Plan Overview

A Data Management Plan created using DMPonline

Title: Mock DMP for RNA-Seq Gene Expression Experiment

Creator: Siobhan Clery

Principal Investigator: Dr. Jane Doe

Data Manager: Dr. Jane Doe

Contributor: Dr John Smith

Affiliation: University College Dublin

Funder: Health Research Board (HRB) Ireland

Template: Health Research Board DMP Template

Project abstract:
This is an example DMP for RNA-Seq gene expression studies. We will be comparing gene expression of cancer versus healthy normal from matched samples of humans. The data will be a mixture of published (eg TCGA) and unpublished work generated in-house.

The in house data will be composed of 20 breast cancer samples plus 20 matched breast normal samples from the same individual. Each sample will be taken in triplicate resulting in a dataset of 180 samples.

** **This DMP is created with the assumption that the project lead has access to a local secure server that is well maintained. The onus is on the P.I. to ensure this or a suitable alternative is in place prior to the project's start.** **

All text in red relates to lab and university-specific policies, resources, and personnel. The onus is on the P.I. to ensure these are in place if you want to incorporate those parts of this example plan into your own DMP.

ID: 124118

Start date: 01-09-2023

End date: 01-09-2027

Last modified: 27-06-2023

Grant number / URL:

Mock DMP for RNA-Seq Gene Expression Experiment

Data description and collection or re-use of existing data

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

In-house generated RNA-Seq dataset:

- This will be conducted using a patient case-control cohort from breast tumor biopsies with matched normal tissue samples received through our collaborators in UCD.

- This dataset will be collected by experts in the project, with their own equipment.

- RNA library preparation will be performed using Illumina Stranded Total RNA Prep kit and single-end sequencing will be performed using a HiSeq 3000 with 75bp reads aiming for an average library size of 50 million reads. The output BCL files will be converted to FASTQ format using bcl2fastq2 v2.20.

- We will perform all bioinformatic analysis using nextflow workflow management software. The quality of FASTQ files will be assessed using FASTQC. Reads will be aligned to rRNA reference genome to remove ribosomal contamination prior to alignment to the GRCh38 reference genome using STAR aligner. Mapping quality will be assessed using Picard. STAR will be used to count the number of reads per gene. Transcript per million (TPM) normalization of the count data will be performed.

Publicly available dataset:

- STAR count data will be downloaded from the Cancer Genome Atlas (https://gdc.cancer.gov).

- This is open-access data and is freely available with an obligation to quote the source (e.g. CC-BY). This data will also be processed using nextflow workflow manager.

Explain how data provenance will be documented:

- We will use RO-Crate (Research Object Crate), to aggregate and describe research data with associated metadata.

- Electronic Lab Notebooks will be kept by all members of the team to document data processing and analysis. These will be stored in a secure shared Google Drive project specific folder which only members of the project will have access to.

- Data provenance for files created during the bioinformatic workflow will be documented using nextflow log command with a template file provided eg:

$ nextflow log nextflow_run_name -t template.html > provenance.html

- With Nextflow, in most cases, you don't need to manage the naming of output files, because each task is executed in its own unique directory, so files produced by different tasks can't overwrite each other. However, we will overwrite this to ensure consistency in file naming and to easily match log file to the processed sample. Therefore the log file name will follow the same format as detailed below in the "Documentation" section.

- We will use metadata schemas containing provenance information in our README files and in any kind of data documentation and metadata files.

- We will also upload our computational workflows to a repository such as WorkflowHub.

- All results files and code will include a version number so that updates will be documented.

Briefly state the reasons if the re-use of any existing data sources has been considered but discarded:
NA- existing data is being reused.

What data (for example the kind, formats, and volumes), will be collected or produced?

Data Type	Data Format	Format Justification	Volume	Purpose of data collection	WP/Institution Responsible
FASTQ Sequence and	FASTQ format is a text-based format	It is the standard	360 files of average size	This data contains all	WP2: Bioinformatic

