Data management plan

Project Name
Functional annotation of human cold adaptation candidate genes in Drosophila
Project Identifier G050822N

Principal Investigator / Researcher Toomas Kivisild

Description: This project will study the function of human genes detected in scans of selection for cold in Drosophila model. The genetic basis and the molecular mechanisms that govern human adaptation to high latitude environments, including adaptations to diet and cold, remain incompletely understood. We previously identified genomic regions with evidence of positive selection in human populations living in the extreme cold conditions in Central and Northeast Siberia. Many of these genomic regions include genes of unknown function while showing a significant enrichment of liver-expression consistent with the proposed role of the liver as an important metabolic hub. We will use Drosophila melanogaster as a genetic model for the functional annotation of poorly annotated genes associated with cold adaptation in humans to gain a better understanding of the genetic basis and the molecular mechanisms behind human cold adaptation.

Institution KU Leuven

1. General Information

Name of the project lead (PI): Toomas Kivisild
Co-PI: Patrick Callaerts

FWO Project number & title:
G050822N
Functional annotation of human cold adaptation candidate genes in Drosophila

2. Data description

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

- Generate new data in Drosophila model
- Reuse existing human genetic data

2.2. You have set a number of objectives in your proposal. Describe the origin, type and format of the data and its (estimated) volume that will be used to obtain each objective, preferably per objective. You might consider using the table in the guidance.

We have previously identified genomic regions with evidence of positive selection in human populations living in the extreme cold conditions in Central and Northeast Siberia. Many of these genomic regions, available in **tabulated form of human and Drosophila gene names and genomic region coordinates**, include genes of unknown function while showing a significant enrichment of liver-expression consistent with the proposed role of the liver as an important metabolic hub. As **the main objective of this project** we will use Drosophila melanogaster as a genetic model for the functional annotation of these regions.
Secondly, as a related objective of the project, additional **scans of selection** will be performed on **genome-scale data from present-day populations** using various publicly available sources, such as 1000 Genome Project, HGDP, SGDP, and EGDP data.

3. Types, format and scale of data:

Genome-scale sequence and genotype data will be stored in the following format:

- Variations data: .vcf(.gz), .bcf, plink files
- Genomic region files: .bed format
- Annotated results of selection scans: .xlsx format

Other types of derived and compiled data: research documentation, manuscripts, algorithms and bioinformatics scripts will be stored in the following formats:

- Text files: Plain text data (Unicode, .txt), MS Word (.doc/.docx), Adobe Portable Document Format (.pdf), LaTex (.tex) format;
- Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx);
- digital photographs of archaeological samples: .bmp, .jpeg
- Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf), .gif; .png
- Digital images in vector formats: scalable vector graphics (.svg), encapsulated postscript (.eps), Scalable Vector Graphics (.svg), Adobe Illustrator (.ai);

4. Ethical and legal issues

4.1. 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 Privacy & ethical review tool (PRET). Be aware that registering the fact that you process personal data is a legal obligation.

Yes. Secondary use of pseudonymized (coded) genomic data from present-day populations will be stored and used for performing tests of natural selection. Data will be stored and managed according to the GDPR requirements. The study has been registered and submitted to the PRET tool for approval (G-2022-4971). Approval for the use of genomic information from present-day sources will also be sought from the Research Ethics committee of UZ/KU Leuven.

4.2. 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, work on publicly available human genetic data is planned along with experimental work on Drosophila. Ethical approval will be sought for the use of the publicly available data.

4.3. 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?

Yes. If there is substantial potential, the invention will be thoroughly assessed, and will be IP protected (mostly patent protection or copyright protection).

4.4. 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 restricting the use of the data.

5. Documentation and metadata

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

Documentation will consist of notes and log files that refer to specific datasets. Those notes will describe the samples used, research methods for data generation, experimental setup and protocols for genotyping, sequencing, the links to the specific computer file location as well as the names of the respective datasets. We also maintain a metadata sheet with the connection between gene annotations in Drosophila and selection target human genes, so that relevant information can be properly linked between two main work packages.

