NIEHS Standard Data Management and Sharing Plans

Element 1: Data Type

1A. 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.*

1A.1 Cryo-EM

Cryo-EM projects produce imaging data (2D and 3D) from transmission electron microscopes equipped with a direct electron detector (camera). Approximately 2 TB of data are generated for each experiment. The following data files will be produced in the course of a typical project:

1. Raw data: Micrographs are "movies" composed of frames collected over the exposure time. The size of each movie varies depending on the detector, the mode in which the images are recorded (counting or super-resolution) mode and number of frames. The format of these files varies depending on the software used for data collection. Typical formats include: half byte LZW compressed TIFF (open standard, non-lossy compression), MRC (UK Medical Research Council, open standard), dm2 (Digital Micrograph, Ametek's proprietary format) and eer (Thermo Fisher Scientific proprietary format). For the most efficient half byte LZW compressed format, current typical sizes range from 200 to 600 MB for each micrograph, and 2000 to 5000 micrographs for each dataset, for a total size that varies from 400 GB to 2.5 TB.
2. Intermediate data: Most intermediate image files are in 2D or 3D MRC format. Image processing software produce a wealth of metadata that can be used to reproduce the calculations performed during the analysis. The format of these files varies greatly depending on the software used (.csv, .xml, .json, scripts and plain text), but they can be compressed and are very small relative to the raw data.
3. Final maps: Final maps are typically 3D images encoded in MRC format. Their size is small relative to the raw data.

1A.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

Metabolomic and Lipidomic projects produce data ranging from 100 MB to 100 GB in size and will consist of the following file types:

1. Raw data: the direct output from the mass spectrometers that we have at the Institute. File types are: .RAW or .d.
2. Processed data: files including peaklist(s) (.mzxml) and search results (.cdResult) will be generated as appropriate.
3. Final result(s): tabular format with the most often stored as .txt, .csv, or .xlsx file.

1A.3 Mass Spectrometry - Quantitative Small Molecule Analysis

Quantitative Small Molecule projects produce roughly 150 MB of raw and processed data per sample per year and will consist of the following file types:

1. Raw data: direct output from the mass spectrometers that we have at the Institute. File types are: .RAW, .raw, and .wiff.
2. Final result(s): tabular format most often stored as .xlsx, .txt, or .csv.

1A.4 Mass Spectrometry - Protein and Peptide Analysis

Protein and Peptide projects produce roughly 1.1 TB of raw and processed data per sample per year. The file types include:

1. Raw data: direct output from the mass spectrometers that we have at the Institute. File types are: .RAW, .raw, .t2d, and .D.
2. Processed data: peaklist(s) (.mgf) and search results (.pdResult) will be generated as appropriate.
3. Final result(s): tabular format most often stored as .xlsx, .txt, or .csv.

1A.5 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations produce between 50 and 80 TB of data per project, per year, depending on system sizes, the number of samples, and the lengths of trajectories. Molecular docking requires download of databases such as the Zinc database for local use that may add an additional 20 TB per year. The file types include:

1. Raw data: coordinate information collected from MD trajectories stored in compressed netcdf or binary formats.
2. Processed data: trajectories will be analyzed using a variety of modules; also stored in compressed netcdf or binary format. Also, quantum mechanical calculations may be performed in combination with MD as used in the QM/MM methodology. Checkpoint files associated with this technique stored in binary format occupy the bulk of additional required data space.

1A.6 NMR – Metabolomics

NMR metabolomics projects generate roughly 0.1 TB of data per project consisting of the following types:

1. Raw data: time-domain NMR data accounting for 45% of the total storage.
2. Processed data: result of Fourier transform and analysis accounts for the bulk of the remainder. Ancillary data (e.g. fitted metabolites) will generate roughly 1MB of tabular data in the form of .csv or .xlsx files. Statistical analyses of the tabular data are both tabular and graphical in nature and typically stored in pdf format.

