From Y-chromosome to microbiome: an interdisciplinary kinship study (version of 19 April 2023)

A Data Management Plan created using DMPonline.be

Creator: Heleen Coreelman

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Heleen Coreelman

Data Manager: Heleen Coreelman

Project Administrator: Heleen Coreelman

Grant number / **URL:** 1154123N

ID: 198413

Start date: 01-11-2022

End date: 31-10-2026

Project abstract:

Both inside and out, our bodies harbor a thriving diversity of micro-organisms each with their own genome, collectively referred to as our microbiome. Several studies state that the microbiome of relatives is more similar than the microbiome of non-relatives. Though, it is not clear whether this is caused by their shared DNA rather than environment. While autosomal DNA inheritance from parent to child undergoes recombination causing kinships to fade over generations, the male-specific Y-chromosome (chrY) lacks recombination making chrY paternally inherited in a rather conserved manner. However, the chrY DNA code can slightly change over generations due to small mutations. Two interesting Y-mutations over time are single nucleotide polymorphisms (Y-SNPs) and short tandem repeats (Y-STRs). Y-SNPs mutate slowly and divide all men around the world in 20 evolutionary Y-haplogroups, while Y-STRs mutate rapidly and allow to characterize a paternal lineage. Using chrY, we can identify evolutionary, distantly and closely related men. Our novel sequencing panel, the CSYseq, is the first panel which identifies both evolutionary and familial kinships. With this interdisciplinary project, I aim to study the association between host genetics and microbiome composition using deep-rooted pedigrees of paternal relatives. Therefore, I will perform in-depth sequencing on microbial and human DNA of males from Belgium and the Netherlands. It is my ambition to provide novel and valuable perspectives across two disciplines: Evolutionary Biology and Forensic Genetics. Through this interdisciplinary approach, I will (a) reveal the link between the microbiome and host genetics, (b) unravel the influence of degree of kinship and environment on microbiome composition, and (c) evaluate the application of microbiome analysis and CSYseq in a forensic context. The key deliverable is the extension of a genetic genealogical database, linking a chrY dataset with microbiome data. Overall, this pioneer study will increase the importance of microbiomics and genetics for interdisciplinary population research worldwide.

Last modified: 19-04-2023

From Y-chromosome to microbiome: an interdisciplinary kinship study (version of 19 April 2023)
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)

According to the KU Leuven comment (i.e., guidance tab), I don't have to fill and submit this section again since my fellowship already started.

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)

According to the KU Leuven comment (i.e., guidance tab), I don't have to fill and submit this section again since my fellowship already started.

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)

According to the KU Leuven comment (i.e., guidance tab), I don't have to fill and submit this section again since my fellowship already started.

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)

According to the KU Leuven comment (i.e., guidance tab), I don't have to fill and submit this section again since my fellowship already started.

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

According to the KU Leuven comment (i.e., guidance tab), I don't have to fill and submit this section again since my fellowship already started.

From Y-chromosome to microbiome: an interdisciplinary kinship study (version of 19 April 2023)
DPIA

DPIA

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

- Not applicable

From Y-chromosome to microbiome: an interdisciplinary kinship study (version of 19 April 2023)

GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

From Y-chromosome to microbiome: an interdisciplinary kinship study (version of 19 April 2023)
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 or research materials	**Description**	**New or reused**	**Physical or digital**	**Digital data type**	**Digital data format**	**Digital data volume (MB/GB/TB)**	**Physical volume**
Informed consent forms	The participants will sign an informed consent form in which they indicate that their data and their samples may be analysed by the researchers and that the results will be published. The participants can indicate whether or not the DNA from their samples and the remaining of the samples may be stored (storage for 10 years is guaranteed).	Newly collected data	Physical and digital	Observational	.pdf	< 100 MB	Paper data will be stored inside KU Leuven - in the office of the primary researcher in a locked drawer or cupboard that can only be accessed by the researcher. The data will be scanned and saved as PDF files on network drives.
Questionnaires	The participants will be asked to complete a questionnaire with data on influencing factors of the microbiome (age, antibiotic usage, dietary and drinking pattern, dominant hand, personal hygiene, residence, occupation, smoking status, traveling lifestyle, etc.) and the genealogical data of his paternal line (name, date and place of birth, profession, and number of biologically related sons/daughters of his father, grandfather and great-grandfather).	Newly collected data	Digital	Observational	.csv; .xlsx; .accdb	< 100 MB	Not applicable
DNA samples	For Y-chromosome and mitochondrial DNA profiling, a buccal mucosa sample will be taken from each participant. For microbiome profiling, a buccal mucosa sample (oral microbiome) and hand palm sample (skin microbiome) will be taken from each participant.	Newly collected data	Physical	Not applicable	Not applicable	Not applicable	The (remaining) material will be stored - Biobank approval obtained. The physical volume is estimated at 25 boxes.
Raw high throughput sequencing data	Original DNA quantity and high throughput sequencing data of the Y-chromosome, mitochondrial DNA, oral microbiome and hand palm microbiome of the participants.	Newly generated data	Digital	Experimental	.xlsx; .accdb; .fastq	< 1 TB	Not applicable
Data processing tools	After de-multiplexing using index-barcodes, primer trimming and quality control, the FASTQ files will be analysed for variant calling (Y-chromosome and mitochondrial DNA) and determination of microbiome composition.	Newly generated data	Digital	Software	.html; .SAM; .BAM; .BAI; .VCF; .py; .R; .xlsx	< 1 TB	Not applicable
Processed data	From each participant, Y-SNP haplogroup, Y-STR haplotype, mtDNA-haplogroup and microbiome data will be generated.	Newly generated data	Digital	Experimental	.xlsx; .accdb; .R	< 100 MB	Not applicable
Genetic genealogical database	Paternally related males will be selected on the basis of data from the in-house genetic genealogical Y-chromosome database containing over 2,700 men or the response on the launched call. Ultimately, the genealogical database will be linked to host DNA (Y-chromosome and mitochondrial DNA) and microbiome data.	Reused existing data and newly generated data	Digital	Observational and experimental	.csv; .xlsx; .accdb	< 100 MB	Not applicable
Data analysis tools	Based on the raw and processed data, data are analysed.	Newly generated data	Digital	Software	.xlsx; .R; .sps; .sav	< 100 MB	Not applicable
Analysed data	Graphs, tables, etc. intended to present the conclusions.	Newly generated data	Digital	Experimental	.xlsx; .accdb; .R	< 100 MB	Not applicable

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:

Paternally related males that previously participated in other projects within the research group will be informed about the current project using their previously obtained contact details. In these previous studies, the participants needed to tick a box indicating whether or not they agree to be informed about further research. Only participants who agreed, will be contacted and informed about the current research project.

At the beginning of 2009, a large-scale genetic genealogical study within the Belgian male population was organised by Familiekunde Vlaanderen vzw and geneticists from KU Leuven. Through the years, the project of Familiekunde Vlaanderen vzw (research contract with LRD) has given rise to various 'citizen science' projects such as the 'Romeins DNA' project (S54010), the 'COR-' project (S55864), the 'Gen-iale Stamboom' (S55864), the identification of 'Sint-Idesbald' (S59085) and the 'My COR* family' project (S65250), which means that in 2023 the Y-chromosome of more than 2,700 Belgian and Dutch families has been analysed and linked to their genealogical origin.

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

Privacy/ethical approval is needed for the current study. The leading ethical committee in this project is EC Research: S67542 (ethical committee procedure is currently pending).

Human subject data:

- Genetic and microbiome data from DNA samples: Y-chromosome data (Y-SNPs and Y-STRs; non-coding regions), mitochondrial DNA data (mtDNA SNPs; non-coding regions), oral microbiome data, skin microbiome data
- Genealogical data: data on the paternal lineage of the participant (name, date of birth, place of birth, occupation, number of biologically related sons, number of biologically related daughters)
- Questionnaire data: data on influencing factors of the microbiome (age, antibiotic usage, dietary and drinking pattern, dominant hand, personal hygiene, residence, occupation, smoking status, traveling lifestyle, etc.)

Personal data will be pseudonymised. Data are coded as there continues to be a link between the data and the individual who provided it. The subject's name or other identifying information are stored separately from the research data and replaced with a unique code to create a new identity for the subject. The sample and the data are given a different code via an algorithm, so that there is no direct link between the personal data and results without using the algorithm. This algorithm will be saved in a password protected Excel file in a folder on the central KU server with access only for the involved KU Leuven investigators.

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

The GDPR questionnaire for the current study was submitted via the Privacy and Ethical Review (PRET) application: G-2023-6269 (date of approval: 15/02/2023).