Sequence Quality Format	containing both the nucleotide sequence and its corresponding quality score. File extension: .fastq.gz FASTQ files are usually compressed using the gzip format. This is an open, non-proprietary format.	format for raw sequencing data and is the format required for most data repositories.	10 GB (i.e. approximately 3600.0 GB in total)	the sequenced reads for the samples.	Analysis of sequenced data generated in-house
Binary Aligment Map (BAM) Format	BAM file is a binary compressed version of a sequence alignment file. File extension: .bam This is an open, non-proprietary format.	It is the standard format for aligned sequence reads. It is accepted by most tools for downstream analysis.	180 files of average size 15 GB (i.e. approximately 2700.0 GB in total)	This data contains the genome coordinates for each read so that they can be mapped to gene coordinates for downstream analysis.	WP2: Bioinformatic Analysis of sequenced data generated in-house
Binary Aligment Map (BAM) Index File	This file contains the index for the corresponding BAM file. File extension: .bai This is an open, non-proprietary format.	It is required in order to use the BAM files.	180 files of average size 10 MB (i.e. approximately 1.8 GB in total)	These files are required in order to use the BAM files as it maps the information in the BAM file to genomic coordinates.	WP2: Bioinformatic Analysis of sequenced data generated in-house
Expression count tab-separated files	This is a tab separated text file which contains read counts for each gene in a sample. File extension: .txt This is an open, non-proprietary format.	This is the standard format of gene expression files. It is a plain text file which allows for easy interpretation and analysis.	180 files from the in-house derived data set plus 1000 samples from TCGA. Average file size is 0.005 GB (i.e. approximately 5.9 GB in total)	These files contain the quantitative counts which are used to compare differences in gene expression between and within samples.	WP2: Bioinformatic Analysis of sequenced data generated in-house WP3: Validation
Reference FASTA file + accompanying files	Reference files are a single reference sequence in FASTA format. It is a text based format. Nucleotide or peptide sequences are represented using single letter codes, split into contains. Each contain sequence is preceded by a description line.	This is the standard format for reference sequences used for alignment.	3 files totalling about 3.5GB.	This file is required in order to map the sequenced reads from the FASTQ files to the human genome to create the aligned BAM files.	WP2: Bioinformatic Analysis of sequenced data generated in-house

File extensions: .fasta or .fa

Each FASTA file requires an index (.fai) and a dictionary (.dict) file to allow efficient random access to the reference base.

Documentation and data quality

What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?

- We will follow the set of well established standards and checklist for the minimum metadata described by the Minimum Information about a high-throughput nucleotide SEQuencing Experiment (MINSEQE) for transcriptomics data.
- We will gather sample-level, assay-level and analysis-level metadata.
- Documentation will be stored within the relevant project folders during the timeline of the project.
- A study level document will be kept detailing the project summary, objectives and all other relevant study details.
- A README plain text file will be kept within each project folder detailing the folder structure and contents of files within the folders.
- Results stored in table format will have an accompanying plain text file describing the values contained within each column.
- Files downloaded from public repositories i.e. TCGA will have an accompanying file with details including the date the data was downloaded.
- Scripts for analysing data will be automated and version controlled to ensure the same data processing steps are applied to all samples.
- For expression data that has been generated in-house a nextflow workflow management pipeline will be used to automate the steps from FASTQ to read alignment to gene expression quantification. The nextflow log file option will be used to document metadata relating to the processing of data such as date, tools and scripts used.
- All versions of tools used for the analysis of data will be documented in a plain text file.
- The versions of all databases used and/or the date the database was downloaded will also be documented.
- We will use standard ontology terms when describing data e.g. Gene and Transcript IDs will be converted to a common gene id (ENSEMBL) to allow easy mapping between resources. ENSEMBL gene and transcript IDs are more stable compared to gene symbol/gene name and will allow longevity and ease of use.
- Expression counts from each sample will be aggregated into a single expression file that will be versioned to facilitate reproducibility.
- Samples IDs will be anonymised to ensure that they cannot be mapped back to an individual patient.

- Day-to-day work will be documented in an online lab notebook.

We will employ the following folder structure to our project data:
project/

--> code/ #code needed to go from input files to final results

--> data/ #raw and primary data (never edit!)

--> --> raw_external/

--> --> raw_internal/

--> --> meta/

--> doc/ #documentation of the study

--> intermediate/ #output files from intermediate analysis steps

--> results/ #output from workflows and analyses

--> --> figures/

--> --> reports/

--> --> tables/

--> scratch/ #temporary files that can safely be deleted or lost

--> README.txt #file and folder description
The following name convention will be used when generating files:
1.Project number / Experiment Acronym
2. SampleID
3. Analysis Type e.g. ExpressionCounts
4. Date of creation
5. File extension
Example File Name: PROJ1_SAMP1_ExpressionCounts_01MAY23.txt

What data quality control measures will be used?

- For data generated in-house we will ensure standard quality control checks such as FASTQC for sequence quality are employed and all samples are assessed before moving onto the next stage of the workflow. Only those samples that have passed the QC criteria will progress to the next stage of analysis.

- For data downloaded from TCGA we will use the data generator's own recommended QC measures for gene expression profiles for selecting samples to include.

- When aggregating data from different sources we will perform batch correction to ensure there are no biases introduced as a consequence of data being generated from different sources.

- We will automate all bioinformatic analysis workflows to ensure all data is processed in the same way.

