseased cells are orchestrated by genomic organization and dynamics. However, our understanding of this is still inadequate, mainly due to the limitations of the available technologies. We aim to revolutionize the field of genomic DNA imaging by bringing together our patent-pending reengineered CRISPR (d)Cas9 with several other cutting-edge DNA nanotechnologies into a single functional complex, delivering long-awaited specificity, sensitivity, multiplexing and combinations thereof, that are particularly challenging in living cells. The ambition of high-throughput imaging of both the genome and transcriptome level will be supported through development of novel bioinformatics tools next to automated microfluidics and imaging workflows. This high risk/high gain project with overarching impact on the fields of cancer, embryogenesis, aging, neurogenesis among many others will be supported by truly interdisciplinary team of 4 groups from 3 different departments.

1. General information

1.1. Name of the project lead (PI)

Jeroen Lammertyn

1.2. C1-C2 Project number & title

IDN/21/006 COLUMBO: Engineered CRISPR toolbox for revolutionizing single-cell genome imaging in embryogenesis and oncology

2. Data description

2.1 Will you generate/collect new data and/or make use of existing data?

Generate new data

2.2 What data will you collect, generate or reuse? Describe the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a numbered list or table and per objective of the project.

WP1. Development of CRISPR-dCas9 toolbox for genome imaging in fixed cells

- DNA/RNA design (.xcel), approximately 13 kb/file (4 years estimate = **520 KB**)
- Spectrophotometer data (raw data and analysis, .xcel), approximately 245kb/experiment (4 years estimate = **49 MB**)
- Cas9 nuclease assay in droplets (images as .tif format and data analysis in .xcel), approximately 2Mb/experiment (4 years estimate = **410 MB)**
- Gel electrophoresis (raw data in .scn format and .tif format, data analysis in .xcel format), approximately 4Mb/experiment (4 years estimate = **800 MB**)

WP2. Bioinformatics and microfluidic tools for high-throughput cell imaging

- Design of microfluidic chips (CAD files, e.g. .svg), 4 years estimate = **2 GB**

- Microfluidic chips operations (video recordings with camera/webcam (e.g., .mp4, .avi), 4 years estimate = **500 GB**
- High-throughput and high resolution microscopy images (OMERO .tiff format, microscope proprietary formats + instrumental metadata), 4 years estimate = **150 TB (generated in WP4 and WP5)**
- Multi-resolution processed and/or summarized microscopy data (*.tiff, delimited text files), 4 years estimate = **5 TB (generated in WP4 and WP5)**
- Single-cell multi-omics sequencing data and alignments (.fastq files, .cram files), 4 years estimate = **1 TB**
- Processed single-cell multiomics data (R data objects), 4 years estimate = **200 GB**
- Analysis code & software packages (flat text, binaries), 4 years estimate = **1-10 GB**
- Containerized analysis workflows (docker containers), 4 years estimate = **1 TB**
- Algorithmically generated in-house gRNA sequence designs (.fasta format), 4 years estimate = **10 GB**
- Trained machine learning models (Keras/PyTorch H5 format), 4 years estimate = **100-200 GB**

WP3. CRISPR-dCas9$_{XNA}$ development for highly-specific genome imaging in live cells

- Nucleic acid designs (.dna, .txt or .fas), 4 year estimate = **10 MB**
- Files detailing oligo purity and identity for in-house synthesized oligos (.pdf), 4 year estimate = **100 MB**
- Files detailing oligo information from commercially synthesized oligos (.pdf), 4-year estimate = **100 MB**
- Images from molecular biology and biochemical assays based on high-resolution gel electrophoresis (.gel and .tiff), 4 year estimate = **2 GB**
- Biochemical assays based on spectroscopical analysis (.xlsx, .pzfx), 4 year estimate = **2 GB**

WP4. Imaging of unique genomic DNA loci and RNA-transcripts in fixed cells

- High-resolution microscopy images of fixed cells (Z-stacking image series in .tif or .nd2 format and data analysis in .xcel), 4 years estimate = **100 TB**
- Multi-resolution processed and/or summarized microscopy data (*.tiff, delimited text files), 4 years estimate = **3 TB**
- Image analysis scripts (GA3 format), approximately 300kb/script (4 years estimate = **5 MB**)