1A.7 NMR – Structural Biology

NMR structural biology projects generate roughly 0.1 TB of data per project consisting of the following types:

1. Raw data: time-domain NMR data accounting for 10-40% of the total storage.
2. Processed data: results of Fourier transform, analysis, and structural calculations account for the bulk of the remainder. Ancillary data (e.g. biochemical, enzymatic, or ELISA) will generate roughly 1MB of tabular data in the form of .csv or .xlsx files.

1A.8 X-ray Crystallography

X-ray crystallography projects produce roughly 20 TB of data per sample and consist of the following types:

1. Raw data: diffraction data that is processed by scaling and merging to generate a reflection file, typically in .mtz format.
2. Processed data: a coordinate file describing the final solved crystal structure stored in protein data bank .cif format as well as tabular data from enzymatic activity assays stored in either .csv or .xlsx files.

1A.9 Neurobehavioral Studies

Three major data types are collected for neurobehavioral experiments 1) behavioral data, 2) in vivo electrophysiological and optical recordings and 3) in vivo physiological signals derived from wireless telemetry:

1. Behavioral data: a project will typically produce behavioral data generated from a variety of experimental apparatuses that present specific stimuli to the animal and measure their responses. Data will be collected from an average of 48 mice generating a data set for each apparatus. Raw data will be processed to determine the behavior of the animal and statistical analyses will be performed on this processed data for publication. The processed data will require approximately 100 kilobytes to 1 gigabyte of data storage per apparatus. This data is usually stored in Excel format with all values from each animal across a single row, all animal identifiers and metadata in separate columns with a separate tab providing a key.
2. In vivo electrophysiological and/or optical recordings: a project will produce electrophysiological and/or optical recording data to measure neural activity during behavioral performance. Data will be collected from an average of 48 mice across multiple recording sessions. Raw data will be processed to determine the neural activity profile during each recording session. The processed data will require approximately 1 to 10 TB of data storage. Such data is typically stored in the following formats: .mat (matlab files), .pickle (binaries), .npz (binaries saved in numpy).
3. In vivo physiological telemetry data: a project will produce wireless telemetry data to measure continuous physiological activity. Data will be collected from an average of 48 mice across multiple recording sessions. Raw data will be processed to determine the physiological activity profile during the recording sessions. The processed data will require approximately 1 TB of data storage. This data is typically stored in .edf format representing a time series of physiological data.

1A.10 Whole-Genome DNA Sequencing of Human Subjects

A whole-genome sequencing (WGS) project will produce approximately 40 GB per sample and consist of the following types:

1. Raw data: sequenced reads in FASTQ format obtained using Illumina NextSeq/NovaSeq instruments. Depending on instrument output, more than one FASTQ file or pair may be generated for each sample.
2. Processed data: alignment of reads to a reference genome will produce an output in CRAM format, calling of single-nucleotide variant (SNV) and small insertions and deletions (indels) will produce an output in GVCF format, and joint genotyping of SNVs/small indels across a cohort as well as jointly called large structural variants (e.g. duplications, inversions, translocations) will produce outputs in VCF format. A joint tab-delimited text matrix describing a variety of sequencing quality metrics will be derived from all of the outputs listed above. Tab-delimited text-based annotations will also be generated for all jointly called variant loci.

1A.11 Single-Cell RNA Sequencing

Single-cell RNA-sequencing projects (scRNA-seq) generate approximately 9 GB of data per sample and consist of the following types:

1. Raw data: generated using an Illumina instrument (Next-seq/Nova-seq) from barcoded libraries prepared using the 10X genomics single cell 3’ Standard or Low Throughput methodology. Sequencing will be performed to capture roughly 1000 cells per sample and 100k reads/cell. The files produced will be stored in compressed FASTQ format.

2. Processed data: QC-filtered per cell outputs, specifically “barcodes.tsv.gz”, ”features.tsv.gz”, and “matrix.mtx.gz”. Further analysis of these files will produce a filtered count matrix which is used to generate visualization of cell clusters and supporting plain text files.

