Plan Overview

A Data Management Plan created using DeiC DMP

Title: Non-sensitive NGS research project template

Creator: Jose Alejandro Romero Herrera

Data Manager: Jose Alejandro Romero Herrera, Sarah Lundregan

Affiliation: Københavns Universitet / University of Copenhagen

Template: Horizon Europe

Project abstract:

This is a DMP template for projects that uses non-sensitive NGS datasets.

This template was made for the Brickman lab at the Center for Stem Cell Medicine (reNEW) at SUND faculty. Nonetheless, everyone is welcome to use it as they see fit.

This template has already prefilled repetitive information regarding RDM practices being implemented and carried out at the Brickman lab. For more information, please visit the Brickman lab Github page.

In the template there will be sections where you will have to fill specific information about your project, samples, metadata, etc. Any text written between <> is a comment or optional text. There will be comments marked with (X), where X is a number. Both should be deleted when you write the DMP.

NOTE: This template is not a replacement for the thought process of understanding what type of data do you have, but to streamline the writing of your DMP.

ID: 5751

Last modified: 13-07-2023

Copyright information:

Non-sensitive NGS research project template

1. Data Summary

Will you re-use any existing data and what will you re-use it for?
State the reasons if re-use of any existing data has been considered but discarded.

<Use either A or B.>
A. The project will reuse several published Next Generation Sequencing datasets in order to benchmark or compare results with our own experiments:
<List datasets (1)>
B. The project will not reuse any existing data.
<No data has been considered but discarded.>

What types and formats of data will the project generate or re-use?

The project will generate <and reuse> Next Generation Sequencing (NGS) quantitative data. The NGS data will consist on unprocessed raw sequencing reads in fastq format and metadata regarding the experiment and sample information in csv format.
The data will be processed using community curated bioinformatic pipelines, such as those from the nf-core community. We will provide pipeline documentations for reproducibility purposes.
In addition, the pipeline will generate different types of intermediate data, which will not be made available. However, we will provide with final results, such as read count matrices in tsv format and QC metrics reports in html format.
Laboratory protocols and Electronic Laboratory Notebooks (ELNs) used to generate the data will be created and managed by the Labguru software. Cell lines, organism models, antibodies, and plasmids used in this project will be identified by their RRID identifier if available.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The purpose of the generated data is to test the hypothesis of the project <state hypothesis here>.
Reused data will be used to compare and benchmark results with our own.

What is the expected size of the data that you intend to generate or re-use?

Depending on the number of the data reused and generated, it might lay anywhere between 50Gb to 500Gb of data (1).

What is the origin/provenance of the data, either generated or re-used?

(1) Biological samples will be generated at the Center for Stem Cell Medicine (reNEW). Laboratory protocols and Electronic Laboratory Notebooks (ELNs) will be managed by the Labguru software. Generated sequencing data will be produced by the reNEW Genomics Platform.
<Reused raw NGS data will be gathered from public and trusted repositories. Since we will preprocess the data, we will be able to assess its quality by using bioinformatic pipelines curated by the research community (nf-core).>

To whom might your data be useful ('data utility'), outside your project?

Other researchers interested in NGS datasets or researchers within the field of study of this project.
<(1)Add here somehting about the big picture of your project. How can the data be used in the future to create innovation / enhance science / save the world, etc? For example:
This research holds significant potential for the future and for humanity. By studying dynamic mechanisms in cell development, particularly in the context of embryonic stem (ES) cells, it offers valuable insights into lineage choice and the factors that influence cell fate determination. Understanding these processes can pave the way for directed differentiation of ES cells into specific functional cell types, such as liver, lung, thyroid, thymus, and pancreas cells. This knowledge can have far-reaching implications, including advancements in regenerative medicine, tissue engineering, and the development of novel therapies for various diseases. Furthermore, it could provide valuable means for studying regulatory networks and transcriptional plasticity, enhancing our understanding of early embryonic development and potentially uncovering new avenues for intervention and manipulation of cellular processes.>

