

Data Management Plan

yvonne_kallberg_test_DMP

Contributors : **Sam Smith** (sam.smith@example.com) Uppsala universitet (PI)

Yvonne Kallberg (yvonne.kallberg@nbis.se) Stockholms universitet (researcher, **contact person**)

Based on: *Intro DMP course module, 1.1.0 (SciLifeLab:intro-DMP:1.1.0)*

Project phase: *Before Submitting the Proposal*

Created by: **Yvonne Kallberg** (yvonne.kallberg@nbis.se)

Generated on: *15 Sep 2021*

Data Management Plan created in Data Stewardship Wizard «<https://dsw.scilifelab.se>»

Projects

We will be working on the following projects and for those are the data and work described in this DMP.

VEGFR2 Y949F mutation

RNA sequencing of lung tissue from transgenic mice in order to investigate the effect of a single tyrosine to phenylalanine exchange in the endothelial receptor VEGFR2 at position Y949.

Start date: 2021-10-12

End date: 2022-05-31

Funding: [Vetenskapsrådet](#): *grant number not yet given (applied)*

Section A: Description of data – reuse of existing data and/or production of new data

1. How will data be collected, created or reused?

Instrument datasets

The following instrument datasets will be acquired in the project:

- **RNA sequences**

This dataset will be collected by National Genomics Infrastructure (NGI).

For this dataset, we are using the following instruments:

- [Illumina HiSeq 2500](#)

The equipment is very well described and known.

Re-used datasets

We will use the following reference datasets:

- **GRCm38 (mm10)**

(https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.20/)

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

Data formats and types

We will be using the following data formats and types:

- [FASTQ Sequence and Sequence Quality Format](#) - We will have only a small amount of data stored in this format.

Section B: Documentation and data quality

3. How will the material be documented and described?

We will use an electronic lab notebook to make sure that there is good provenance of the data analysis.

We will use the following metadata standards to describe our datasets:

- [Minimal Information about a high throughput SEQuencing Experiment](#).

In addition, the following metadata will be provided as documentation: File organisation and naming conventions will be documented in a README text file that will be put in the root folder of the project. Also, a dictionary of the sample metadata will be created.

4. How will data quality be safeguarded?

Instrument datasets

- **RNA sequences**

We will be using the following quality processes:

- calibrating measurements
- repeat samples/measurements

Section C: Storage and Backup

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

Data will be stored at SNIC center Uppmax. An application for project time will be made.

6. How is data security and controlled access to data safeguarded?

Section D: Legal and ethical aspects

7. How is data handling according to legal requirements safeguarded?

We will be working with the philosophy *as open as possible* for our data.

The data cannot become completely open immediately because we want to

publish a paper first.

Data that is not legally restrained will be released after a fixed time period, unconditionally.

For the reference and non-reference data sets that we reuse, conditions are as follows:

- **GRCm38 (mm10)** – freely available for any use (public domain or CC0).

8. How is correct data handling according to ethical aspects safeguarded

Data we collect

None of the collected data sets are personal.

Section E: Accessibility and long-term storage

9. How, when and where will research data or information about data (metadata) be made accessible?

The datasets will be available as follows:

- **RNAseq**

Mouse RNA sequences.

- Open (shared with anyone) using a domain-specific repository: [European Nucleotide Archive](#).

Restrictions on data sharing

Ethical and legal restrictions are documented under Section D.

Data cannot be completely open due to legal reasons. But data that is not legally restrained will be released after a fixed time period. The embargo period will be 1 year.

10. In what way is long-term storage safeguarded, and by whom?

We plan to produce the following datasets:

- RNAseq (published):

11. 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?

- RNAseq (published)

The distributions will be stored in:

- Domain-specific repository: [European Nucleotide Archive](#).

12. How will the use of unique and persistent identifiers be safeguarded?

The published datasets will have the following persistent identifiers:

- RNAseq

Section F: Responsibilities and Resources

13. Who will be responsible for data management?

Yvonne Kallberg is responsible for overall data management including implementing the DMP, and ensuring it is reviewed and revised.

14. What resources will be required for data management?

To execute the DMP, additional specialist expertise is required which will be accomplished by training existing staff.