Algorithms, scripts and pipelines will be documented locally on the server and when finalized, they will be additionally described in either manuscripts or open public domains, e.g. github.

In case of the secondary use of genotype and sequence data from individuals from the present-day population all information provided to the study team will be previously coded and no identifying keys will be provided to the team. The data will be stored in password and key protected VSC servers and operations with the data will be documented either locally on the server on in the password protected computers of the team members.

5.2. 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.

Selection scan output data will be stored in form of .bed files that annotate the chromosome, start and end position of targeted regions and their genic content.

6. Data storage and backup during the FWO project

6.1. Where will the data be stored?

- Sequence and genotype data: on KU Leuven servers, on the Flemish Supercomputer Centre (VSC), or on secured cloud-based platforms
- Algorithms, scripts and software codes: in private online git repositories of the PIs. After publication, the repository will be changed to a public repository.
- Other data files (digital images, etc..): local KU Leuven servers and PI computers

6.2. How will the data be backed up?

- Project team will be encouraged to store their data on secured KUL/UZ drives. Data temporarily stored on the personal computers will be copied to a backup folder on the KUL drive on a weekly basis. KU Leuven drives are automatically backed-up at least daily.
- All omics data stored on the VSC will be transferred on a weekly basis to the archive area which is mirrored.

6.3. 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.
There is sufficient storage and back-up capacity on all KU Leuven servers and on the VSC to cover the project needs.

6.4. What are the expected costs for data storage and backup during the project? How will these costs be covered?
The total cost of active data storage for 5 years of the project plus archive storage for 10 years after the project has been budgeted at 5,000 EUR.

6.5. Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?
All genotype and sequence data will be stored and analyzed on the VSC storage folders the access to which is password and security key protected and only granted to the project team members with VSC accounts.

6.6. **Physical sample storage.**
There are no physical human samples planned to be processed as part of this project.

7. Data preservation after the end of the project

7.1. 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 minimum preservation term of 10 years after the end of the project will be applied to all datasets.

7.2. Where will these data be archived (= stored for the long term)?
Project data, developed algorithms and software will be stored on VSC archive, as well on public repositories such as Github.com, and will be maintained or updated by the (co-)promotor for at least 10 years.

7.3. What are the expected costs for data preservation during these 10 years? How will the costs be covered?
The total estimated cost of data storage during the 10 years after the end of the project is 1,750 EUR, as budgeted in the FWO project.

7.4. **Long-term storage of the physical samples.**
No human physical samples will be stored long-term.

8. Data sharing and re-use

8.1. 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 factors restricting the sharing of project data can be identified.

8.2. Which data will be made available after the end of the project?
The following data will be made available at the end of the project:

- New genetic data from Drosophila
- Research protocols, algorithms, scripts and software
- Manuscripts

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

- In an Open Access repository

- Upon publication, relevant information about the samples supporting a manuscript will be made publicly available as supplemental information.
- All protocols used to generate published data will be described in the corresponding manuscripts. These data and all other documents are accessible to the PI and the research staff, and will be made available upon request.
- Algorithms, scripts and software: github.
- Manuscripts: All scientific publications will be shared openly. Publications will also be automatically added to Lirias 2.0 (the metadata will be added, not the full manuscripts).

8.4. When will the data be made available?
Upon publication of the research results.

8.5. Who will be able to access the data and under what conditions?
Whenever possible, datasets and the appropriate metadata will be made publicly available. A CC-BY license will be opted for when possible. In case of other sources of data that the project gains access to

for the secondary use references to the source (URL where relevant) will be provided in the project publications.

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

It is the intention to minimize data management costs by using free-to-use data repositories and dissemination facilities whenever possible. Data management costs will be covered by the laboratory budget.

9. Responsibilities

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

Metadata will be documented by the research and technical staff at the time of data collection and analysis.

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

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

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

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

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

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