2. FAIR data

2.1. Making data findable, including provisions for metadata:
Will data be identified by a persistent identifier?

Yes. raw NGS datasets, as well as metadata regarding how the samples were created, will be uploaded to the Gene Expression Omnibus (GEO) repository. GEO will provide a unique persistent identifier to our submission.
Data analysis scripts and results, as well as the pipeline used to preprocess the data, will be version controlled as a Github repository and archived in Zenodo. Zenodo will also provide a unique persistent identifier to github repository.
Cell lines, organism models, antibodies, and plasmids generated in this project will be registered at the Research Resource Identifiers and given a RRID identifier.

2.1. Making data findable, including provisions for metadata:
Will rich metadata be provided to allow discovery?
What metadata will be created?
What disciplinary or general standards will be followed?
In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

For raw NGS data archived in GEO:
Rich metadata will be provided as requested on the GEO repository. Such as: keywords, submission responsible, publication title and DOI, dataset description (including generation).
Although GEO does not strictly use a metadata standard, it does encourage the use of NGS data standards and it is compliant to them. Thus, we will follow the MINSEQE standard(1), as recommended by GEO.

For data analysis and results archived in Zenodo:
Rich metadata will be provided such as: keywords, author, publication title and DOI, original dataset (GEO ID). A description file will be provided as well specifying the contents of the archive.

2.1. Making data findable, including provisions for metadata:
Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

NGS, <NGS type, like RNAseq>, <study field keywords>, <Organism>, <Cell line>, <Cell type>, <protein/gene of interest>.

2.1. Making data findable, including provisions for metadata:
Will metadata be offered in such a way that it can be harvested and indexed?

For raw NGS data archived in GEO:
Yes, GEO provides the possibility to access and fetch data and metadata. Datasets uploaded to GEO are indexed and loaded into GEO Profiles and GEO DataSets, which allows users to query gene names, visualize charts and clusters, and more"
(https://www.ncbi.nlm.nih.gov/geo/info/faq.html#retrievals).
The metadata is provided in different txt formats. We will adhere to the MINSEQE standard.
For data analysis and results archived in Zenodo:
Zenodo policies specifies that metadata is licensed under CC0, except for email addresses. All metadata is exported via OAI-PMH and can be harvested.

2.2. Making data accessible - Repository:
Will the data be deposited in a trusted repository?

For raw NGS data and its metadata archived in GEO:
Yes, the data will be deposited in the Gene Expression Omnibus (GEO), which is supported by the USA National Center for Biotechnology Information (NCBI) .

For data analysis and results archived in Zenodo:
Zenodo is a trusted repository supported by CERN (European Organization for Nuclear Research) and is part of the European Open Science Cloud (EOSC).

2.2. Making data accessible - Repository:
Have you explored appropriate arrangements with the identified repository where your data will be deposited?

For raw NGS data and its metadata archived in GEO:
GEO only requires to create a free account. Data can be deposited and embargoed until a publication is published/end of project. Data deposited in GEO cannot be made private and will have open access.

For data analysis and results archived in Zenodo:
Zenodo also only requires to create a free account. Data can be deposited and embargoed until a publication is published/end of project. Data will be available freely afterwards.

2.2. Making data accessible - Repository:
Does the repository ensure that the data is assigned an identifier?
Will the repository resolve the identifier to a digital object?

For raw NGS data and its metadata archived in GEO:
Yes, GEO provides an accession number that is unique to the deposited data.

For data analysis and results archived in Zenodo:
Zenodo will provide a DOI to the archived data.

2.2. Making data accessible - Data:
Will all data be made openly available?

If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

For raw NGS data and its metadata archived in GEO:
Datasets will be made openly available in the GEO repository (public domain).

For data analysis and results archived in Zenodo:
Datasets will be made openly available under CC-BY license.

2.2. Making data accessible - Data:
If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

All datasest will be embargoed until publication is accepted in a journal/end of the project.
Release will happen as soon as the publication is published. Datasets not used for publications will be released as soon as the project ends.