Storage and backup during the research process

How will data and metadata be stored and backed up during the research process?

- Data will be stored on a local UNIX/LINUX server hosted within our research institute.
- All data will be automatically backed up each week to restricted access resources hosted in UCD's Daedalus Data Centre.
- There will be restricted access to the data so only those authorised to work on the project will have access to it during the data generation and analysis phase. This will be maintained by the project data manager using UNIX/LINUX file access permissions system.
- A copy of the original raw data will be kept in a separate folder with "read-only" access so that it cannot be deleted or moved during the project analysis stage. Only one copy of the raw data will be kept at the end of the project.

How will data security and protection of sensitive data be taken care of during the research?

- All members of the research team will install operating system and application updates when they are available, install Sophos endpoint protection software provided by UCD and use a secure network connection when connected to the Internet. While on campus either the cabled network or the Eduroam Wi-Fi network will be used and when off-site the Staff Virtual Private Network (VPN) will be used to secure the network connection. www.ucd.ie/itservices/ourservices/security/protectingyourdevice/protectyourlaptopanddesktop
- Expression counts will not be patient identifiable. SampleIDs will be anonymised so that they cannot be linked back to an individual. The nature of expression profiles means that it cannot be attributed to a specific individual.
- Raw sequencing and alignment files will also be anonymised when assigned an identifier. However, as this data contains sequence variants, the very nature of the data means it is unique to an individual and therefore can be attributed to a person. This data will be kept on a secure server where only authorised access is allowed.
- Data will be backed up weekly so that in the event of an incident it will be recoverable with support from our IT service.
- Project members will be granted access to the data on the server. They will not have permission to copy raw data from this folder in order to ensure the data stays in a secure environment.
- If transfer of genomic data is required (e.g to a data repository) the data will first be encrypted using tools such as the EgaCryptor. Data will be transferred securely using ftp or aspera.
- Other datasets that do not contain genetic information can be transferred using HEAnet FileSender.
- *Relevant UCD policies, procedures & resources*:

IT Services policies and procedures www.ucd.ie/itservices/ourservices/security/policiesandprocedures

UCD Password Protection Policy hub.ucd.ie/usis/!W_HU_MENU.P_PUBLISH?p_tag=GD-DOCLAND&ID=182

Information Security Awareness Online Course www.ucd.ie/itservices/ourservices/security/training/onlinetraining

Top security tips www.ucd.ie/itservices/ourservices/security/topsecuritytips

Legal and ethical requirements, codes of conduct

If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

- Our data will not contain name or date of birth of the patient. Instead, we will have an anonymised patient ID and age of the participant at the time the specimen was taken. However, we will be working with special category data because we have genetic information from the patients.

- We will have informed consent from participants whose data we are using to ensure that we have the correct permissions to use, preserve and share the data. We will make sure to remove any patient-identifiable information before making any data open access and any genomic data which can be used to determine the identity of a patient will be kept in a controlled access location both during and after the project has ended. Controlled access to the data will be made available through a Data Access Committee.

- For transferring data to recipients with approved access, genomic data will be encrypted using the Crypt4gh python tool and the recipient will be provided with a key to de-encrypt with Crypt4gh when the data is received. Only the recipient will have the key and it will expire within a specified time frame. Data will be transferred securely using sftp or aspera.

- All project members will receive UCD Data Privacy & Security Training.

- Upon project completion data will be archived in a secure repository located within the EEA (ie EGA repository for genetic data).

- We will receive approval from the UCD ethics board for research using human data.

- A Data Protection Impact Assessment will be completed as per UCD policy.

- *Relevant UCD policies, procedures & resources*:

UCD GDPR www.ucd.ie/gdpr

UCD GDPR Key Terminology of GDPR www.ucd.ie/gdpr/about/keyterminologyofgdpr

UCD Data Protection resources www.ucd.ie/gdpr/guidanceresources

UCD Data Protection Policy www.ucd.ie/gdpr/t4media/UCD_Data_%20Protection_%20Policy.pdf

How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

- We will be working with the philosophy as open as possible for our data. Data that is not legally constrained will be released as open access upon publication, with the obligation to quote the source (CC-BY license).

- Sequencing data cannot be released as open access due to the ability to identify individuals using genetic variants. Therefore, this data will be released as controlled access to researchers who have approved access, under a general research use license.

- Personal data is owned by the participant whose data we are using. Except for publications that stem from this project, all data generated during the project will be owned by UCD.

