Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D18AFFde93

Title: Rhode Island IDeA Network of Biomedical Research Excellence April 2024

Creator: Christopher Hemme - **ORCID:** 0000-0002-4092-211X

Affiliation: University of Rhode Island (ww2.uri.edu)

Data Manager: Christopher L. Hemme

Project Administrator: Bongsup Cho, Brett Pellock

Funder: National Institutes of Health (nih.gov)

Grant: P20GM103430

Template: NIH-GEN DMSP (Forthcoming 2023)

Start date: 04-01-2024

End date: 03-31-2029

Last modified: 07-08-2024

Copyright information:

Rhode Island IDeA Network of Biomedical Research Excellence April 2024

Data Type

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.*

Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

In the proposed project, data from Core facilities will be generated via the following methods: next-gen sequencing, mass spectrometry, HPLC, microscopy, single-cell omics, and spatial omics. The experimental designs will vary based on the individual researchers using the facilities. The total size of the data collected will vary based on the type of experiment conducted but is not expected to exceed 10 terabytes per experiment.

We expect to generate the following data file types: images (.TIFF, .JPG, .PNG), sequencing (.FASTQ, .FASTA), bioinformatics (.SAM, .BAM, .BED, .VCF, .WIG), mass spec (.RAW and other), tabular (.CSV, .TSV). Raw data files will be analyzed individual researchers based on their own protocols, through vendor-provided software packages, or through established Core facility workflows.

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.*

For scientific data, the responsibility for preserving and sharing data will generally be left to individual researchers. All researchers are expected to make every reasonable effort to make their data publicly available at the earliest opportunity. For large datasets such as sequencing, imaging, or mass spec data, the Core facilities will implement a data storage policy consistent with the University's cybersecurity policies that will include a combination of in-house and cloud storage. The Core facilities will commit to storing data for at least one year and individual researchers will commit to storing data for the lifetime of their projects. Human subject data will be properly anonymized before release of data to the public.

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.

To facilitate interpretation of data, particularly for omics data, relevant information about the data will be released. This includes metadata (environmental, clinical, etc.), protocols, code, statistical models. Documentation and support materials related to clinical information will be compatible with the clinicaltrials.gov Protocol Registration Data Elements.

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.

Code generated by the MIC will be stored on the MIC GitHub repository using the MIT license. Bioinformatics workflows are typically coded in Snakemake (Python) and R using Anaconda and containers. MIC workflows are typically deployed on URI HPC resources. Individual researchers operating under the RI-INBRE program will be expected to follow similar procedures.

Sequencing data generated by the Illumina MiSeq is automatically transferred to the Illumina BaseSpace system where it can then be transferred to the user or to the URI HPC systems. The CRCF, MIC and URI College of Pharmacy will develop workflows for proteomics and metabolomics data analysis. Additional workflows will include data generated for single cell or spatial omics methods.

Use of artificial intelligence, machine learning, and deep learning (AI/ML/DL) tools in all funded projects will be comprehensively documented. Creative endeavors including research projects are expected to be substantively original unless the use of AI/ML/DL is integral to the project. All use of AI/ML/DL tools, including prompts for generative AI algorithms, should be documented and made available through publication or by other means (e.g. GitHub). Researchers should indicate in the acknowledgements section of all publications the use of AI/ML/DL tools and the degree of their use.

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

For all data generated by the core facilities or by RI-INBRE researchers, FAIR (**F**indability, **A**ccessibility, **I**nteroperability, and **R**euse) principles for data will be followed including the use of open file formats and persistent unique identifiers.

Omics data generated by the core facilities or by individual researchers will follow common data standards for the process. Users may use workflows developed by the MIC or a third party. NGS workflows will be designed to use standard omics data formats (.fastq, .gff., .gtf., .sam, .bam, .bed). Proteomics and metabolomics workflows will be designed to use standard omics data formats (.raw). Count data from omics workflows will be stored as .csv files and relevant metadata will be stored as plain text or csv files (.txt, .csv).

For other types of experiments, data will as much as possible follow conventional data standards for the instruments/workflows in question.

Data Preservation, Access, and Associated Timelines

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)

Sequencing and proteomics/metabolomics data and related protocols and metadata will be required to be deposited in public repositories (e.g. Genbank, Sequence Read Archive (SRA), Gene Expression Omnibus (GEO), and Proteomics Identification Database (PRIDE)). The MIC will post all relevant data (e.g. code) to the MIC GitHub account, and use of Github to share code, protocols, and small non-standardized datasets will be encouraged for individual RI-INBRE researchers. Data and notebooks generated on cloud resources (e.g. All of Us) will be publicly available according to the protocols of the individual platforms.