2.2. Making data accessible - Data:
Will the data be accessible through a free and standardized access protocol?

For raw NGS data and its metadata archived in GEO:
Yes, GEO offers API services. You can also manually download each dataset. Standard download is possible via http and ftp protocols.

For data analysis and results archived in Zenodo:
Access to metadata and data files is provided over standard protocols such as HTTP and OAI-PMH.

2.2. Making data accessible - Data:
If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

Data will be restricted to collaborators before publication/end of the project.
Since the data generated is non-sensitive, all data will be made available public upon publication/end of the project.

2.2. Making data accessible - Data:
How will the identity of the person accessing the data be ascertained?

Non-applicable.

2.2. Making data accessible - Data:
Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

No, the project does not use or reuse personal/sensitive data.

2.2. Making data accessible - Metadata:
Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement?
If not, please clarify why.
Will metadata contain information to enable the user to access the data?

Yes, the metadata will be accesible as open access, with no conditions.
Data and metada deposited in GEO is open data with no restrictions (public domain). Zenodo's policies state that all metadata is licensed under CC0, except for email addresses.

2.2. Making data accessible - Metadata:
How long will the data remain available and findable?
Will metadata be guaranteed to remain available after data is no longer available?

For raw NGS data and its metadata archived in GEO:
GEO promises long-term archiving of the data, but does not specify for how long.

For data analysis and results archived in Zenodo:
Zenodo's policies explains that the data will be retained at least for the next 20 years.

2.2. Making data accessible - Metadata:
Will documentation or reference about any software be needed to access or read the data be included?
Will it be possible to include the relevant software (e.g. in open source code)?

Yes, metadata including how to process the reads and the pipeline used to process the reads will be made available as part of the submission to GEO.
In addition, we will provide the code (mainly jupyter notebooks and/or Rmarkdown) used to analyse the data through the Brickman lab Github repository. This includes an execution report in html formtat with all the software and their versions used to preprocess the data. The Github repository will be automatically archived in Zenodo. The Zenodo repository will generate a DOI, which will link the GEO dataset submission and manuscript. The manuscript will also include the DOI of the Zenodo repository in the appropiate "Data Availability" section.
All tools are free to use.

2.3. Making data interoperable:
What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines?
Will you follow community-endorsed interoperability best practices? Which ones?

We will use the community curated nf-core pipelines to process the reads in fastq standard format.
GEO submissions will be compliant with the MINSEQE standard.
We will use Ensembl IDs for Gene IDs <or Transcript IDs> whenever we refer to genes, as well as their gene name. In the case of proteins, we will refer to their uniprot ID.
Organisms used will have the taxonomy vocabulary (e.g. *Mus musculus*) and NCBI ID (e.g. NCBI:txid10090).
Genome reference version will be also provided.

2.3. Making data interoperable:
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?
Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

The project will not generate specific ontologies or vocabularies. We will register, whenever possible, generated plasmids, organism models or cell lines into the RRID database.

2.3. Making data interoperable:
Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/).

(1)
Data analysis and results archived in Zenodo will link to the original Github repositories where the data analysis and results are hosted. The Zenodo metadata will also contain the GEO ID and link to the raw NGS data and its metadata.

2.4. Increase data re-use:
How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning,

analyses, variable definitions, units of measurement, etc.)?

For raw NGS data and its metadata archived in GEO:
As part of the submission to GEO, users need to provide protocol descriptions to recreate the samples and raw data. We will provide GEO with protocols in the form of STAR methods, based on information fetched from the ELNs and lab protocols managed by the Labguru software. Protocols are then available through GEO as part of the samples' metadata.

For data analysis and results archived in Zenodo:
We will, in addition, provide information on how to preprocess the data using the community curated workflows (nf-core). Furthermore, downstream data analyses will be deposited to a Github repository, and link it to Zenodo as to provide a DOI for the Github repository.

