DMP_SofieClaerhout_1265622N

Project Name FWO junior postdoc - DMP - DMP_SofieClaerhout_1265622N

Grant Title 1265622N

Principal Investigator / Researcher Sofie Claerhout

Description The male-specific and non-recombining Y-chromosome (chrY) serves as a powerful tool for genetic-genealogy. But, as a chrY-specific sequencing panel did not exist, it has not been linked to a genealogical database yet. During my PhD, I developed the first chrY-specific sequencing panel, the CSYseq, providing a large dataset of 15,813 interesting Y-markers. This CSYseq is useful to determine evolutionary biogeographical ancestry and to identify close paternal kinships. In this postdoctoral project, I aim to reveal the genetics behind historical socio-demography in Flanders through chrY analysis. It is my ambition to provide novel and valuable perspectives across three disciplines: historical demography, evolutionary biology and forensic genetics. I will sequence 450 males linked to the unique socio-demographical COR* database, which is representative for the Flemish population. Through this interdisciplinary genetic-genealogical approach, I will (a) explore surname founder history to gather more perceptions on surname origins, (b) unravel the influence of chrY and demographic variability on offspring gender ratio, and (c) determine human chrY diversity to estimate familial searching success rates. The key deliverable is a genetic-genealogical database, linking a universally exchangeable chrY dataset with extensive socio-demographic data. Overall, this pioneer study will increase the importance of the highly underused chrY for interdisciplinary population research worldwide.

Institution KU Leuven

1. General Information

Name applicant

Sofie Claerhout

FWO Project Number & Title

1265622N

The genetics behind historical socio-demography in Flanders revealed by the human Y-chromosome: an interdisciplinary approach

Affiliation

- KU Leuven
- Other

Currently also affiliated as long term research visitor in URadboud, Nijmegen, the Netherlands

2. Data description

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

- Generate new data
- Reuse existing data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Type of data	Format	Volume	How created
1. MPS (massive parallel sequencing) dataset	.fastq	150 GB (857 x 300 bp per library)	Experimental: Y-chromosomal DNA will be sequenced with our CSYseq panel on MiSeq (2x300). 857 x300bp fragments on the non-coding regions of the Y-chromosome.
2. Genealogical data	csv, xlsx, accdb	max 2 GB 400 participants, 50 surname groups, 50 spreadsheets, 1 database	Recent genealogical information is digitally collected through Google Forms (https://forms.gle/izFktLLQuVdBKKro9), extension to their patrilineage is done through websites: Geneanet.be and search.arch.be
3. COR database	accdb, xlsx	max 1 GB, +2.000 rows	Digital: the existing Antwerp COR* database (2020, https://hlcs.nl/article/view/9301) is shared by prof. Koen Matthys (Center Sociological research, KU Leuven). This database will be expanded, including all obtained COR-ancestors and paternal lineages to the existing database
4. Physical samples	Buccal swabs	Per participant 2 swabs will be extracted towards 500 µl	DNA buccal swabs are obtained through DNA-kit which is send to the participants address and resend to the Interdisciplinary Research Facility of KULAK. A DNA-kit consists of 2 buccal Helix swabs, informed consent, information letter and a guideline with pictures to explain how they should swab.

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- Yes

Privacy Registry Reference: G-2021-3216

Short description of the kind of personal data that will be used:

Genetic data: non-coding Y-chromosomale regions 202 Y-STRs and 15.000 Y-SNPs; no medical or phenotypical information

Genealogical data: name, occupation, date and place of birth/wedding/death, number of sons/daughters of ancestors in paternal lineage.

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

Privacy Registry Reference: G-2021-3216

S-number: S65250

Human DNA samples and genealogies > Pseudonymization and secure storage

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- No

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

4. Documentation and metadata

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

1. Sequencing: (a) protocols: Logbooks, protocols and laboratory notes are kept and written in detail, in order to repeat every step in the lab. Bioinformatics pipeline is documented for every analysing step in a WORD document. (b) files: all raw fastq files as well as edited files or tool outputs (FASTQC, BAM, SAM, BWA, IGV, FDStools, Yleaf) are generated and kept in a secure folder. A clear description file will also be available to understand this data and the process behind the data. A manual to handle the different tools (SAMtools, IGV, FASTQC, FDStools, CSY.analYser, Yleaf postions file, etc.) is provided.

2. Genealogical data: all paternal lineages with their detailed genealogical data are visualised into pedigrees in one Excel file (per surname group a different spreadsheet). All data is collected in one overview Excel file. This will be accompanied with a how-to-description file for all future researchers to understand the file in detail.

3. COR: The extended COR access database will be saved as an Excel database and an Access database. This is part of our interdisciplinary collaboration with Center of Sociological Research, CeSO, KU Leuven (prof. Koen Matthijs and prof. Paul Puschmann, Uradboud, Nijmegen). A manual to understand all different tables, characters and explanation was already provided by the CeSO, but will be updated in case some new features or data is included in the database.