WP5. Live-cell imaging of chromosome instability using the CRISPR-dCas9XNA complex

- High-resolution microscopy images of live cells (Z-stacking and/or time lapse image series in .tif or .nd2 format and data analysis in .xcel), 4 years estimate = **50 TB**
- Multi-resolution processed and/or summarized microscopy data (*.tiff, delimited text files), 4 years estimate = **2 TB**
- Image analysis scripts (GA3 format), approximately 300kb/script (4 years estimate = **5 MB**)

WP6. Project coordination and management

- Observational data: written down in electronic lab notebook (eLABJournal, Bio-ITech)
- Consortium meetings: update presentations, meeting reports (e.g. .pptx, .docx). Expected volumes = **5 GB**
- Scientific publications and doctoral dissertations (e.g. .docx and .pdf). Expected volumes = **10 GB**

3. Ethical and legal issues

3.1 **Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.**

Yes. We will reuse sequencing data previously generated and will also generate new sequencing and/or spatial data from commercially available human cell lines and human embryos. In relation to the embryo analyses, we will potentially also use patient information (parents) retrospectively. This includes personal data, clinical parameters and (epi)genetic information. Personal data includes: age, gender, clinical history at the Leuven University Fertility Center (LUFC), received treatments at LUFC. This data is pseudonymized. The key to personal/patient information of pseudonymized data will always be kept with UZ Leuven (LUFC) on their secured server. For personal and sensitive data, we will abide by the Belgian law on the protection of individuals with regard to the processing of personal data (30th July 2018) and the General Data Protection Regulation 2016/679. Privacy registrations via the KU Leuven PRET tool has been applied for and approved.

The reference of the file in the KU Leuven privacy register regarding the collection and use of human cell lines for method development, including spatial (multi-)omics methods, is: G-2020- 2313-R2(AMD)

The reference of the file in the KU Leuven privacy register regarding multi-omics and spatial profiling of human pre- and post-implantation embryos is: G-2021-3372

3.2 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. We have already acquired The Ethics Committee Research UZ / KU Leuven (EC Research: S64403) approval for technology development on human cell lines. Remaining approvals in this project are for: (1) year 2 - fetal, adult, and ageing healthy tissues as well as diseased tissues, e.g. cancers; (2) year 3 - ECD approval for work with animal (e.g. mouse) embryos and (3) year 4 - work on the human embryos requiring local EC approval as well as Federal EC approval. We are in the process of acquiring these additional ethical approvals.

3.3 Does your research 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 expect new IP generation in this project, more specifically to expand our IP portfolio around patent-pending CRISPR-dCas9$_{\text{DNAzyme}}$ tool and its applications for genomic imaging in fixed and live cells. Additionally, we can expect that the proposed work will result in research data with the potential for tech transfer and valorization, especially with regards to newly developed spatial (multi-)omics protocols and data analysis pipelines, as well as the potential discovery of molecular causes of chromosome instability in the human embryo. When there is concrete potential for tech transfer, the IP related to these research data will be protected, with the support of KU Leuven LRD and the IOF managers supporting this project, i.e. IOF manager linked to the Biosensors group (Dr. Francesco Dal Dosso) and the Voet group (Genomics Medicine, Dr. Gregory Maes). The IP protection will not withhold the research data from being made public. In the case a decision is taken to file a patent application, it will be planned as such not to delay publications longer than needed.

3.4 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 regarding reuse and sharing are in place?

There are no 3rd party agreements in place regarding this project.

