RI-INBRE DATA MANAGEMENT AND SHARING PLAN

ELEMENT 1: DATA TYPE

A. Types and amount of scientific data expected to be generated in the project: Data will be generated from Core facilities via the following methods: next-gen sequencing, mass spectrometry, HPLC, microscopy, single-cell and spatial omics. Data sizes will vary based on the experiment but are not expected to exceed 10 terabytes per experiment. Data formats include: images (.TIFF), sequence (.FASTQ), bioinformatics (.SAM, .BAM, .BED), mass spec (.RAW), tabular (.CSV). Raw data files will be analyzed by individual researchers based on their own protocols, through vendor-provided software packages, or through established Core workflows.

B. Scientific data that will be preserved and shared, and the rationale for doing so: Researchers are ultimately responsible for their own data and are expected to make their data publicly available at the earliest opportunity. For large datasets (e.g. omics), the Core facilities will implement a data storage policy including in-house and cloud storage and will store data for at least one year. Individual researchers will commit to storing data for >3 years.

C. Metadata, other relevant data, and associated documentation: Relevant project data will be released including metadata, protocols, code, statistical models. Documentation related to clinical information will be compatible with the clinicaltrials.gov Protocol Registration Data Elements.

ELEMENT 2: RELATED TOOLS, SOFTWARE AND/OR CODE

Code generated by the Cores will be stored in a GitHub repository using the MIT license. Bioinformatics workflows are coded in Snakemake and R using Anaconda and are deployed on URI HPC resources. Individual researchers will follow similar procedures. Illumina sequencing data is stored in Illumina BaseSpace and then transferred to the user. Use of artificial intelligence and machine learning (AI/ML) tools in all funded projects will be documented. Research outputs are expected to be substantively original unless the use of AI/ML is integral to the project. All use of AI/ML tools, including prompts for generative AI algorithms, should be documented.

ELEMENT 3: STANDARDS

FAIR data principles will be followed including the use of open file formats and persistent unique identifiers. Omics data will follow common data standards and workflows will use standard data formats. Count data and metadata from omics workflows will be stored as plain text or .csv files. Other generated data will follow conventional data standards.

ELEMENT 4: DATA PRESERVATION, ACCESS, AND ASSOCIATED TIMELINES

A. **Repository where scientific data and metadata will be archived:** Omics data and related materials will be deposited in public repositories (e.g. NCBI repositories). The Cores and researchers will post all relevant data (e.g. code,

protocols, etc.) to GitHub or other repositories. Materials generated on cloud resources (e.g. All of Us) will be publicly available according to the protocols of the platforms.

B. **How scientific data will be findable and identifiable:** RI-INBRE will use Persistent Unique Identifiers (PIDs) to improve data findability including ORCID iDs, DOIs (e.g., datasets, protocols), and Research Resource IDentifiers (RRIDs), 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.).

C. **When and how long the scientific data will be made available:** The Cores will assist users on depositing data to relevant repositories. For large datasets, the Core facilities will archive and maintain data for at least one year. Researchers will maintain their data for a minimum of 3 years. All data generated using RI-INBRE funding will be subject to all federal data sharing policies will be available at the time of publication. Users will properly acknowledge the RI-INBRE grant in all publications and presentations and will comply with NIH Public Access Data policies (i.e., deposition of the manuscript in PubMed Central).

ELEMENT 5: ACCESS, DISTRIBUTION, OR REUSE CONSIDERATIONS

A. **Factors affecting subsequent access, distribution, or reuse of scientific data:** We 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, the privacy of the subjects will be protected through anonymization of data and use of certificates of confidentiality. 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.

B. **Whether access to scientific data will be controlled:** All human subject data 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.

C. **Protections for privacy, rights, and confidentiality of human research participants:** 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 choose the proper methods to deanonymize human subject data.

ELEMENT 6: OVERSIGHT OF DATA MANAGEMENT AND SHARING

The Director of RI-INBRE Molecular Informatic Core will oversee RI-INBRE data management and sharing policies, will be responsible for disseminating relevant policies to network participants. The individual researchers will ultimately be responsible for their own data.