2.4. Increase data re-use:
Will your data be made freely available in the public domain to permit the widest re-use possible?
Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes:
Data and metada deposited in GEO is open data with no restrictions (public domain).
Data analysis and results deposited in a Github repository and archived in Zenodo will all be under CC BY 4.0 license.

2.4. Increase data re-use:
Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes, all data will be made usable after the end of the project.

2.4. Increase data re-use:
Will the provenance of the data be thoroughly documented using the appropriate standards?

Protocols used to generate samples will be provided in the form of STAR methods. This can be fetched from Labguru.
Data preprocessing will be documented throughly and documentation of the nf-core pipeline will be provided directly as part of the pipeline results.
Downstream data analysis will be provided in the form a report in a html file.

2.4. Increase data re-use:
Describe all relevant data quality assurance processes.

Standard QC metrics will be used to ensure the quality of the raw NGS dataset, such as FastQC, and MultiQC tools/reports.
Furthermore, other QC metrics will be used for downstream analysis according to the specific data type (bulk RNAseq, single cell RNAseq, ATACseq/ChIPseq, etc.).

3. Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Software and pipeline used to process the data will be provided in a Github repository, which will be archived in Zenodo and made available to use freely.
Cell lines will be stored in liquid nitrogen (-180 ºC approximately). Organism models will be preserved by freezing sperm (at -20 ºC approximately). Antibodies and reagents will be preserved according to the manufacturer specifications.
Whenever possible, new cell lines, plasmids or organism models will be registered at RRID and provided an identifier.
Protocols used to generate samples will be provided in the form of STAR methods, as part of the GEO submission.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

Protocols used to generate samples will be fetched from Labguru and made available in the form of STAR methods upon publication/end of project as part of the GEO submission and/or the publication. Primers and adapters used to generate samples will be made available as well.
Whenever possible, new cell lines, plasmids or organism models will be registered at RRID.org and provided an identifier.

4. Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

Github, GEO and Zenodo are free to use.
Local storage is provided by the IT department of the University of Copenhagen. This includes storage backup. Access to local storage is controlled and secured by mandatory login and it is open only to collaborators. Cost is covered by the University of Copenhagen.
Labguru costs are covered as well by the University of Copenhagen.
Making sure the output adheres to FAIR principles will require mostly time resources from the researchers.

How will these be covered?

Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).

Costs related to research data and output management are covered by the University of Copenhagen.

Who will be responsible for data management in your project?

The roles are as such (1):

- Collection and quality control of samples: <insert name and email>
- Analysis and documentation: <insert name and email>
- Publication and curation of research data: <insert name and email>
- Preservation and archiving: <insert name and email>

How will long term preservation be ensured?
Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long).

Protocols and ELNs used to generate the data will be managed by Labguru and retained for at least 5 years.
Raw NGS data and its metadata will be preserved in the GEO repository. GEO promises long-term archiving of the data, but does not specify for how long.
Data analysis will be preserved in Zenodo. Zenodo terms of services explains that the data will be retained at least for the next 20 years.
Data file formats follow domain standards and are in open format (csv, txt, pdf, html, fastq, etc).

5. Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Secure local storage and backups will be provided by the IT department of the University of Copenhagen. Data will be either archived in Zenodo or GEO.
This project does not include sensitive data.

Will the data be safely stored in trusted repositories for long term preservation and curation?

Yes, Zenodo and GEO are trusted repositories for long term preservation.

6. Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?
These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Non-applicable for data sharing.
Regarding ethical approval for animal work, all animal work was carried out in accordance with European legislation. All work was authorized by the Danish National Animal Experiments Inspectorate (Dyreforsøgstilsynet, license no. <lab license number for working with animals>) and performed according to national guidelines.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Non-applicable.

7. Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management?
If yes, which ones (please list and briefly describe them)?

Yes. The Brickman lab is currently implementing procotols for RDM management. For more details, see here.
We will also adhere to publisher procedures for submitted manuscripts.
Finally, we will also adhere to the University of Copenhagen RDM policy.