cfDNA methylome analysis: novel applications

A Data Management Plan created using DMPonline.be

Creators: Mio Aerden, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: n.n. n.n.

Grant number / URL: 1S38923N

ID: 198141

Start date: 01-11-2022

End date: 30-11-2027

Project abstract:

Cell-free DNA (cfDNA) methylation analyses are increasingly being implemented for noninvasive screening, diagnosis and monitoring of physiological and pathological conditions. Here, I propose to expand these applications by focussing on 3 groups: healthy individuals, pregnant women and twins. First, I will assess intra-individual variation in cfDNA methylation by generating longitudinal and age-specific methylome data from healthy individuals, thus providing essential insights into the physiological variability and dynamics of cfDNA. In addition, the host lab developed a cfDNA methylation signature to predict early-onset preeclampsia presymptomatically in the first trimester of pregnancy. I will test if this signature also predicts late-onset preeclampsia.

Next, I will develop a cfDNA methylation-based test to detect vanishing twins. This will provide not only basic knowledge but also a better interpretability of non-invasive prenatal tests.

Lastly, I will assess if vanishing twin survivors carry a specific DNA methylation imprint, and test if that imprint elucidates the etiology of oculo-auriculo-vertebral-spectrum, a poorly understood, non-genetic developmental disorder.

Gains to be expected from this project include early detection of preeclampsia and vanishing twins, thus reducing healthcare costs. In addition, charting cfDNA methylome variation will provide valuable information for screening programs and more tailored monitoring regimes in the near future.

Last modified: 21-04-2023

cfDNA methylome analysis: novel applications
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

Question not answered.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

Question not answered.

What’s the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

Question not answered.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

Question not answered.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

Question not answered.

cfDNA methylome analysis: novel applications
DPIA

DPIA

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

- Not applicable

cfDNA methylome analysis: novel applications
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

cfDNA methylome analysis: novel applications
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

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

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
LongVar	Longitudinal (n= 150) and diurnal (n= 16) medical, socio-demographic and cell-free DNA methylome data of healthy individuals	Generate new data	Digital and physical	Observational and experimental	.txt, .doc/.docx, .xls/.xlsx, .pdf, .tex, .csv, .tab, .txt, .jpg/jpeg/png, .svg, .eps, .svg, .ab1, .fasta/.fa, .qual, .gb/.gbk, .bcl, .fastq(.gz), .bam, .bed, .bg, .bedGraph, .bw, .bigwig, .vcf(.gz), .bcf, .tsv(.gz), .mtx, .loom, .rds(.gz), .cov	< 10TB	Paper questionnaires (n~ 364) + plasma samples (n~ 2300 - 2400 x 2ml) + WBC (n~ 1150 - 1200)
LOPE	Maternal and pregnancy-related health information of women with late-onset preeclampsia (n= 100) + methylome analysis of leftover cfDNA from NIPT	Generate new data + reuse existing data	Digital and physical	Observational and experimental	.txt, .doc/.docx, .xls/.xlsx, .pdf, .tex, .csv, .tab, .txt, .jpg/jpeg/png, .svg, .eps, .svg, .ab1, .fasta/.fa, .qual, .gb/.gbk, .bcl, .fastq(.gz), .bam, .bed, .bg, .bedGraph, .bw, .bigwig, .vcf(.gz), .bcf, .tsv(.gz), .mtx, .loom, .rds(.gz), .cov	< 10TB	Leftover cfDNA samples from NIPT (n~ 100)
VTwin	Maternal and pregnancy-related health information of vanishing twin (n= 50 + 8), twin (n= 50) and singleton (n= 50 + 40) pregnancies + methylome analysis of cfDNA from maternal blood (n= 150 + 48) and gDNA from cord blood (n= 48)	Generate new data + reuse existing da	Digital and physical	Observational and experimental	.txt, .doc/.docx, .xls/.xlsx, .pdf, .tex, .csv, .tab, .txt, .jpg/jpeg/png, .svg, .eps, .svg, .ab1, .fasta/.fa, .qual, .gb/.gbk, .bcl, .fastq(.gz), .bam, .bed, .bg, .bedGraph, .bw, .bigwig, .vcf(.gz), .bcf, .tsv(.gz), .mtx, .loom, .rds(.gz), .cov	< 10TB	Leftover cfDNA samples from NIPT (n~ 150) + plasma samples (n~ 160 x 2ml) + WBC isolated from cord blood samples (n~ 48)
ADULTwins	Medical, socio-demographic and DNA methylome information on VT survivors (n= 48), dizygotic twins (n= 48 - one of pair), monozygotic twins (n= 48 - one of pair) and non twins (n= 48)	Generate new data	Digital and physical	Observational and experimental	.txt, .doc/.docx, .xls/.xlsx, .pdf, .tex, .csv, .tab, .txt, .jpg/jpeg/png, .svg, .eps, .svg, .ab1, .fasta/.fa, .qual, .gb/.gbk, .bcl, .fastq(.gz), .bam, .bed, .bg, .bedGraph, .bw, .bigwig, .vcf(.gz), .bcf, .tsv(.gz), .mtx, .loom, .rds(.gz), .cov	< 10TB	WBC isolated from whole blood samples (n~ 384)
OAVSet	Pregnancy-related, clinical and DNA methylome information of oculo-auriculo-vertebral spectrum (OAVS) patients (n= 24)	Generate new data	Digital and physical	Observational and experimental	.txt, .doc/.docx, .xls/.xlsx, .pdf, .tex, .csv, .tab, .txt, .jpg/jpeg/png, .svg, .eps, .svg, .ab1, .fasta/.fa, .qual, .gb/.gbk, .bcl, .fastq(.gz), .bam, .bed, .bg, .bedGraph, .bw, .bigwig, .vcf(.gz), .bcf, .tsv(.gz), .mtx, .loom, .rds(.gz), .cov	< 10TB	WBC isolated from whole blood samples (n~ 48)