4. Documentation and metadata

4.1 What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?

- Protocols, the research progress and clear descriptions of obtained data, what they represent and how they were generated, will be collected **in the Biosensors group** in an electronic notebook (eLABJournal, Bio-ITech). Here, folders will be provided for all subtasks of the project. In each folder, a new file will be made for each experiment, named with the date and subject, and including information about the responsible person (i.e., the person who created the file) as well as version tracking. Each experimental file will contain a section on the objective, protocol, results (a description of results and observations rather than all raw and analyzed data) and conclusions. For each experiment, all raw and analyzed data files will be stored in a folder on the shared server, using the same hierarchical folder structure as the electronic lab notebook. By using the same structure on the server and in the electronic lab notebook, contextual information on the experimentally obtained data can be easily searched and used by a secondary analyst via the electronic notebook.
- **For Voet lab:** Each experiment is registered in the (E-) lab journal of the scientist performing the respective experiment. Standard operating procedures (SOPs) have been and will be written for all the techniques used in the lab. Notes will describe the biological/clinical samples used, experimental setup and protocols used, data generated, links to the specific location of sample derivatives and data, as well as the names of the datasets. Data obtained from experiments will be stored in specific folders that also contains a README.txt file explaining the design/protocol, analysis methods, results and labels used in the data analysis file, and a reference to the (E-) lab journal of that particular experiment. The information provided will allow another researcher to follow all steps in the data processing. Also, algorithms, scripts and software usage will be documented. When scripts, algorithms and software tools are finalized, they will be additionally described in manuscripts and/or online git repositories (e.g. GitHub).
- **For Pinheiro lab**: Protocols and experimental data are registered in the cloud-based e-lab notebook used by the group (Benchling). Data is organised in folders by objective with notebooks detailing experimental parameters (description, design, linked resources) and

results (input directly or as attachments to the notebook). All raw data (< 10 Mb) is stored on university cloud-based drives with larger datasets maintained in data repositories (e.g. Github) accessible only to members of the Pinheiro group. Periodic backups (yearly) to PDF are carried out and stored on cloud-based university servers (OneDrive).

- Administrative items (e.g. project proposal, project reports, update presentations, contracts) will be stored on the shared folder created on the shared drive (J:\SET-MEBIOS-BIOSENSORSSHAREDPROJ-DI0433\KUL-0005\IDN\IDN COLUMBO). This folder is open to all the consortium members and is secured and backed-up by the ICTS service of KU Leuven.

4.2 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.

- Being a highly interdisciplinary project, it is not possible to use a standard metadata system. The Biosensors group will use the electronic lab notebook in which a number of predetermined topics have to be described for each experiment (objective, protocol, results, and conclusion). The electronic lab notebook facilitates searching for particular metadata through a search engine. By mimicking the folder structure of the electronic lab notebook in the server-based folder with the experimental data, linking of the metadata to the actual data will be facilitated. Most of these things also apply for the Voet and Pinheiro lab, although no electronic lab notebooks are used. In the Voet lab, metadata sheet will be maintained with the connection between lab samples and files on our data storage so that data files, lab samples, and experimental notes (including descriptions of equipment, setting and used experimental settings) remain properly linked. In Pinheiro lab, a cloud-based lab book (Benchling) is used.
- Metadata with the connection between lab samples and files on our data storage so that data files, lab samples, and experimental notes (including descriptions of equipment, setting and used experimental settings) will remain properly linked via the iRODs management system.
- As a general rule, datasets will be made openly accessible, whenever possible via existing platforms that support FAIR data sharing (www.fairsharing.org). Examples of current public repositories for (epi)genome and transcriptome sequencing data include the European Genome-phenome Archive (EGA), the database of Genotypes and Phenotypes (dbGAP) or the Gene Expression Omnibus (GEO) database, each requiring their own type of metadata. For spatial (multi-)omics data, currently no mature public repositories exist, but if they emerge in the future, we will make use of these and their metadata standards, e.g. OMERO database for imaging data and SpatialDB database for spatial omics data.
- When depositing data in a local or public repository, the final dataset will be accompanied by this information in a README.txt document, following the Dublin Core Metadata standard if no other meta-standard is available yet. 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 each peer-reviewed article, a separate folder will be made on the server, containing the latest Word version and all raw and processed data used in the article. In addition, a separate file will be made in the electronic lab notebook for each article, containing clickable links to all metadata files of data that were used in that particular article, to facilitate tracing back of protocols, results and conclusions.