- Re-used data can be used with the following conditions:
    - *TCGA expression counts is freely available with obligation to quote the source.*
    - *Human reference genome used for alignments is freely available for anyone to use.*
- *Relevant UCD policies, procedures & resources*:

UCD Intellectual Property Policy sisweb.ucd.ie/usis/!W_HU_MENU.P_PUBLISH?p_tag=GD-DOCLAND&ID=157

NovaUCD - Knowledge Transfer Supports www.ucd.ie/innovation/knowledge-transfer/researcher-supports/what-can-i-expect

NovaUCD (Intellectual Property) www.ucd.ie/innovation/knowledge-transfer/researcher-supports

UCD Research and Innovation Signing contracts and agreements www.ucd.ie/research/portal/win/signcontractsandagreements

What ethical issues and codes of conduct are there, and how will they be taken into account?

- We will comply with all UCD policies regarding research ethics and good practice in research with human data
- We will follow best practices in terms of data management and security, following best practices set out by UCD
- We will ensure we are adhering to EU GDPR for working with sensitive or special category data.
- We will receive ethics approval prior to the collection of data and will perform a Data Protection Impact Assessment.
- We will ensure that all members working on the project receive training on research integrity, IT security and GDPR and make sure they are aware of all UCD policies relating to working with human research data.
- *Relevant UCD policies, procedures & resources*:

UCD Office of Research Ethics www.ucd.ie/researchethics

UCD Policy on Research Ethics hub.ucd.ie/usis/!W_HU_MENU.P_PUBLISH?p_tag=GD-DOCLAND&ID=218

UCD Code of Good Practice in Research with Humans and Animals, 2019 hub.ucd.ie/usis/!W_HU_MENU.P_PUBLISH?p_tag=GD-DOCLAND&ID=14

UCD Research Integrity Policy: hub.ucd.ie/usis/!W_HU_MENU.P_PUBLISH?p_tag=GD-DOCLAND&ID=184

Research Ethics policies & guidelines www.ucd.ie/researchethics/policiesguidelines

Research Integrity resources www.ucd.ie/researchintegrity/resources

Data sharing and long-term preservation

How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

- Raw sequencing and aligned files cannot be completely open due to legal reasons.
- Data that is not legally restrained will be released upon publication.
- Restricted sequence data, along with their metadata, will be uploaded to the European Genome-Phenome archive (EGA) which provides restricted access to researchers who request data for specific research uses. The data will be archived behind a controlled access firewall and will only be made available to researchers after approval by a Data Access Committee and access will be made available through authorised consent.
- A data use certification agreement will be required for controlled access data.

How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

- Raw sequencing data and alignment files will be preserved which will allow users to reproduce our results or allow users to repurpose our data for other related analyses, under a restricted use policy. This data will be stored in EGA upon publication.
- Final expression data will be made available as open access which will allow researchers to use our results to carry out additional analysis. This data will be stored in EGA upon publication.

What methods or software tools are needed to access and use data?

- The EGA download client pyEGA3 is required for downloading data to which the user has been granted access.
- A secure server is required to host downloaded data and to reuse the data.
- The pyEGA3 client is compatible with any OS with Python 3.6+ installed. The client requires a connection to the internet, sufficient space on the destination drive, and the EGA download account credentials.
- Files downloaded from EGA are transferred over secure HTTPS connections.
- The python download client cannot be used to download dataset metadata. Metadata will need to be downloaded through the EGA Archive webpage and will therefore require a web browser for access. Users will be required to login into the EGA webpage to download data they have access to.

How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

- The data repository which we will submit our data to will provide a globally unique persistent identifier for the data.
- For individual samples within the data set we will create non-patient identifiable unique IDs which will be used to trace all files generated related to an individual sample.

Data management responsibilities and resources

Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

- The P.I. will have overall responsibility for data management. A Data Information Manager will support the P.I. to ensure all data is stored properly on secure servers which will be backed up regularly.
- The Data Information Manager will help to administer access control policies and the curation of data resources created and shared internally and to open-access online resources and ensure the security and integrity of both the data resources and infrastructure on which they will be processed and reside.
- The P.I will also be responsible for implementing the DMP and ensuring it is reviewed and revised as the project progresses. A member of the group will also be nominated as data steward for the project to ensure all policies and plans are implemented throughout the project.
- The P.I. is also responsible for maintaining the finished resource.

What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

- We will work with the university's Information Manager to ensure computational resources are kept secure, including providing training to members of the project on how to keep data safe and secure.
- The research institute where the PI is based has a dedicated Data Information Manager who manages and maintains the computational resources and will provide support on managing resources, long-term storage and maintaining all hardware.
- We will ensure all members of the project undergo training on making data FAIR. This will be supported by the institution's own Data Manager as well as by resources offered through ELIXIR Ireland.
- We will assess the FAIRness of our data using tools such as FAIR evaluator or FAIRshake.
- Data will be archived through EGA which provides submitters with a completely free, secure and permanent archiving solution for sharing data worldwide.