DMP title

Project Name C1 DMP 2021 Koh - DMP title

Project Identifier ZKE0470

Grant Title C1421117

Principal Investigator / Researcher Kian Koh

Project Data Contact kian.koh@kuleuven.be

Description TET family dioxygenases erase DNA methylation. Loss of Tet1 in the pre-gastrulation mouse epiblast causes dysregulation of chromatin accessibility and DNA methylation at neural fate genes prior to their gene activation in development, resulting in congenital and post-natal neurodevelopmental disorders in mutant mice. Here, we aim to translate our findings from mouse to human, by exploiting in vitro differentiation of human embryonic stem cells into neuronal progenitor cells and 3D micro-patterned organoids. We will examine temporal changes in gene expression, chromatin accessibility and DNA methylation in the neuronal cultures and their dependence on TET and co-factors. We will use next-generation sequencing to discover how disease-associated genetic variants, the epigenetic dysregulation related to non-genetic factors, or the combination of both downstream of TET or co-factor deficiency in early life can tip the balance between health and disease later.

Institution KU Leuven

1. General Information

Name of the project lead (PI)

Kian Koh

Internal Funds Project number & title

C1421117: The early embryonic function of TET DNA dioxygenases in human neurulation and disease.

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.

Description of data by work packages

WP1

Data type 1: **Digital images.** (<1 GB)

- Fluorescence images of human embryonic stem cell lines;
- Gel photos and scans of blots to validate CRISPR-Cas9 mutations.

Data type 2: **Flow cytometry data.** (<1 GB)

- FACS acquisition plots for GFP sorting of CRIPSR-Cas9 transfected cells.

Data type 3: **Next-generation sequencing data.**

- RNA-seq of time-course differentiation comparing control and mutant hESC lines, 48 samples at 2 GB per sample (96 GB total).
- ATAC-seq of time-course differentiation comparing control and mutant hESC lines, 48 samples, 6 GB per sample (288 GB total).
- Whole-genome bisulfite sequencing in biological duplicates at up to 3-4 time-points, 16 samples, 45 Gb for 15x coverage per sample (720 GB).
- PLAC-seq, 24 samples, 5 GB per sample (120 GB).

Data type 4: **Quantitative measurements** (<1 GB)

- RT-qPCR of gene expression in neuronal cultures

WP2

Data type 1: **Digital images** (<1 GB)

- 5hmC DNA dot blots for L-AA bioactivity
- 5mC mass spectrometry (MS)

Data type 2: **Next-generation sequencing data.**

- RNA-seq of hESC-neural differentiation cultures treated versus untreated with L-ascorbic acid, 12 samples, 2 GB per sample (24 GB total).
- Reduced representation or targeted bisulphite sequencing of cells treated versus untreated with L-ascorbic acid, 12 samples, 10 GB per sample (120 GB).
- CUT&RUN of cells treated versus untreated with L-ascorbic acid, 48 samples, 3 GB per sample (144 GB total).

WP3
Data type 1: **Digital images** (<1 GB)

- Microscopic images of 3D neuruloids;
- Immunofluorescence staining of neuronal derivatives.

Data type 2: **Flow cytometry data.** (<1 GB)

- FACS acquisition plots for single cell sorting of neuruloids.

Data type 3: **Next-generation sequencing data (scNMTseq).**

- SMART-seq transcriptome datasets of single-cells, 4x2x100=800 cells, 40 GB
- DNA methylome 150PE sequencing of single-cells, 4x2x40=320 cells, 2 TB

WP4
Data type 1: **Digital images.** (<1 GB)

- Gel photos and scans of blots to validate CRISPR-Cas9 mutations and cDNA constructs.

Data type 2: **Flow cytometry data.** (<1 GB)

- FACS acquisition plots for GFP sorting of dCas9 transfected cells.

Data type 3: **Quantitative measurements** (<1 GB)

- RT-qPCR of gene expression in ESCs
- ChIP-qPCR of histone modifications and signal effectors at target genes
- Statistical analysis

Data type 4: **Sanger DNA sequencing chromatograms** (<1 GB)

- Sequence validation of DNA constructs

Data type 5: **Next-generation sequencing data**

- CUT&RUN of histone modifications, 3 GB per sample.
- targeted bisulfite amplicon sequencing, 2 GB.