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:

Reuse of data includes existing health-related information extracted from patient electronic health files through the 'Klinisch Werkstation or KWS'.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

All datasets include personal data, clinical parameters and (epi)genetic information.
For personal and sensitive data, I 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.
Ethical approval for the following datasets has already been obtained:
- LongVar: S66450
- LOPE: S61883
- VTwin: S67260
Data collection for the ADULTwin and OAVSet dataset has not yet started. Ethical approval for these datasets will be sought at a later stage.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

- **LongVar**:
1. All study participants are asked to fill-in a self-completing health questionnaire with standard socio-demographic (age, sex, ethnicity, weight, and height) and health-related questions (awareness of any diseases and smoking/drinking habits).
2. Methylation data generated from plasma cfDNA, and (possibly) genomic DNA extracted from the buffy coat.
- **LOPE**:
1. Patient information that is required includes: Gestational age at time of delivery, gestational age at time of preeclampsia diagnosis, diagnostic criteria for preeclampsia.
2. Relevant maternal characteristics (maternal age at time of conception, BMI, preeclampsia risk factors such as diabetes and chronic hypertension and smoking status).
3. Obstetrical information (twin pregnancy, mode of conception and obstetrical history).
4. Sex newborn and intra uterine growth restriction.
5. Methylation data generated from plasma cfDNA.
- **VTwin**: The following personal data will be generated and stored:
1. Time of vanishing twin detection.
2. Relevant maternal characteristics (maternal age at time of conception, BMI, vanishing twin risk factors such as maternal infection, comorbidities, smoking and medication).
3. Obstetrical information (gestational age, twin pregnancy, mode of conception and obstetrical history, sex of the newborn, fetal growth).
4. Methylation data generated from plasma cfDNA, and genomic DNA extracted from the buffy coat.
For the **ADULTwin** and **OAVSet** relevant personal information will be specified later.
All personal data will pseudonymised, either through generation of a study-specific code by the investigator or through generation of a pseudonymisation number in the KWS. Patient information that can uniquely identify individuals (date of birth, address, date of diagnosis, …) will explicitly not be requested.