Participants will be selected on the basis of data from the in-house genetic genealogical Y-chromosome database containing over 2,700 men or the response on the launched call. We are specifically looking for closely (i.e., twin brothers), distantly and evolutionarily related males.

(Special) categories of personal data:

- Contact details: personal data used for organising the research (name, surname, email address, etc.)
- Genetic and microbiome data from DNA samples: Y-chromosome data (Y-SNPs and Y-STRs; non-coding regions), mitochondrial DNA data (mtDNA SNPs; non-coding regions), oral microbiome data, skin microbiome data
- Genealogical data: data on the paternal lineage of the participant (name, date of birth, place of birth, occupation, number of biologically related sons, number of biologically related daughters)
- Questionnaire data: data on influencing factors of the microbiome data on influencing factors of the microbiome (age, antibiotic usage, dietary and drinking pattern, dominant hand, personal hygiene, residence, occupation, smoking status, traveling lifestyle, etc.)

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.

- No

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.

- No

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).

Paper data (informed constent forms, etc.) and the physical lab books that contain the lab protocols and results will be stored in the KU Leuven office of the primary researcher in a locked drawer or cupboard that can only be accessed by the researcher. All experimental data will be written down in detail in the physical lab books with the chronological reporting of all protocols, laboratory notes and results, in order to repeat every step in the lab.

Digital data (scanned informed consent forms, questionnaires, pseudonymisation key, protocols, manuals, raw and processed sequencing data, genetic genealogical database, data analysis results, etc.) are stored on secure KU Leuven network drives, which are password protected so that only members of the research team will have access to the documents. A clear folder structure will be provided to simplify retrieving and consulting the overall documentation. A meaningful file name will be provided, starting with the date (year/month/day), description of its content and ending with the initials of the investigator that created the file. If appropriate, a clear description file will also be available to keep all data understandable and usable for ourselves and all future researchers.

The (remaining) material of the DNA samples will be stored at KU Leuven - Kulak. Biobank approval is already obtained. Each tube will be labeled with the code of the participant, description of its content, date (year/month/day) and initials of the investigator that performed the protocol.

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.

- No

A README file will be created for this project to make it easier to find and reuse the data by ourselves and all future researchers. This file will introduce and explain the project. It will contain all the information that is commonly required to understand what the project is about. Secondary README files will be created in subfolders to document the specific parts of the data, which includes the content and structure of the dataset. Elements (general information, data collection, data organisation, processing software, quality control procedures, etc.) are only included in the README files if they are useful and/or necessary to correctly interpret, evaluate and reuse the dataset.

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

Where will the data be stored?

Paper data (informed constent forms, etc.) and the physical lab books that contain the lab protocols and results will be stored in the KU Leuven office of the primary researcher in a locked drawer or cupboard that can only be accessed by the researcher.

Digital data (scanned informed consent forms, questionnaires, pseudonymisation key, protocols, manuals, raw and processed sequencing data, genetic genealogical database, data analysis results, etc.) are stored on secure KU Leuven network drives, which are password protected so that only members of the research team will have access to the documents.

The (remaining) material of the DNA samples will be stored at KU Leuven - Kulak. Biobank approval is already obtained.

How will the data be backed up?

Paper data (informed constent forms, etc.) will be scanned and saved as PDF files on secure KU Leuven network drives.

Digital data (scanned informed consent forms, questionnaires, pseudonymisation key, protocols, manuals, raw and processed sequencing data, genetic genealogical database, data analysis results, etc.) 'from the past' can be restored on the KU Leuven network drives as files stored on the KU Leuven network drives are automatically backed up.

Besides, all the data will also be backed up on two hard drives of the research group.

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

Sufficient storage and backup capacity is available as the KU Leuven network drives have a capacity of 1 TB and the two hard drives of the research group have a capacity of 2 TB.

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

Paper data (informed constent forms, etc.) and the physical lab books that contain the lab protocols and results will be stored in the KU Leuven office of the primary researcher in a locked drawer or cupboard that can only be accessed by the researcher. The office will always be closed when leaving the desk.

Digital data (scanned informed consent forms, questionnaires, pseudonymisation key, protocols, manuals, raw and processed sequencing data, genetic genealogical database, data analysis results, etc.) are stored on secure KU Leuven network drives, which are password protected so that only members of the research team will have access to the documents. The screen lock will always be turned on when leaving the desk.