Data Type	Format	Estimated Volume	Data Origin
1. Digital images	Raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif.	4 GB	Microscopy, photoimaging and scanning
2. FACS	Raw data: Flow Cytometry Standard (.fcs); Analysis: FlowJo.	2 GB	FACS Aria III, Canto, Symphony
3. NGS	Raw data: .fastq(.gz); Sequence alignment data: .sam, .bam; Coverage data: .bed, .bg, .bedGraph, .bw, .bigwig; Quantitative tabulation: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx).	8 TB	HiSeq4000, NextSeq and NovaSeq platforms with multiplex DNA library input.
4. Quantitative	Tabular: MS Excel (.xls) Statistical analysis: Prism (.pzfx)	1 GB	StepOnePlus real-time PCR ; Graphpad Prism software
5. Chromatograms	Nucleotide sequences: raw sequence data trace (.ab1), text-based format (.fasta/.fa) and accompanying QUAL file (.qual), Genbank format (.gb/.gbk)	1 GB	GATC sequencing services

6. Presentation	Text files: MS Word (.doc/.docx), Adobe Portable Document Format (.pdf), LaTex (.tex) format;
	Presentations: MS Powerpoint (.ppt);
	Digital images in vector formats: Encapsulated Postscript (.eps), Scalable Vector Graphics (.svg), Adobe Illustrator (.ai);
		1 GB	Word processing, presentation and graphical softwares

Raw as well as processed data will be submitted to a public repository in the afore-described standard (preferably open) formats.

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.

No.

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).

Licensing approvals for the use of human ESC lines H1 and H9 in biomedical research have been obtained from WiCell. Ethical clearance have been given by the Ethics Committee Research UZ/KU Leuven under Project S65684 entitled “TET DNA dioxygenase gene function in human pluripotency transition and neural induction.”

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?

This project is part of a collaborative effort to understand the impact of peri-conceptual nutrition and epigenome perturbation on early human neurodevelopment using human ESCs. The outcomes inform the potential causal roles of epigenetic aberrations during early embryonic development on disease susceptibilities; alternatively, a correlative role of disease-associated epigenetic variations may serve as biomarkers. Part of the work will be performed in collaboration with Prof. Richard Finnell and co-workers at the Baylor College of Medicine, Houston, TX, U.S.A. under a Research and Collaboration agreement signed by KU Leuven Research & Development and Baylor College of Medicine Office of Research on July 12, 2019. We foresee that some of the research results may have potential for tech transfer and valorisation. This will be evaluated together with KU Leuven LRD.

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?

No.

4. Documentation and metadata

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

Digital data:

All experimental procedures are documented as protocols in MS Word and saved on KU Leuven shared drives in folders classified by methodology subject (tissue culture, cellular assays, protein, nucleic acid, molecular cloning, …) and accessible to all group members. Specific experimental details (time, consignee, protocol, samples names, conditions, ...) are recorded in hard-copy laboratory notebooks, compiled with hard-copy and electronic versions of content pages listing titles of experiments by page and date to allow quick reference. Computational scripts are saved in Plain text data format (Unicode, .txt) or Hypertext Markup Language (.html) on GitHub. Bioanalyzer reports for all NGS library preparations are filed as hard copies per submission with attached readme doc. sheets recording library preparation parameters. Raw and analyzed data files (specific file format according to data type) are saved with a structured file name (experimental code name, number and date) on KU Leuven storage drives in a folder structure organized by WP/Experimenter Name/Type of Experiment. Index/read me files (.txt file) for each WP contain links and location (folder and subfolder on BOX/server/hard disk) of above-mentioned files. In the concluding stage of the project, a master index file containing the combined information for all the WPs will be compiled to list the final storage location and organization in the KU Leuven archive Large Data storage drive.

Physical data:

Viral vectors, cell lines (frozen stocks) are filed in a database (xls) available via the central server through a folder that is available for members of the Stem Cell Institute. In addition, an electronic inventory of frozen cell lines (labelled with user initial, description, passage number and freeze date) organized by stack/box/grid position is maintained to facilitate retrieval and distribution in the ON4 Biobank. All oligonucleotides synthesized by IDT are named according to standard lab nomenclature containing gene name and initials for user and application and stored at -20°C. RNA can be stored up to 5 years at -80°C. Sample tubes are labelled with date+ sample name in storage boxes organized by experiment. Storage location (in box grid position) of individual samples are recorded on Excel in .xls format. All plasmid constructs generated by the laboratory are deposited at Addgene under Kian Peng Koh Lab Plasmids.

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.

NGS data types are ultimately deposited in public repositories, such as GEO, SRA, ArrayExpress, or ENA, using sequencing metadata spreadsheet templates required by each repository. In these metadata schemes, technical and analytical methods, in addition to sample preparation details, will be documented in sufficient detail to allow for independent reproduction. When depositing data in a repository, 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 allow the data to be understood by others and add context to the dataset for future reuse.

For imaging files., metadata are inherently included in the header files of the original images (author, date, ...) as provided by the manufacturer of the machinery· (confocal, Nikon Ti2, ...). These files contain information regarding the acquisition settings.

5. Data storage and backup during the project

5.1. Where will the data be stored?

Digital data will be stored on KU Leuven servers.

- NGS data and microscopic images will be stored on Large Volume Storage (L-drive) at KU Leuven, maintained by ICTS. In addition, we have contributed to two purchased nodes on the Flemish Supercomputer Centre (VSC) and purchased 4 TB of staging space, with future plans to transfer published data to archive area.
- Other smaller data files (manuscript files and reports) will be stored on KU Leuven shared J-drive.
- Algorithms and scripts are stored in private online git repositories owned by the PI (https://github.com/KianKoh-Lab).