1A.12 Illumina RNA Sequencing

Illumina RNA sequencing (RNA-seq) projects produce data obtained using NextSeq and NovaSeq instruments consisting of approximately 12 GB per sample of the following types:

1. Raw data: sequenced reads in FASTQ format. Depending on instrument output, more than one FASTQ file or pair may be generated for each sample.
2. Processed data: aligned reads in BAM format, tab-delimited text matrices of read counts associated with genes, transcripts, or other genomic features, bigWig or bedGraph format read coverage files, and tab-delimited text results of differential expression analyses.

1A.13 Nanopore RNA Sequencing

Nanopore projects produce direct RNA sequencing data generated by the Oxford Nanopore Technologies (ONT) GridION instrument approximately 160 GB in size and consist of the following types:

1. Raw data: FAST5 files containing raw signal data that can be used for base-calling and calling poly(A) tail length and base-called read data available in FASTQ format.
2. Processed data: BAM files resulting from read alignment, read count tables, poly(A) tail length tables and differential gene expression tables. Read count data, differential expression tables and poly(A) tail length determination tables will be presented a tabular format such as Excel spreadsheets.

1A.14 Microarray Projects

These projects will produce microarray data generated/obtained from an Affymetrix GeneChip, Agilent SureScan, or Illumina iScan Microarray Scanner. Less than 50 MB will be produced per sample in the following formats:

1. Raw data: CHP/CEL (Affymetrix), TXT (Agilent), or IDAT (Illumina). Raw data will be normalized and exported to a more readable/portable format.
2. Processed data: TXT and XLSX files consisting of normalized probe intensities and the results of statistical analyses.

1A.15 Nanostring Projects

These projects will produce Nanostring data generated/obtained from a Nanostring Scanner. Less than 10 MB will be produced per sample in the following formats:

1. Raw data: RCC files which will be normalized and exported to a more readable/portable format.
2. Processed data: TXT and XLSX files consisting of normalized counts and the results of statistical analyses.

1A.16 Flow Cytometry

Flow cytometry projects generate data obtained from BD Fortessa, BD FACSAriaII, BD FACS Melody, BD Symphony S6, and Sony ec800 instruments. Roughly 1 GB of data is collected per experiment of the following types:

1. Raw data: instrument output stored in .fcs format.
2. Processed data: mean fluorescence intensities/percentages summarized in spreadsheets (.xlsx) and used for statistical analysis., multiple images in .pdf or .png formats.

1A.17 Animal Imaging

These projects will produce animal imaging data from two instruments:

1. Faxitron Bioptics Ultrafocus DXA: This instrument produces bone density and radiographs and creates images and tissue decomposition statistics. Images are saved as .jpg or .bmp and the measurements that are generated by built-in image analysis are saved as .txt files. Sizes of individual jpg files are 90 KB for radiographs and 800 KB for each bone density image. The resulting dataset will include one .jpg image file per animal and a single file in .xlsx format containing the experimental measurements for all animals within the project and statistical analysis.
2. FUJIFILM VisualSonics Vevo Ultrasound Imager: This instrument collects ultrasound video/still images such as echocardiographs or pregnancy ultrasound. Files produced for each imaged animal are saved as .bimg (46,000 KB), .pimg (270 KB), .png (600 KB) and are stored in the instrument database. The resulting dataset will include one .png image file per animal and a single file in .xlsx format containing the experimental measurements for all animals within the project and statistical analysis

1A.18 Microscopy Imaging

These projects will produce imaging data obtained using brightfield, epifluorescence, confocal, lightsheet, or dSTORM microscopes. Data may be collected using one or a combination of instruments and the size will vary from 1 MB to 100 MB for small projects, 100 MB to 50 GB for medium-sized, and 50 GB to 5+ TB for large projects. Files are generated of the following types:

1. Raw data: file types vary by microscope manufacturer. These include: Zeiss (.lsm, .czi), Olympus (.oif, .oib), Leica (.lif), Nikon (.nd2), Aperio (.svs), Andor Dragonfly (.ims, .xml). Homebuilt microscopes may generate .tif as well as other formats.
2. Processed data: file types vary by the software used for processing. These include: ImageJ/FIJI (.tif), Metamorph (.nd2), Imaris (.ims, .xml), Arivis (.tif, .sis), Aivia (.tif, .py), Zeiss Zen (.czi), Huygens (.tif, .hdr, .ics, .ics2). Additional generic file formats include .bmp, .mp4, .mpg, and .avi.

