s

# DATA MANAGEMENT AND SHARING PLAN

If any of the proposed research in the application involves the generation of scientific data, this application is subject to the NIH Policy for Data Management and Sharing and requires submission of a Data Management and Sharing Plan. If the proposed research in the application will generate large-scale genomic data, the Genomic Data Sharing Policy also applies and should be addressed in this Plan. Refer to the detailed instructions in the application guide for developing this plan as well as to additional guidance on [sharing.nih.gov.](https://sharing.nih.gov/) The Plan is recommended not to exceed two pages.

Text in italics should be deleted. There is no “form page” for the Data Management and Sharing Plan. The DMS Plan may be provided in the *format*

shown below.

An example DMS plan proposing to collect gene expression data from human or non-human subjects.

# Element 1: Data Type

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

Summarize the types and estimated amount of scientific data expected to be generated in the project.

This transcriptomics study will generate next-generation sequencing (NGS) data using RNA-sequencing technology. We will use four conditions (that includes a control group) with four replicates per condition resulting in a total of 16 samples for total RNA-sequencing. The raw data (FASTQ and TSV format) and the

processed data (TSV format), as well as the associated metadata (TXT format) will be generated and the total size of the data is estimated to be about 130-160 GB.

# Scientific data that will be preserved and shared, and the rationale for doing so:

Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.

All raw and processed NGS data from the transcriptomic study will be shared on the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), which is an NIH-funded repository, and freely distributes these data to the broad scientific community. To facilitated file transfer and save storage space, the raw and processed data will be stored in as compressed files (GZ format or TAR.GZ format). Even though, we share the processed data (such as intermediate files, read counts, differentially expressed genes, etc.), it is important to share the original raw data with the community to enable analysis using newer tools that would be available in the future.

# Metadata, other relevant data, and associated documentation:

Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

A brief study protocol with corresponding metadata fields and instrumentation details will also be submitted along with the raw sequencing data to facilitate data interpretation. The Institutional Certification will be submitted to NIH once we have been told that a grant award is likely. Within the first six months following the award, we will submit the Data Submission Agreement to {Name of the NIH Institute funding this project}.

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

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

Transcriptomic data undergo an extensive automated data cleaning process in the laboratory using Quality Control (QC) tools such as FastQC. While all sequencing data from this proposal will be generated using {Illumina} instrumentation, the number of reads per sample, differences in sequencing coverage and the kits used for RNA-seq library preparation varies between studies. We will use STAR for aligning reads to the reference genome and the RSEM tool for annotation and quantification of transcripts at the gene and isoform level to generate normalized TPM and FPKM values. We will use DESeq2 to perform two-group comparison of the RNAseq data, and the output from DESeq2 can be used for downstream analyses such as generating heatmaps and volcano plots. All the tools used in this study are open-source and available in the public domain.

# Element 3: Standards:

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist.

Data will be deposited and shared on the GEO repository. The accession number (GSE) assigned to the data by GEO will be used as a persistent unique identifier, so that the data are findable and identifiable to the general public. We will prepare the metadata in compliance with the MIBBI (Minimum Information for Biological and Biomedical Investigations) standards) using the guidelines in the GEO metadata spreadsheet (<https://www.ncbi.nlm.nih.gov/geo/info/examples/seq_template.xlsx>) to enable interoperability.

The following common data elements will be collected to facilitate aggregation of this data set with other data sets. At the sample level, we will provide details on the organism, cell line, cell type, genotype, treatment, total RNA or mRNA, single or paired-end, sequencing instrument and model. Regarding protocols, we will provide the growth, treatment and extract protocols, library construction and strategies used, data processing steps, tools and the genomic build/assembly of the reference genome used for alignment.

# Element 4: Data Preservation, Access, and Associated Timelines

1. **Repository where scientific data and metadata will be archived:**

Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived; see [Selecting a Data Repository](https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository)).

High throughput transcriptomic sequencing data generated in this study will be deposited into Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), which is an NIH-funded repository, and freely distributes these data to the broad scientific community. All data will be deposited to *GEO* starting 12 months after the award begins and will be deposited every six months thereafter following the usual data submission dates.

# How scientific data will be findable and identifiable:

Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

Data will be findable for the research community through the NIH supported Gene Expression Omnibus (GEO) repository using the unique GEO identifier (GSExxxxx). From the GEO page, links to access sample level data will be available. At the time of publication, GSE ID will be provided to the publisher, which is a requirement for most of the journals these days. Also, links to the GEO ID will be provided in the supplementary data files of the publications.

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

Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

The research community will have access to data at the end of the grant award or when a publication has been submitted, whichever comes first. Once the data are submitted to GEO, it will be accessible the research community permanently. We will comply with the broad and rapid data sharing as required for NIH funded projects and also any additional requirements by the funding institution.

**Element 5: Access, Distribution, or Reuse Considerations**

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

NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See [Frequently](https://sharing.nih.gov/faqs%23/data-management-and-sharing-policy.htm) [Asked Questions](https://sharing.nih.gov/faqs%23/data-management-and-sharing-policy.htm) for examples of justifiable reasons for limiting sharing of data.

All research participants will be consented for broad data sharing.

# Whether access to scientific data will be controlled:

State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

All data access will be permitted as per the NIH Genomic Data Sharing Policy guidelines, and any additional outlined by the repositories. To request access of the data, researchers will use the standard processes outlined by the repositories.

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

If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

An institutional IRB approval will be obtained before collecting data from human subjects. Consent will be obtained from all research participants for research use of genomic data either in a de-identified or identified format, as relevant to the project. Samples consented only for the de-identified usage will be processed differently from those consented for identifiable use and stored on a separate server. All the HIPAA variables will be removed, and de-identified patient IDs will be created using a one-way hash system. Only the de-identified patient IDs will be used in all data files containing phenotypic or clinical data, which will be used for sharing with the research community. All data will be stored on secure servers located at the HIPAA-compliant data center at the University of Nebraska Medical Center campus that maintains strict enterprise-level firewall security measures.

# Element 6: Oversight of Data Management and Sharing:

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

The Office of Sponsored Programs at University of Nebraska Medical Center administering this award has created a data management and sharing plan compliance system as part of their process for submitting this application. In addition, the following individuals will monitor and manage the implementation of this Plan on a day-to-day basis:

Lead PI, Jane Doe PhD, ORCID: xxxx-xxxx-xxxx-xxxx, will be responsible for the day-to-day oversight of data management activities and data sharing. Broader issues of DMS Plan compliance oversight and reporting will be handled by the PI and Co-I team as part of general stewardship, reporting, and compliance processes. The following individuals will be responsible for data collection, management, storage, retention, and dissemination of project data, including updating and revising the Data Management and Sharing Plan when necessary

John Doe, Database manager, UNMC, ORCID 0000-000x-xxxx-xxxx, johndoe@unmc.edu, will be responsible for…