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

There is potential tech transfer/valorisation in both the novel models we develop (genetic profiles with clinical information) and the putative assays/biomarkers we measure. We will contact with LRD regarding these issues.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

We intend to seek timely protection of the intellectual property generated in this project: prior to submission of manuscripts describing our findings, to presentation of our findings at external scientific meetings or to the end of the project. Hence, no restrictions on data sharing will be required at the moment when data will be made available (at publication or at the end of the project).

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

For every dataset the following information will be kept for facilitation of collaboration and reuse:
- Folder with ethical approval, informed consent documents, study protocol, questionnaires used/patient characteristics, collected, etc.
- Sample tracking (.xlsx) document with information on enrolment, sample collection, sample quality, storage location, etc.

- An electronic laboratory notebook (E-notebook) to document laboratory experiments and protocols.
- Data will be shared in formats that are routinely used in the field (e.g. FASTQ files for raw sequencing data). Samples will be analysed following a controlled vocabulary.
- Should custom scripts be generated in the current project, these will be made accessible on a dedicated GitHub page, along with documentation.
- Software tools provided by external developers cannot be shared as source code by us. We will log and share version numbers, to facilitate reproducibility of the data analyses.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Metadata will be documented by the research 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.
From the onset of the project, data documentation will be tailored for their ultimate deposition in public repositories, with spreadsheet headers corresponding to fields required by these public repositories. Technical and analytical methods used to generate the data 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 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.

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

Where will the data be stored?

The data will be stored on KU Leuven servers, where it can only be accessed by the involved researchers. All drives are managed by KU Leuven personnel, bound by the KU Leuven general and ICT codes of conduct. Omics data generated during the project will either be stored on KU Leuven servers or on The Flemish Supercomputer Centre (VSC).

How will the data be backed up?

* KU Leuven servers
Backups of KU Leuven drives are made using “snapshot” technology, which is the online storage of incremental data changes. For the standard backup regime, as specified below, 10% of the requested storage capacity will be reserved:
- An hourly backup (at 8 AM, 12 PM, 4 PM and 8 PM) the last 6 of which are stored on the KU Leuven servers
- A daily backup, at midnight, the last 6 of which are stored on the KU Leuven servers
- A weekly backup, Saturday night at midnight, the last 12 of which are stored on the KU Leuven servers
The end user can use his own Windows PC to restore files to an older version using the “previous versions” function. According to the backup system above, it is possible to go back in time upto 12 weeks (~3 months).
* VSC ($VSC_HOME and $VSC_DATA)
The backup consists of snapshots that are created at regular intervals:
- for the past 24 hours, one snapshot per hour is avialable (hourly.<timestamp>);
- for the past week, 6 snapshots are available, created once a day (daily.<timestamp>);
- for the past month, 4 snapshots are available, created weekly (weekly.<timestamp>).
An offsite backup is maintained for older data, but operator intervention is required to restore that.

**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

Yes. The necessary funding for storage and backup for the contracted service has been foreseen.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

All ICT solutions at KU Leuven are subject to the university-wide ICT information security standards. The faculty’s ICT service organizes the raw network storage it procures from central ICT services in such a way that access permissions are limited, fixed, delegated to and audited by data managers who do not need to have an IT background.
Digital data will be stored in a restricted network share on the KU Leuven servers, which can only be accessed by the involved researchers.
Access to code files with pseudonymisation keys is controlled by the responsible PhD student (with the PI as a back-up). All other researchers who participate in the project have access to the pseudonymised data only.
The VSC storage is only accessible to VSC accounts, and specifically our volume will only be accessible to group members.

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

Data storage on the KU Leuven servers costs either €104,42/TB/year (for L- and K-drive) or €503,66/TB/year (J-drive). Implementation of data back-up and storage is currently being organised and implemented in the lab. VSC costs are €20/TB/year for HPC staging storage (during the project) and €70/TB/year for HPC archive (after the project). The necessary funding for storage and backup for the contracted service has been foreseen.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All generated data will be preserved for 10 years according to KU Leuven RDM policy.