1A.19 Biostatistics Collaborative Projects

These projects involve data analysis in collaboration with colleagues that generate data. As the data generators, our colleagues will distribute and document the data themselves according to FAIR principals and NIH policies. Final analysis code and data for any publication will be shared upon request where such data is 1) not in violation of participant consent and 2) not publicly available. This includes code, relevant metadata, and in cases where computational overhead necessitated retention, intermediate datasets.

1A.20 Biostatistics Methods/Software Projects

Scientific output of these research projects are new statistical methods and code. No scientific data as defined by NIH is generated by these projects, as such, their outputs do not fall under the scope of the DMS plan requirement.

1B. 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.

1B.1 Cryo-EM

Based on technical considerations, the following data produced in the course of a project will be shared:

1. Raw data files: An entire dataset can comprise a few thousand micrographs amounting to several TB. These datasets are of limited utility for secondary analysis with few exceptions (e.g. the development of methods). Therefore, raw data will be preserved locally at NIEHS for at least seven years from the time of publication and shared upon request but not deposited in a

repository. Selected datasets deemed useful for method development or required by scientific journals will be made available at the time of publication.

2. Maps: The final result of processing cryo-EM data produce at least one electron scattering density map. Maps will be made available at the time of publication.
3. Atomic Models: In many cases, the maps mentioned above will be interpreted as atomic models. These results will be made available at the time of publication.

1B.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

All raw metabolomic and lipidomic data will be preserved at NIEHS and all processed files used to generate published analysis results will be deposited in a publicly available repository.

1B.3 Mass Spectrometry - Quantitative Small Molecule Analysis

All raw quantitative small molecule data will be preserved at NIEHS and all processed files used to generate published analysis results will be deposited in a publicly available repository.

1B.4 Mass Spectrometry - Protein and Peptide Analysis

All raw protein and peptide data will be preserved at NIEHS and all processed files used to generate published analysis results will be deposited in a publicly available repository.

1B.5 Molecular Dynamics Simulations

Due to the large data volume, raw molecular dynamics data cannot be shared via public repositories. All trajectory coordinate data will be preserved locally at NIEHS for further reuse and/or re-analysis. Starting, final, or representative sample coordinates files from MD simulations in .pub format will be released to the public upon publication. Additional file types (e.g. coordinate files or calculated energy data files) will be shared upon request.

1B.6 NMR – Metabolomics

Metadata, Fourier transformed NMR, and tabular analysis using Metabolomics Workbench, (https://www.metabolomicsworkbench.org) will be shared publicly upon publication. All statistical analyses will be reported in accompanying publications. Additional raw data will be preserved at NIEHS and provided upon request.

1B.7 NMR – Structural Biology

The final structural product data will be submitted to the Protein Data Bank (PDB) (www.rcsb.org). For NMR structures, the PDB also requires submission of analyzed atomic chemical shift data to the BioMagResBank (BMRB) (https://bmrb.io), and a list of NMR experiments used, but not the data due to the large file size. Project metadata and raw data will be stored and backed up at NIEHS and available upon request.

1B.8 X-ray Crystallography

All raw diffraction data is preserved at NIEHS. All .mtz files that used to generate coordinate files will be deposited in the publicly available Protein Database, PDB (https://www.rcsb.org/) released to the public upon publication of the associated manuscript.

1B.9 Neurobehavioral Studies

All data collected will be preserved on local NIEHS servers, however due to the technical constraints of storing extremely large files, we will only share the processed data (used for statistical analysis) on public servers.

