Plan Overview

A Data Management Plan created using DMPonline

Title: Application of a Novel Telomere-to-Telomere Reference Genome in a cross-sectional Swedish Cohort

Creator: Daniel Schmitz

Affiliation: Uppsala University

Template: Uppsala University - data management plan

ID: 111231

Last modified: 15-11-2022

Application of a Novel Telomere-to-Telomere Reference Genome in a cross-sectional Swedish Cohort

General information

Project title

Application of a Novel Telomere-to-Telomere Reference Genome in a cross-sectional Swedish Cohort

Project leader

Åsa Johansson

Project identifier

N/A

Version and date

v1.0, 2022-11-14

Description of data – reuse of existing data and/or production of new data

How will data be collected, created or reused?

The main dataset is sequencing data from the SweGen project, which has already been collected and is available as a resource on Bianca. Additional data used for this project is publicly available:

- The T2T-CHM13 reference
- Annotation databases for variant annotation

The generated data, namely alignments, variant calls and statistics of these, will be stored on Bianca. The results will mainly be calculated using the Sarek workflow, which includes quality control and checks for consistency.

What types of data will be created and/or collected, in terms of data format and amount/volume of data?

Existing sequencing data from the SweGen study (FASTQ files) will be used. We will request access to the data from UPPMAX.

We will generate alignments (BAM/CRAM), variant calls (VCF) and annotations (tabular) for all 1000 individuals in SweGen, totaling around 100 TB of data.

The data will be reproducible through the use of existing pipelines. All additional code used for downstream analyses will be saved and documented.

Documentation and data quality

How will the material be documented and described, with associated metadata relating to structure, standards and format for descriptions of the content, collection method, etc.?

The material will be described in a README file on Bianca. Public code and results will be documented in the place of their publication.

How will data quality be safeguarded and documented (for example repeated measurements, validation of data input, etc.)?

There are no measurements involved. Sarek and Nextflow include provisions to protect data integrity. Downstream analyses will mostlye be performed using established libraries.

Storage and backup

How is storage and backup of data and metadata safeguarded during the research process?

All data will be stored on the Bianca cluster at UPPMAX. SweGen is available as a shared resource on Bianca. Access is restricted through the SNIC-SENS project that will be used for the computations.

How is data security and controlled access to data safeguarded, in relation to the handling of sensitive data and personal data, for example?

All data will be stored on Bianca, which is specifically designed for storage and analysis of sensitive information.

Legal and ethical aspects

How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?

All sensitive information is kept on Bianca which prohibits unauthorized access. There anre no intellectual property rights to consider.

How is correct data handling according to ethical aspects safeguarded?

All data except for population-level statistics will remain on Bianca and not be shared.

Accessibility and long-term storage

How, when and where will research data or information about data (metadata) be made accessible?
Are there any conditions, embargoes and limitations on the access to and reuse of data to be considered?

Population-level results will be made available on public repositories and/or returned to SweGen for inclusion in their database. Custom code will be made available publicly. Individual-level data cannot be shared due to privacy reasons.

In what way is long-term storage safeguarded, and by whom?
How will the selection of data for long-term storage be made?

All data will be saved in established formats or plain-text files. All data will be stored on Bianca, which is managed by UPPMAX and regularly backed up.

Will specific systems, software, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?

No specific system is necessary for the published data. Mappings will require the use of software that supports CRAM files.

How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?

A data repository that supports DOIs will be considered.

Responsibility and resources

Who is responsible for data management and (possibly) supports the work with this while the research project is in progress?
Who is responsible for data management, ongoing management and long-term storage after the research project has ended?

During project: Daniel Schmitz

After the project has concluded, the data will be returned to SweGen.

What resources (costs, labour input or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)? What resources will be needed to ensure that data fulfil the FAIR principles?

No resources needed.