5. Data storage and backup during C1-C2 project

5.1 Where will the data be stored?

- The time-stamped digital data will be stored in an already created project folder on the shared drive (J:) of KU Leuven. The time-stamped digital metadata of the **Biosensors group** will be stored on the server of the electronic labbook (eLABJournal, Bio-ITech). The folder is open for the members participating in this KU Leuven ID-N project and is secured and backed-up by the ICTS service of KU Leuven. Copies can be made and kept on personal devices. In the **Pinheiro group** raw experimental data are stored in the local machines with back up copy in cloud (KUL OneDrive, shared folder, Github or Benchling – depending on intended frequency of access and file size). For this project, we are storing non-sensitive data in Benchling but we will move data to the ICTS shared drives for IP sensitive material. In the **Voet group** digital files will be stored temporarily on local computers and redundant hard drives connected to microscopy systems and/or on longer term on KU Leuven servers with back-up capacities (KU Leuven large volume storage (L-drive), KU Leuven J-drive, KU Leuven OneDrive, the VSC (Flemish Supercomputer Centre) and UZ Leuven server: sample logs, experimental protocols, library QC reports & log files → KU Leuven OneDrive; automation protocols & log files → Local computer, KU Leuven OneDrive and VSC storage; raw sequencing data (.bcl) → stored by Genomics Core on professional Google cloud storage and/or protected UZ Leuven server; sequencing (.fastq) and analysis data → VSC staging and archive storage area; raw and processed imaging data → Temporarily on local computer/device and redundant additional external hard drives connected to the microscopy systems, with gradual transfer to the KUL large volume storage and/or VSC staging and archive storage area; analysis pipelines → KU Leuven OneDrive, VSC storage and online git repositories (e.g. GitHub). In the **Sifrim group,** for the duration of the project the data will be stored on an ICTS-managed, access-controlled, redundant storage system with an iRODS data management system. During active analysis the data will be transferred to large-scale scratch data volumes at the VSC and analysis results will be backed up and version-controlled in the iRODs system. After the project, the data will be archived in glacial data storage to reduce storage costs (currently being investigated in collaboration with ICTS).
- Large-scale imaging and sequencing data will be stored in a KU Leuven ICSTS-managed error-redundant iRODS data management system for long-term archival. Experimental meta-data (e.g. sample, instrument, experimental conditions, instrument logs) will be linked directly with the dataset through this system for organized and programmatic retrieval and user-access control. Processed data and trained machine learning models will be version controlled and linked with analysis code available on a KUL-hosted, version-controlled git repository. Linking version-controlled and documented python notebooks and data management systems through an API will allow for easy reproduction and reuse of data.
- Upon publication, all data and protocols supporting the manuscript(s) will be made available via public repositories (see below). The decoding key to pseudonymised personal data is always kept at UZ Leuven on their secure servers.
- Physical samples will be stored in fridges and freezers located in the respective laboratories or liquid nitrogen tanks managed at the departmental level and/or the UZ/KU Leuven biobank. All samples will be tracked using electronic lab notebook (Biosensors group), a LIMS database system or alike, for the purpose of traceability (Voet group) or registry in Benchling (Pinheiro group).

5.2 How will the data be backed up?

All data belonging to the project is stored on snapshotted, failure-proof systems managed by KUL ICTS.

5.3 Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

Yes. KU Leuven ICTS and VSC provide sufficient storage and archival capacity during and after the project. Cost-effective alternatives are currently being worked-out in active discussions with ICTS/VSC.

5.4 What are the expected costs for data storage and backup during the project? How will these costs be covered?

We have estimated a ramp up to 200 TB storage over the duration of the project; at current ICTS/VSC storage pricing of 20 Euro/(TB*year) we estimate storage costs at 14.000 Euro. This amount has been budgeted as part of the project. If additional storage would be required, partners Sifrim and Voet have acquired a large-scale VLIR infrastructure grant to procure high-volume data storage as part of the Leuven Institute of Single-Cell Omics (LISCO) which could be used for this project.

5.5 Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The data is secured by the ICTS service of KU Leuven and secure user access is managed through the iRODs system. Confidential data can and will be protected with a password (available only for PI Jeroen Lammertyn). Visitors, MSc thesis students and internship students in the groups as well as other unauthorized persons will not have access to the data on the shared folder.