1B.10 Whole-Genome DNA Sequencing of Human Subjects

All previously described processed data will be preserved and shared. The raw FASTQ files are larger than lossless CRAM formatted alignments, even when compressed, and all original read data can be recovered from CRAM files if necessary. Further, large-scale human whole-genome sequencing is typically provided to repositories in CRAM format, and this is the usual delivery vehicle for such data when an external sequencing vendor is utilized. To ensure CRAM files associated with the project are easily usable by third party researchers and compatible with repositories, they will be generated based on the Broad Institute's hg38 resources available on the Google Cloud Platform (https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/). All preserved data will be shared by submission to the NCBI database of genotypes and phenotypes (dbGaP).

1B.11 Single-Cell RNA Sequencing

The major datasets used in scRNA-seq analysis are the original FASTQ files and the filtered raw per-gene read counts. These counts can be used as input by a variety of scRNA-seq analysis software including the widely utilized Seurat. All previously described data types, specifically FASTQ, the three files representing QC-filtered raw data, and the filtered count matrix will be publicly shared upon publication.

1B.12 Illumina RNA Sequencing

All previously described data, with the exception of the BAM formatted aligned reads, will be preserved and shared. These excluded files are large and contain the sequenced reads, which are also present in the FASTQ files, and it is these raw files that are typically provided to public repositories. Further, the alignments may be reproduced by a third party based on knowledge of the alignment algorithm utilized, its parameters or configuration, and the underlying reference genome assembly. Other derived, reproducible files are preserved for the convenience of external researchers, as these are relatively small and facilitate closer inspection or reanalysis with minimal investment of computational resources. All preserved data will be shared by submission to the NCBI Gene Expression Omnibus (GEO) repository which will in turn deposit raw sequencing files to the Sequence Read Archive (SRA).

1B.13 Nanopore RNA Sequencing

All processed data produced in the course of a project will be preserved and shared to enable analysis, validation, and further reuse. The raw data is too large to be shared via public repository and will be archived on site.

1B.14 Microarray Projects

All data produced in the course of these projects will be preserved and shared.

1B.15 Nanostring Projects

All data produced in the course of these projects will be preserved and shared.

1B.16 Flow Cytometry

All raw data will be preserved locally at the NIEHS. It is not a common practice in the flow cytometry field to share the raw data. The summarized data, statistical analysis and representative images will be shared in the associated publication. For data generated from human participant samples, only deidentified data will be made available for sharing.

1B.17 Animal Imaging

Due to technical limitations for both instruments, raw data images will be preserved locally at the NIEHS and only .jpg and .png images and compiled experimental measurements for the project and statistical analysis in .xlsx file format will be shared.

1B.18 Microscopy Imaging

Imaging data will be stored on NIEHS servers or computers in PI labs. Finalized data will be shared upon publication according to the specific journal's publication criteria.

1B.19 Biostatistics Collaborative Projects

N/A

1B.20 Biostatistics Methods/Software Projects

N/A

1C. Metadata, other relevant data, and associated documentation: *Briefly list 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. Indicate if none.*

1C.1 Cryo-EM

There are two types of metadata associated with cryo-EM projects:

1. Metadata needed for the interpretation of the electron density maps and atomic models resulting from image processing. These are usually required by repositories and publications and will be shared along with the relevant results.
2. Metadata pertaining to the computational protocol used during the processing of raw data. These are not useful unless shared together with raw datasets. The later will be stored as the raw data and shared upon request.

1C.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

All relevant metabolomic and lipidomic metadata such as sample preparation methods, acquisition methods, equipment used, software and version numbers, etc., will be collected and disseminated as part of the publicly accessible description of the dataset.

1C.3 Mass Spectrometry - Quantitative Small Molecule Analysis

All relevant quantitative small molecule metadata such as sample preparation methods, acquisition methods, equipment used, software and version numbers, etc., will be collected and disseminated as part of the publicly accessible description of the dataset.

