DMP TEMPLATE FOR A GENOMICS STUDY

(Examples were extracted from a fictional study and the template was adapted from the Portage general template - https://dmp-pgd.ca/template_export/878086536.pdf)

This document is a working draft and will continue to undergo piloting and refinement. It represents an ongoing effort to improve data management plans in the biomedical sciences.

The individuals involved in the development of the template include Dr. Anna Catharina V. Armond, Dr. Theodore J. Perkins, Dr. Michael Hoffman, and Dr. Kelly D. Cobey.

1. DATA DESCRIPTION AND COLLECTION

1a. Describe the study for which the data are being collected.

Example:
The GAB study (Gene regulation in Acute lymphoblastic leukaemia and normal Blood) has the goal of understanding differences in gene regulation and identifying biomarkers between normal hematopoiesis and patients with acute lymphoblastic leukaemia (ALL). The study will involve analysis of blood drawn from patients and healthy volunteers, as well as a mouse model.

1b. What types of data will you collect, create, link to, acquire and/or record?

Example:

The GAB study will generate new data on gene expression at the RNA and protein levels in 20 ALL patients (10 female, 10 male) and 20 healthy age- and sex-matched volunteers. Specifically:
- RNA-level expression will be generated by single-cell RNA-sequencing (scRNA-seq)
- Protein-level expression will be generated by tandem mass spectrometry (MS/MS)
- Patient sex, age at diagnosis, and age at blood draw will be recorded
Similar experiments will be conducted in a mouse model of ALL and compared to healthy mice (3 biological replicates of each condition).
Data will further be analyzed in comparison with other public datasets on gene expression from The Cancer Genome Atlas (TCGA).

1c. How will new data be collected or produced and/or how will existing data be re-used?

Example:

Each blood sample (human or mouse) will be split into a fraction to be used to analyze RNA and one to analyze protein levels. For RNA, blood cells will be processed by standard 10xGenomics/Illumina protocols for single-cell RNA sequencing (scRNA-seq). Sequencing will be performed on an Illumina NextSeq 2000. Data will be analyzed by the Seurat pipeline and other in-house scripts. Proteins will be extracted from cells and digested using trypsin and the resulting peptides will be labelled with tandem mass tags (TMT) for multiplexed quantification. The labelled peptides will be analyzed by tandem mass spectrometry (Thermo Scientific Orbitrap Fusion Lumos). The output data will be analyzed using the Trans Proteomic Pipeline for protein identification and quantification, and other in-house scripts. An electronic laboratory notebook system will be used to record which team members perform major steps of: patient consenting, sample collection, sample preparation, sequencing/mass spectrometry, and data analysis. All data generated in this project are new, no existing data will be used.

1d. What file formats will your data be collected in? Will these formats allow for data reuse, sharing, and long-term access to the data?

Example:

Raw scRNA-seq data will be stored in FASTQ format. Raw protein data will be stored in RAW format. Processed RNA and protein data, representing quantitative expression levels per gene and sample/cell will be stored in CSV files. Patient metadata will be stored in a CSV file.

1d. What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?

Example:

A centralized spreadsheet will be used to assign sample IDs to all human and mouse samples. Files pertaining to individual samples, such as raw data files and intermediate stages of processing, will follow a naming convention as follows: sampleID_fileType_version_YYYYMMDD. fileType describes the type of data or stage of analysis, while version number allows for the possibility of reanalysis with new or alternate pipelines. The centralized spreadsheet, as well as summary files (e.g. cross-sample RNA expression, protein expression, figures, etc.), will be version-controlled, and named in the format fileType_version_YYYYMD. In case of any changes to the analysis pipeline itself, the pipeline will also be version-controlled.

1e. How will the research team and other collaborators access, modify, and contribute data throughout the project?

Example:

All data entry and access will be controlled by usernames and passwords. Staff will have access restricted to the functionality and data that are appropriate for their role in the study, as decided by the PI and senior analyst. Generally, to ensure data integrity, raw data files will be deposited but never overwritten, except by password-confirmed action of the PI or senior analyst

2. DOCUMENTATION AND METADATA

2a. What documentation will be needed for the data to be read and interpreted correctly in the future?