Where will these data be archived (stored and curated for the long-term)?

The data will be archived on the KULeuven server. Conform the KU Lueven RDM policy, research data will be made available on request to the KU Leuven at any time during the research, and will be stored for a minimum of 10 years after the publication of the results or the end of the period of the project funding.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Data will either be deposited in public repositories (see below), which is usually free or stored at the KU Leuven servers. The price for long time storage on the KU Leuven servers is €52,21/TB/year (= 50% of €104,42/TB/year as the the Group Biomedical Sciences sponsors 50% of the cost price). Data management costs will be covered by the laboratory budget.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository
- Yes, in a restricted access repository (after approval, institutional access only, …)

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).

Re-use of patient data is restricted by the GDPR and by the informed consent forms, and patient data will therefore be made available under restricted access, with access rights being decided by a local data access committee following the submission of a data analysis proposal.

If access is restricted, please specify 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. 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 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. A CC-BY license will be opted for when possible. Participants' personal information (e.g., contact information, names, etc.) will never be shared.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Ethical aspects
- Yes, Privacy aspects

Yes. Re-use of patient data is restricted by the GDPR and by the informed consent forms, and patient data will therefore be made available under restricted access, with access rights being decided by a local data access committee following the submission of a data analysis proposal.

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

Patient data: Upon publication, all (linked) anonymized patient details supporting a manuscript will be made publicly available as supplemental information.

Genetic, transcriptomic and methylation datasets will be deposited in open access repositories such as the EBI ArrayExpress databases for functional genomics data or the EBI European Genome- phenome Archive (EGA) for personally identifiable genetic and phenotypic data. Data at EGA will be collected from individuals whose consent agreements authorise data release only for specific research use to bona fide researchers.

Research documentation: All protocols used to generate published data will be described in the corresponding manuscript(s), and the related documentation will be included as supplementary information. These data and all other documents (daily logs, raw data) deposited in the E-Notebook are accessible to the PI and the research staff, and will be made available upon request.

Manuscripts: All scientific publications will be shared under open access. Manuscripts submitted for publication will be deposited in a pre-print server such as bioRxiv, arXiv, Nature Precedings or ASAPbio. At the time of publication, research results will be summarised on the PI's website (https://gbiomed.kuleuven.be/epigenetics), through social media channels, through press releases where appropriate, and post-print pdf versions of publications will be made available there if allowed by copyright agreements, possibly after an embargo as determined by the publisher. Before the end of the embargo or in cases where sharing the post-print is not allowed due to copyright agreements, a pre-print version of the manuscript will be made available. Publications will also be automatically added to our institutional repository, Lirias 2.0, based on the authors name and ORCID ID (the metadata will be added, not the full manuscripts).

When will the data be made available?

Upon publication of research results.

As a general rule all research outputs will be made openly accessible at the latest at the time of publication. No embargo will be foreseen unless imposed e.g. by pending publications, potential IP requirements.

Which data usage licenses are you going to provide? If none, please explain why.

All data will be self generated and will therefore not be bound to a data usage license. Should data be reused later in the project, we will comply to the license under which we received the data. Generated (sequencing) cannot be licensed and will be deposited according to the repositories standard terms and conditions.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

A permanent identifier will be added to the data upon deposit in a repository.

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

Deposition of the data in repositories will be either free of charge or costs will be covered by the laboratory budget. If physical data (such as samples) should be shared, the costs will be paid by the researcher requesting the materials.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Metadata will be documented by the PhD students 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 manage data storage and backup during the research project?

The research and technical staff will ensure data storage and back up, with support from ICTS, gbiomed-IT staff, and UZ-IT staff.

Who will manage data preservation and sharing?

The PI is responsible for data preservation and sharing, with support from ICTS, gbiomed-IT staff, and UZ-IT staff.

Who will update and implement this DMP?

The PI is ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.