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.

RI-INBRE will use Persistent Unique Identifiers (PIDs) to improve data findability . PIDs used will include ORCID iDs for people, DOIs for outputs (e.g., datasets, protocols), Research Resource IDentifiers (RRIDs) for resources, and Research Organization Registry (ROR) IDs and funder IDs for places, as much as possible to make data identifiable and findable. Data placed in public repositories will use the PIDs assigned as by those repositories (e.g. PubMed ID, accession numbers, BioProject ID, etc.).

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 core facilities will assist users on depositing data in the relevant repositories. For large datasets, the Core facilities will archive and maintain he data for a minimum of one year with options available for longer-term storage. Researchers who maintain their own data will be expected to maintain the data for the life of their project plus additional time after completion of the project (minimum 3 years total). All data generated using RI-INBRE funding will be subject to all relevant federal data sharing policies (e.g. 2023 NIH Data Management and Sharing policies, 2025 White House mandate, etc.). It is expected that all such data will be available at the time of publication of the data. Users are also expected to properly acknowledge the RI-INBRE grant in all publications and presentations and to comply with NIH Public Access Data policies (i.e., deposition of the manuscript in PubMed Central).

Access, Distribution, or Reuse Considerations

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 Asked Questions for examples of justifiable reasons for limiting sharing of data.

RI-INBRE will follow all relevant data privacy laws and regulations (i.e., HIPAA, FERPA, IRB policies). In the event of research generating clinical and/or human subject data, all efforts will be made to protect the privacy of the subjects including but not limited to anonymization of data and use of certificates of confidentiality. All such research will follow federal inclusion policies to ensure the research benefits individuals of all sexes/genders, races, ethnicities, and ages.

Research conducted on vertebrate animal subjects will be conducted by the standards of Public Health Service (PHS) Policy on Humane Care and Use of Laboratory Animals and the Animal Welfare Act.

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 human subject data funded by RI-INBRE will follow federal policies on the use of human subjects and is ultimately the responsibility of the individual researcher. Consent of participants on data sharing and preservation of data will be required. Anonymization and managed access procedures will be employed to protect participant privacy. All relevant regulations and laws (e.g., HIPAA) will be followed.

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).

All work on human subject data will follow standard IRB protocols for the investigator's institution and HIPAA regulations which includes informed consent documentation, plans for data management and sharing, and anonymization of data. Researchers will individually chose the proper methods to deanonymize human subject data.

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 Director of RI-INBRE MIC will oversee RI-INBRE data management and sharing policies, will be responsible for disseminating relevant policies to network participants, and will manage program metrics tracking with RI-INBRE administrators. The MIC Director will be responsible for data generated by the MIC. The CRCF Director and CRCF Manager will be responsible for data generated by CRCF until it transferred to the individual researcher. The individual researchers will ultimately be responsible for their own data.

Planned Research Outputs

Workflow - "Bioinformatics Workflows"

Standard bioinformatics workflows generated by the MIC (typically Snakemake workflows).

Dataset - "Next-Gen Sequencing Datasets"

MiSeq next-gen sequencing reads (.FASTQ) or data from functional genomics data analysis (.BAM, .BED).

Dataset - "Proteomics Datasets"

Raw mass spec files for proteomics and/or metabolomics

Dataset - "Single Cell and Spatial Omics Datasets"

Datasets associated with single cell or spatial omics, including sequences (.FASTQ) and image files.

Interactive resource - "Virtual/Augmented Reality Applications"

Virtual and Augmented Reality applications developed in-house (typically Unity)

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Bioinformatics Workflows	Workflow	Unspecified	Open	GitHub		None specified	None specified	No	No
Next-Gen Sequencing Datasets	Dataset	Unspecified	Open	SRA - Reads Gene Expression Omnibus		None specified	None specified	No	No
Proteomics Datasets	Dataset	Unspecified	Open	PRIDE		None specified	None specified	No	No
Single Cell and Spatial Omics Datasets	Dataset	Unspecified	Open	SRA - Reads		None specified	None specified	No	No
Virtual/Augmented Reality Applications	Interactive resource	Unspecified	Open	GitHub		MIT License	None specified	No	No