Example:

The metadata will include: ethics approval for the study; for human samples, unique patient identifier, sex, age at diagnosis, age at the time of blood draw, and consent forms; for mice, unique animal identifier, sex, date of birth, date of blood draw, and genotype; experimental protocols for sample acquisition and processing; outline of statistical analysis plan; and links to data file locations and analysis pipeline(s).

2b. Describe in detail the data-level descriptions, codes, and definitions.

Example:

We will record data in the following formats for human patients:

1. *Unique patient identifier (character string)*
2. *Sex (Female/Male/Other)*
3. *Age at diagnosis (years, only for ALL patients, -1 for healthy volunteers)*
4. *Age at time of blood draw (years)*
5. *RNA and protein data in formats described above*

We will record data in the following formats for mice:

1. *Unique animal identifier (character string)*
2. *Sex (Female/Male)*
3.. *Date of birth (dd/mm/yyyy)*
4. *Date of blood draw (dd/mm/yyyy)*
5. *Genotype (ALL or healthy)*

6. *RNA and protein data in formats described above*

2c. How will you make sure that documentation is created and captured consistently throughout your project?

Example:

All data will be collected in a documented directory structure, shared and accessible to researchers participating in the project, but with written permission granted only to those depositing data and performing data analysis. The PI will audit relevant electronic laboratory network entries, quarterly, to ensure consistent and correct metadata capture. The senior analyst will monitor the consistency and correctness of genomic data.

2d. If you are using a metadata standard and/or tools to document and describe your data, please list here.

Example:

Whenever possible, data will be standardized to CDISC standards. The generated csv datasets are machine-readable and openly-accessible. Shareable raw sequencing data and proteomics datasets will be produced in standard machine-readable FASTQ and RAW formats respectively. In csv datafiles, genes will be identified by unique ENSEMBL IDs, alongside gene symbols for easier human understanding.

3. STORAGE AND BACKUP

3a. What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?

Example:

The estimated storage-space is 20TB. This includes raw data, intermediate stages of analysis, final results, and all metadata. Raw omics data and any intermediate forms that contain potentially personally-identifying information will be retained for three years after the completion of the study, to allow time for publication and any need for re-analysis. After this, any personally-identifying omics data will be destroyed. All other derived data, which is associated with sample IDs but is not personally identifiable, will be retained indefinitely.

3b. How and where will your data be stored and backed up during your research project?

Example:

Raw data RNA and protein data will be stored in two duplicate, secure locations: one on the PI s computing hardware and one on institutionally provided infrastructure. Large intermediate data files, which can be recreated by analysis pipelines, will be kept only on the PI s computing hardware. All other smaller, summary data files will be stored using the 3-2-1 back up rule, using the institutionally provided infrastructure, the PI s computer, and an external hard drive.

4. PRESERVATION

4a. Where will you deposit your data for long-term preservation and access at the end of your research project?

Example:

When ready to make available, the data will be shared via the Open Science Framework (OSF). Shareable raw scRNA-seq and proteomics data, along with summary gene expression data, will be shared in appropriate formats on the Sequence Read Archive (SRA), Gene Expression Omnibus (GEO), Protein Identifications Database (PRIDE), and ProteomeXchange.

4b. Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.

Example:

FASTQ and RAW files will be retained for raw high-throughput sequencing and mass spectrometry data respectively. Other datasets will be stored in the comma-separated CSV format. After completion of the data collection, data will be checked for inconsistencies, de-identified, and locked.

4C. If your study involves in-house or bespoke analysis pipelines, how will those pipelines be documented and will they be shared so that other researchers can reproduce your results or apply the same analyses to their data?

Example:
FASTQ and RAW files will be retained for raw high-throughput sequencing and mass spectrometry data respectively. Other datasets will be stored in the comma-separated CSV format. After completion of the data collection, data will be checked for inconsistencies, de-identified, and locked.