The VSC storage is only accessible to VSC accounts, and specifically, our volume is only accessible to group members. No personal data will be stored on the VSC or other servers, except for the pseudonymized nucleic acid/protein sequences. The key to personal/patient information of pseudonymized data will always be kept with UZ Leuven (LUFC) on their secured server.

6. Data preservation after the end of the C1-C2 project

6.1 Which data will be retained for the expected 10 year period after the end of the project? If only a selection of the data can/will be preserved, clearly state why this is the case (legal or contractual restrictions, physical preservation issues, ...).

The data to be retained during 10 years after the project's end are dissemination data (source files of publications and presentations) and the most relevant measurement data. A possible exception to this could become the raw imaging data files from the high-resolution spatial analyses, but only if it will be judged that keeping the reconstructed/processed imaging data will be sufficient and the cost of storing the large amount of raw imaging data will be unaffordable.

6.2 Where will these data be archived (= stored for the long term)?

The research data will be stored in a glacial archive storage system after the end of the project. Dissemination data, namely files corresponding to papers and presentations, will be stored on the PCs of involved PIs, and backed-up daily on the departmental server for long term storage. Analysis code will be stored on ICTS-hosted code repositories.

6.3 What are the expected costs for data preservation during these 10 years? How will the costs be covered?

- The volume corresponding to dissemination data is expected to be relatively low (<10 GB), and therefore can be seamlessly embedded in the PIs’ allocation on the departmental server. The costs (1000 EUR/year) will be covered by other on-going projects at that point in time.

- For research data, at current archiving costs of 10 Euro/(TB*year), we estimate a cost of 2000 Euro/year. These costs will be covered by funding acquired by the project PIs in the context of other research projects. VLIR infrastructure funding for long-term data storage has been awarded to Prof. Voet and Prof. Sifrim.

7. **Data sharing and re-use**

7.1 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 or because of IP potential)?

7.2 Yes. Personal data (including sequencing and/or spatial omics data obtained from human material) will only be published after pseudonymisation, and identifiers will not be published. Patients are informed via the informed consent forms about the policies regarding data sharing. Also data sharing restrictions might potentially apply due to generation of IP. Regular meetings with KU Leuven LRD will be held to evaluate and protect possible IP generated during the project that could lead to valorization actions. If deemed necessary, data that fall under IP will either not be shared, put under embargo, or a suitable license will be applied to the data when published (e.g. Creative Commons License).

7.3 Which data will be made available after the end of the project?

Relevant digital data will be published and made available after the end of the project. Data with valuable IP will be protected prior to publication. We will comply with open access regulations of KU Leuven.

7.4 Where/how will the data be made available for reuse?

The approach to share data is upon request by e-mail. Due to the data volume, access will then be granted to a restricted access repository.

7.5 When will the data be made available?

As soon as the research results have been published, the data can be made available to other researchers.

7.6 Who will be able to access the data and under what conditions?

All project collaborators will be authorized to have access to all obtained digital and physical data after the project. In case the question originates by researchers outside the consortium, the data can be made available upon e-mail request, and on condition that the users agree to give proper credit, such as co-authorship on their papers building on these data. Usage for commercial purposes will require obtaining a license, or equivalent arrangement.

7.7 What are the expected costs for data sharing? How will these costs be covered?

A restricted access repository can be implemented on a free tool, such as Dropbox, up to a certain volume. If this volume does not suffice, time-limited storage will be considered, thus limited to the time needed to download the data.

8. **Responsibilities**

8.1 Who will be responsible for the data documentation & metadata?

Research and technical staff working on this KUL ID-N project will be responsible for the data collection, documentation and metadata. They will be trained in data management.. Supervisors will manage the data storage facilities.

8.2 Who will be responsible for data storage & back up during the project?

The PIs, research and technical staff working on this KUL ID-N project will be responsible to store the data on the appropriate accommodation provided by KU Leuven. The ICTS service of KU Leuven is responsible for the back-up of the network drives at KU Leuven. The folders will be managed by the supervisors.

8.3 Who will be responsible for ensuring data preservation and sharing?

The PIs of this project will be responsible for the data preservation and eventual reuse of obtained data.

8.4 Who bears the end responsibility for updating & implementing this DMP?

The PIs bear the end responsibility of updating and implementing the DMP.