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

There are no expected costs for data storage and backup as the costs are already covered by the research group.

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 data will be retained for 10 years as storage for 10 years is guaranteed. Yet, the participant can indicate on the informed consent form whether or not the DNA from his samples and the remaining of his samples may be stored. In the latter case, the DNA from his samples and the remaining of the samples will be destroyed after the project. Nevertheless, the obtained data will

remain preserved and pseudonymised.

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

Long term data archives of paper data (informed constent forms, etc.) and the physical lab books will be maintained in the KU Leuven office of the promotor of the research project in a locked drawer or cupboard that can only be accessed by the promotors of the research project.

Long term data archives of digital data (scanned informed consent forms, questionnaires, pseudonymisation key, protocols, manuals, raw and processed sequencing data, genetic genealogical database, data analysis results, etc.) will be maintained on secure KU Leuven network drives, which are password protected so that only members of the research team will have access to the documents. The data will also be archived on the KU Leuven repository Lirias and two hard drives of the research group. Part of the data will be published in scientific journals and Open Access databases.

Long term data archives of the (remaining) material of the DNA samples will be maintained at KU Leuven - Kulak. Biobank approval is already obtained.

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

There are no expected costs for data preservation as the costs are already covered by the research group.

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

Only the involved KU Leuven researchers have access to the data during and after the project. If the participant chooses on the informed consent form that the DNA from his samples and the remaining of his samples may be stored and if he is eligible for a new study of the research group that is approved by an ethical committee, he will be contacted with an information letter regarding this new project and has to sign a new consent form for approval.

Part of the data will be published anonymously in scientific journals and Open Access databases. Non-published data will remain confidential until a final decision on publication of the data has been taken.

After the project, pseudonymised data will be made available for or shared with researchers after written permission of the involved KU Leuven researchers of the current project, approval of an ethical committee and informed consent of the participants.

If access is restricted, please specify who will be able to access the data and under what conditions.

Access to anonymised data will be open to readers of the journal in which the paper is published.

Access to pseudonymised data after the project will be open to researchers after written permission of the involved KU Leuven researchers of the current project, approval of an ethical committee and informed consent of the participants.

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, Privacy aspects
- Yes, Ethical aspects

Human subject data will be analysed, which is subject to privacy and ethical aspects. The Y-chromosome and mitochondrial DNA markers that will be investigated are completely neutral (non-coding regions) and contain no medical information. In addition, the Y-chromosome is passed on quite intact from father to son, which typifies not one individual, but a paternal line. Mitochondrial DNA is transmitted in a conserved manner from mother to child, which typifies not one individual, but a maternal line. By pseudonymisation, the DNA profile cannot be linked to one person. Of note, only pseudonymised data will be shared upon request and and after written permission of the involved KU Leuven researchers of the current project, approval of an ethical committee and informed consent of the participants.

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

Part of the data will be published anonymously in scientific journals and Open Access databases.

Pseudonymised data will be made available for or shared with researchers upon request and after written permission of the involved KU Leuven researchers of the current project, approval of an ethical committee and informed consent of the participants.

The data will be archived on the KU Leuven repository Lirias.

When will the data be made available?

The anonymised/pseudonymised data will be made available upon publication of the research results at the end of the project.

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

A Creative Commons Attribution license (CC-BY) will be provided as data usage license.

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

To add a dataset to Lirias, the dataset should have received a persistent identifier, such as a DOI or Handle.

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

The data will be published in Open Access journals using the provided FWO bench fee, which costs about €2,000 per paper.

6. Responsibilities

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

Heleen Coreelman will be responsible for data documentation and metadata during the research project, under supervision of Ellen Decaestecker, Ronny Decorte and Sofie Claerhout.

Who will manage data storage and backup during the research project?

Heleen Coreelman will be responsible for data storage and backup during the research project, under supervision of Ellen Decaestecker, Ronny Decorte and Sofie Claerhout.

Who will manage data preservation and sharing?

Ellen Decaestecker, Ronny Decorte and Sofie Claerhout will be responsible for data preservation and sharing.

Who will update and implement this DMP?

Heleen Coreelman will be responsible for updating and implementing this DMP, under supervision of Ellen Decaestecker, Ronny Decorte en Sofie Claerhout. Ellen Decaestecker, Ronny Decorte and Sofie Claerhout will take the end responsibility.