The sample Data Management and Sharing Plan below is for a proposal conducting non-human basic research. It is one of four examples provided by NIDDK.

NIDDK Example Data Management and Sharing Plan – Non-Human Basic Research

Element 1: Data Type:

A. Types and amount of scientific data expected to be generated in the project:

This project will produce clinical measures, phenotypic characteristics, microscopic imaging, and transcriptomic gene expression profiles from mouse models of nonalcoholic steatohepatitis (NASH)-driven hepatocellular carcinoma (HCC). Data will be collected for up to 50 mice generating a total of three datasets. The following data files will be produced during the course of this project:

- Clinical and Phenotypic data including animal sex, body weight, specific organ weights, age and developmental stage, tissue profiled, and disease outcome.
- RNA sequencing datasets including normalized transcript and gene level expression counts. This dataset will also include a number of quality control metrics including total reads, clipped reads, sequencing platform, and any other relevant metrics that arise during the project.
- Light and confocal microscopy images of formaldehyde-fixed paraffin-embedded tissue slices.

B. Scientific data that will be preserved and shared and the rationale for doing so:

All three datasets described in A will be preserved and shared through public repositories.

C. Metadata, other relevant data, and associated documentation:

- A detailed methods section outlining the collection of each scientific data generated with this work will be provided. Any step-by-step protocols developed in this project will be shared as a supplementary protocol document. Specifications about instruments and technologies used to produce this data will also be provided.
- All steps in the data analysis pipelines and workflows will be characterized and documented on GitHub.
- A data dictionary describing all phenotypic and clinical variables collected will be provided in MS Excel format and uploaded to the repository with the associated dataset.
- A set of well-established standards and minimum metadata checklists exist for various aspects of transcriptomics. Minimum Information about a high-throughput nucleotide SEQuencing Experiment (MINSEQE) describes the minimum metadata that is needed to enable the unambiguous interpretation and facilitate reproduction of the results of the experiment and will be followed to the extent possible.

Element 2: Related Tools, Software, and/or Code:

- Raw data files from the RNAseq experiments include FASTQ files. This is a text-based sequencing data file format for storing next-generation sequencing data (both raw sequence and quality scores) from Illumina sequencing instruments. FASTQ files are the standard format and can be used as input for a wide variety of secondary data analysis pipelines.
- Dockerized analysis pipelines for generating gene expression counts from FASTQ files will be shared on Docker Hub and GitHub.
- Microscopy images will be available as .TIFF files and can be viewed with common image viewing software.
- Phenotypic and clinical variables will be stored as tab-separated files and can be managed using common spreadsheet-based software such as Microsoft Excel.
- Statistical analyses of the data will be performed using R and Python programming languages. This code will be shared on GitHub for public access. Code will be available at the time of publication or at the end of the award.

Element 3: Standards:

Mouse anatomical and developmental descriptions will use language specified by the mouse adult gross anatomy ontology and the mouse developmental stages ontology available from the EMBL-EBI Ontology Lookup Service. The Animal History common data elements (CDE) from the National Institutes of Health (NIH) CDE repository will be used to record relevant information about mouse subjects. Community standard file formats will be provided, including FASTQ files for RNA-sequencing files results, spreadsheets for normalized gene expression counts, .TIFF files for microscopy imaging, and tab-separated spreadsheets for clinical measures and phenotypic outcomes.

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:

Transcriptomic datasets that can be shared will be deposited in Gene Expression Omnibus (GEO, National Center for Biotechnology Information [NCBI]). Clinical and phenotypic datasets will be associated with the RNAseq submission to GEO. GEO is an NIH-supported repository that archives and freely distributes microarray and next-generation sequencing data. Light and confocal microscopy images of liver tissue will be made available at The Cancer Imaging Archive (TCIA), which is an NIH-supported Scientific Data Repository. TCIA accepts radiology and microscopy imaging data related to cancer generated with any NIH institute funding. It currently houses many datasets on liver cancer.

B. How scientific data will be findable and identifiable:

GEO provides metadata, persistent identifiers (accession numbers), and long-term access. This repository is supported by NCBI. Datasets are available under an open access policy. Unique identifiers associated with the data will be referenced in the corresponding publications.

C. When and how long the scientific data will be made available:

Scientific data will be shared as soon as possible. Scientific data included in published manuscripts will be made available at the time of publication; all other scientific data will be made available no later than the end of the award. Data will be preserved and available for at least 5 years' duration. Raw data, intermediate data, and the code/software/tools used to develop the published or submitted dataset will be shared at the time of data submission or publication and for at least 5 years' duration.

Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data:

There are no use limitations associated with the scientific data generated in this study. There are no ethical or legal issues that can have an impact on data sharing. No personal data will be published in this project.

B. Whether access to scientific data will be controlled:

Data will be made publicly available and data access will not be controlled.

C. Protections for privacy, rights, and confidentiality of human research participants:

No human samples or research participants will be used in this study.

Element 6: Oversight of Data Management and Sharing:

The Principal Investigator for this project, Dr. ABC, will ensure that this Data Management and Sharing (DMS) Plan is followed. The institutional official (title and role), will be responsible for oversight of compliance with the accepted DMS Plan. Compliance will be evaluated annually during the award period and progress towards the plan's DMS activities will be included in the annual Research Performance Progress Report (RPPR) submitted to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Project Officer. At the project conclusion, the final progress report will summarize how the DMS objectives were fulfilled and provide links to the shared dataset(s).