5. SHARING AND REUSE

5a. What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final, and metadata). Will sharing be open or by restricted access request? If a restricted access request process is used, describe this process.

Example:

Raw high-throughput patient data will not be shared. Raw high-throughput mouse data will be shared openly as described above. De-identified results spreadsheets and metadata will be shared openly, along with documentation describing variables, naming conventions and standards.

5b. What type of end-user license will you use for your data?

Example:

De-identified data will be shared under Creative Commons' CC-BY-4.0 license.

5c. What steps will be taken to help the research community know that your data exists?

Example:

We are committed to the FAIR principles. To make our data findable and accessible, As described above, summary data will be archived and shared via the Open Science Framework. OSF assigns and promotes persistent identifiers - Digital Object Identifiers (DOIs) – to OSF-hosted studies. This initiative promotes academic credit, direct citation and tangible metrics for researchers who collect and share data. DOIs contain metadata that identify and acknowledge the staff and sites that contribute to the dataset, as well as information on location of the study, funders, and publications or documents associated with the data. Raw mouse omics data will be shared in the databases described above, where it can be cited via DOIs or accession numbers. Summary data will also appear as supplementary information in any articles published based on this work. To make the data interoperable, we will share detailed metadata (workflows, vocabularies, processes, and standards) and the data will be shared in preservation friendly formats. To make the data reusable, we will share the de-identified data publicly under Creative Commons CC-BY-4.0 license.

6. RESPONSIBILITIES AND RESOURCES

6a. Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

Example:

The team will include:
Prof. Poly Merase - Principal Investigator – (CONTACT DETAILS)
Dr. Gene E. Us - Co-investigator– (CONTACT DETAILS)
Prof. Nucleo Philer - Senior Analyst – (CONTACT DETAILS)

Sponsor (ADD ROLES, IF ANY)

Principal investigator will be responsible for:
(i) Agreement of the final Protocol and the Data Analysis Plans;
(ii) Reviewing progress of the study and, if necessary, deciding on Protocol changes;
(iii) Review and approval of study publications and substudy proposals;
(iiii) Oversee and analyze the project s data;

Co-investigator will be responsible for:
(i) Data collection and entry;
(ii) Data processing, analysis, storage, and deposition in public databases as appropriate;
(iii) Data documentation

Statistician will be responsible for:
(i) Agreement of the final Protocol and the Data Analysis Plans;
(ii) Data processing and storage;
(iii) Oversee and analyze the project s data.
(iiii) Making, testing, and validating changes to the database

6b. How will responsibilities for managing data activities be handled if substantive changes occur in the personnel overseeing the project's data, including a change of Principal Investigator?

Example:

Prior to the study conduct, the co-investigator will be tasked with this responsibility. In case of departure of the PI, efforts will be made to identify a new PI to carry on the project, and data management responsibilities will be re-negotiated as appropriate between the new PI and the Senior Analyst. In case of departure of the Senior Analyst, a new analyst will be recruited, and/or the PI will take over all remaining responsibilities.

6b. What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

Example:

Our project includes a data management budget of 500 CAD. We will use an open data repository. Therefore, there will be no cost.

7. ETHICS AND LEGAL COMPLIANCE

7a. If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project? If applicable, what strategies will you undertake to address secondary uses of sensitive data?

Example:

Acquired patient samples will be labelled with unique sample IDs, which will be used to link all data and metadata. No other patient personal information will be retained among the study data. Raw genomic data, which can itself be personally identified, will be retained only until three years after completion of the study. During that time, all data will be password protected, and raw data will be accessible only to the PI and Senior Analyst. De-identified summary data, indexed by sample IDs, will be shared upon publication.

Only de-identified data will be used for secondary uses, and all study participants will be informed about data sharing in the informed consent. The informed consent will describe that the personal identifiers will be removed and cannot be easily linked back to the identity of the participant. They will also be informed that the data will be shared publicly, unless they opt out of having their data shared.

7b. How will you manage legal, ethical, and intellectual property issues?

Example:

The research protocol will be approved by the respective Research Ethics Board. All the participants will be informed about what data will be collected, how the data will be used and reused.