4. Samples: Samples will be collected through DNA-kits send to the participants. Details on the DNA-kit will be provided. Picture guidelines for sampling, information letters and informed consent process will be documented in detail. Both receiving the DNA kits from the participants as well as pseudonymizing the DNA samples is documented in detail in lab book as well as in one overview WORD file.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

Metadata will be created at the first sheet in all overview Excel files using a detailed checklist per sheet. Abbreviations, tabs, columns will be fully described. Readme files will be created for all programs used and an overview WORD file will list all tools and files to have one major metadata standard in order to make it easier to find and reuse all my data by colleagues in the future. These will be shared with students, colleagues and collaborators to ensure all metadata is collected in a structured way for them to understand all input and output files and reperform all analysis.

5. Data storage and backup during the FWO project

Where will the data be stored?

1. Our MPS data and output variant calling data will be saved and stored on our desktop file storage, personal and shared UZ/KU Leuven network folders and 2 hard drives (raw sequencing data) of our research group.

2. The genealogical data will only be available to the PI (Sofie Claerhout) and will be stored on her desktop file storage as well as personal KU Leuven network folder during the project. Afther the project this will also be available for the promotors and will be stored on the central servers of UZLeuven.

3. COR database will be stored on the desktop folder of the PI (Sofie Claerhout) and will be shared with our collaborating partners (prof. Koen Mathijs, prof. Paul Puschmann) via secured OneDrive weblink to make co-editing easier.

4. Participants samples and DNA extracts will be stored in the biobank (Biobank application form: BB-GEN002-FO03)

How is backup of the data provided?

The data in our database will be protected with a password and saved and stored on the central servers of UZLeuven. This has automatic back-up procedures. Besides, data will be stored on two

hard drives of our research group (to have all the data in duplo).

Non-digital informed consents will be scanned and saved as a PDF in a folder on the central UZ server.

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

Central Server storage and two hard drives (to have back up) is sufficient to store our large volume storage of MPS.

Other small digital files are stored on our personal networkfolder (UZLeuven).

Weekly secured back up on the desktop file storage of the PI (Sofie Claerhout)

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

None. The two hard drives are already available in our research group.

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

All input and output data will be stored on the central servers of UZ Leuven and kept in a database which is protected with a login and password.

The paper version of all informed consents will be destroyed after digitilizing into one PDF file. The online document will be archived and back-ups will be made on the servers and hard drives.

All non-digitilized documents (e.g. lab notes) will be kept in a closed room with batch control that only persons from the Forensic Medicine service (UZLeuven) have access to.

Samples will be stored in the UZ/KU Leuven biobank.

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

DNA samples will be conserved for 10 years in the UZ/KU Leuven biobank (unless applicant requested to destroy DNA on informed consent). Data will be archived on UZ/KU Leuven network drives, KU Leuven repository Lirias and hard drives of the group. Part of the data will be published in data journals and open access databases.

Where will the data be archived (= stored for the longer term)?

1. Data will be archived on UZ/KU Leuven network drives, KU Leuven repository Lirias and hard drives of the group. Part of the data will be published in data journals and open access databases.

2. The genealogical data will be archived on UZ/KU Leuven network drives and hard drives of the group. Part of the data will be published (anonymously) in international journals and open access databases.

3. COR database will be archived on UZ/KU Leuven network drives, KU Leuven repository Lirias and hard drives of the group. Part of the data will be published in international journals and open access databases.

4. Samples will be conserved for 10 years in the UZ/KU Leuven biobank (unless applicant requested to destroy DNA). (Biobank application form: BB-GEN002-FO03)

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

No costs.

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- No
- Yes. Specify:

The personal data of participants might only be usable within the project because the informed consent does not specifically allow this for other projects. Nevertheless, sharing personal genealogical data is possible with the genealogical COR* database, but this is included as part of the project.

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

Only the anonymized sequencing Y-STR reads and Y-SNP results used for data analysis when published in peer-reviewed journals. Its availability is possible after approval of the ethics committee.

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

- In an Open Access repository

Depends on the previous point.

When will the data be made available?

- Upon publication of the research results

The anonimyzed sequencing data will be available after open access publication. Other data (e.g. pseudonymised characteristics (such as occupation, age, gender), surname data and genealogical data) will be available after finishing the project, see further.

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

Access to anonymized data will be open to readers of the journal where the paper is published.

Access to pseudonymized data will be open to researchers after contacting my promotors and approval of ethical commission.

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

Through my FWO bench fee, I will have the opportunity to publish in three open access journals (€7,000).

8. Responsibilities

Who will be responsible for data documentation & metadata?

Dr. Sofie Claerhout and promotors prof. Ronny Decorte (Forensic Genetics, KU Leuven), prof. Jan van Bavel (Sociology, KU Leuven) and prof. Ellen Decaestecker (Evolutionary Biology, KULAK)

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

Dr. Sofie Claerhout and promotors prof. Ronny Decorte (Forensic Genetics, KU Leuven), prof. Jan van Bavel (Sociology, KU Leuven) and prof. Ellen Decaestecker (Evolutionary Biology, KULAK)

Who will be responsible for ensuring data preservation and reuse ?

The promotors prof. Ronny Decorte (Forensic Genetics, KU Leuven), prof. Jan van Bavel (Sociology, KU Leuven) and prof. Ellen Decaestecker (Evolutionary Biology, KULAK)

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

The PI bears the end responsibility of updating & implementing this DMP.