5.2. How will the data be backed up?

Data stored on KU Leuven's secure network drives are automatically backed-up by daily procedures. Additionally, a mirror of the data is provided in a second ICTS data center for business continuity or disaster recovery purposes.

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, currently we have 5Tb of purchased storage space on the KU Leuven L drive, which can be expanded in blocks of 5Tb when required. J- drive storage is available in expandable blocks of 100GB for smaller datasets. In addition, we have purchased 3 TB of Staging and Archive on VSC.

5.4. 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 about 3,000 EUR. This estimation is based on the following costs:

- The costs of data storage on KU Leuven servers: 173,78 EUR/TB/year for the "L-drive" and 519 EUR/TB/year for the "J-drive".
- The cost of VSC staging: 140 EUR/TB/year.

We expect costs to drop slightly during the coming four years.

These costs will be covered by the project budget and/or SCIL institutional funds.

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

The L- and J-drive servers are accessible only by laboratory members. The VSC storage is only accessible via VSC accounts which operate with password access-protection by users and person-based decision on rights to access and modify data. Specifically our volume will only be accessible to project researchers and only the PI can grant access. For extra safety, raw sequencing data are backed-up on external hard disks kept at the office of the PI.

6. Data preservation after the end of the 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, ...).

Digital data: Upon acceptance of manuscripts for publication, data will be made publicly available on repositories such as GEO, submitted via EBI-ENA/NCBI-SRA, where they will be permanently archived to preserve access to the public. All computational scripts stored on Github repositories pertaining to published datasets will be made public. At the end of the project, all published data will be transferred from L-/J- drives to VSC archive for a minimum storage term of 5 years. Imaging data will be submitted to Image Data Resource.

Physical data: Plasmid vectors are stored as purified DNA in -20°C freezer and as a bacteria glycerol stock at -80°C as back-up. 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.

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

Published sequencing data will be deposited in public repositories with GEO accession IDs for public access. For all other datasets, long term storage will be ensured as follows:

- Large sequencing/omics data: will be stored on VSC archive
- Small digital files: files will be stored on the KU Leuven archive ''K-drive".
- Developed algorithms and software will be stored on VSC archive and/or L drive,

as well on public repositories such as Github.com.

- Third-party software and algorithms that are used are referenced by their version numbers (e.g., in our Github repository) and are installed as modules on the VSC and/or containers on the VSC, to ensure reproducibility.

All departing staff submit a closing inventory list of all reagents and samples stored in freezers. All laboratory notebooks and microfilms obtained from chemiluminescent or autoradiographic blots will be stored in the PI’s office.

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

- The cost of digital data storage on the K-drive is 128, 84 EUR/TB/year.
- The cost of VSC archive is 70 EUR/TB/year.

Published NGS datasets and scripts are deposited in public repositories for free. We estimate that up to 2TB of additional data will be preserved on VSC archive for 10 years after the end of the project, at an additional cost of 1000 EUR. These costs will be covered by the project budget and/or SCIL institutional funds and paid up-front.

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)?

No personal data are involved in this project. There are no restrictions on the sharing of data.

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

All published data will be publicly available.

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

- In an Open Access repository
- Upon request by mail

Omics datasets will be deposited in open access repositories such as NCBI Gene Expression Omnibus (GEO).

Other digital datasets that support publications (including image, cytometry data) will be made publicly available via an open research data platform such as Mendeley Data or Image Data Resource (http://idr.openmicroscopy.org/about/.)

All the relevant algorithms, scripts and software code driving the project will be stored in a private online Github repository managed by the PI, and limited to project members during the project. As soon as the manuscript is publicly available, the repository will be changed to public.

7.4. When will the data be made available?

- Upon publication of the research results

Upon publication of the research results. All post-acceptance manuscripts are uploaded on the KU Leuven repository LIRIAS to be immediately available for free under Green Open Access to the public.

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

Only researchers participating in the project will be able to access the data before publication. Upon publication all datasets and the appropriate metadata will be made publicly available through repositories that support FAIR sharing (eg. via NCBI-SRA using GEO accession IDs).

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

Publishing costs, including Gold Open Access as required by specific journals (eg. Cell Reports) will be covered by project grants. No costs are expected for public free-to-use data repositories.

8. Responsibilities

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

The PhD and postdoctoral researchers associated with this project will be responsible for data documentation & metadata, under supervision of the PI.

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

Research staff will be responsible for data storage. KU Leuven storage drives are maintained by ICTS. Administrative support is provided by SCIL secretariat, Christina Vochten.

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

The PI, Kian Koh, will be responsible for ensuring data preservation and reuse, with support from ICTS.

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

The end responsibility for updating and implementing the DMP is with the supervisor (promotor).