

## Plan Overview

---

*A Data Management Plan created using DMPTool*

**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](https://orcid.org/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:** 04-04-2024

### Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

---

## Rhode Island IDeA Network of Biomedical Research Excellence April 2024

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.

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.

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](https://clinicaltrials.gov) Protocol Registration Data Elements.

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.

For all data generated by the core facilities or by RI-INBRE researchers, FAIR (Findability, Accessibility, Interoperability, and Reuse) 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.

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.

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.).

The core facilities will assist users on depositing data in the relevant repositories. For large datasets, the Core facilities will archive and maintain the 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).

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.

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.

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.

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