1C.4 Mass Spectrometry - Protein and Peptide Analysis

All relevant protein and peptide metadata such as sample preparation methods, acquisition methods, equipment used, software and version numbers, protein databases, database search methods and parameters, etc., will be collected and disseminated as part of the publicly accessible description of the dataset.

1C.5 Molecular Dynamics Simulations

Detailed methods will be included in the publication associated with each data set.

1C.6 NMR – Metabolomics

Accessible metadata will be comprised of the following elements recommended by The NMR Interest Group within the Metabolomics Association of North America (MANA), https://metabolomicsna.org/index.php/mana-interest-groups/nmr-metabolomics:

1. Number and type of experimental groups, number of replicates, number of controls
2. Instrumentation details: manufacturer, software, spectrometer, nuclei, NMR probe, automation used
3. Sample data: pH, solvent, temperature, reference standard, sample handling to remove large biomolecules
4. QC strategy
5. Data processing including: NMR parameters, peak identification/binning, software
6. Statistical methods

1C.7 NMR – Structural Biology

Metadata associated with structural depositions are recorded in the PDB and BRMB submissions, which include: equipment, software, biomolecular sequence, species of origin, structural coordinates, chemical shift assignments, experiments performed, and a link to the publication in PUBMED (www.pubmed.gov). The PDB and BMRB provide quality checks for public comparison on the website.

1C.8 X-ray Crystallography

When depositing the coordinates into the PDB, all relevant metadata such as crystallization conditions, equipment used, software and version numbers, protein sequence, species used, species expressed in etc., is collected and can be found in the header of the .cif coordinate file that is released to the public upon publication.

1C.9 Neurobehavioral Studies

To facilitate interpretation of the data we will include all necessary metadata with each publicly shared data set. This includes parameters used in data collection, behavioral training, subject ID, and treatment group. The format of the data will be determined by the requirements of the specific public data repository.

1C.10 Whole-Genome DNA Sequencing of Human Subjects

To facilitate interpretation and reanalysis of the data, a description of the library preparation protocol and the full analytical pipeline, including software release versions and reference genome assembly, will be made accessible. Phenotypes, disease status, and other participant-specific metadata collected during a project will also be shared to the extent permitted by the informed consent for data reuse granted by each individual along with a full accounting of the type and scope of consent.

1C.11 Single-Cell RNA Sequencing

To facilitate interpretation of the data, a brief explanation of the experimental design, metadata, sequencer information, library construction kit, and code will be shared and associated with the relevant data sets. The brief explanation of the experimental design will include cell line used, timeline from seeding to treatment to collection, treatment doses, and single cell suspension protocol prior to library preparation. The metadata will include library name, cell line/strain, sex, organism, treatment and software version used for analysis. Custom annotations will be shared as a GTF file.

1C.12 Illumina RNA Sequencing

To facilitate interpretation and reanalysis of the data, the strain, cell line, or other identifiable origin of each sample will be provided as well as a brief description of all courses of treatment or variable conditions. A description of the library preparation protocol will also be made accessible along with the full analytical pipeline, including software release versions, reference genome assembly, and feature set (e.g. versioned gene annotation). Any custom feature sets utilized for a project will be provided alongside the data itself in a standard text-based format (e.g. GTF, BED) and any custom reference sequences will be provided in FASTA format.

1C.13 Nanopore RNA Sequencing

To facilitate interpretation of the data, metadata, protocols and data collection instruments will be shared and associated with the relevant datasets. The metadata will include library name, organism, tissue, cell type, genotype, treatment. Protocols will include cell growth protocol, treatment protocol, RNA extraction protocol and library construction protocol. The collection of data will include sequencer and flow cell type, and the software required for the different steps of the analysis, including base caller, mapping tool, poly(A) caller tool, and differential expression analysis tool. The name of the genome assembly used to map the reads will also be provided, together with the format of the output files.

1C.14 Microarray Projects

To facilitate interpretation of the data, metadata, including sample name, model organism, sample characteristics (e.g., tissue type, genotype, age), growth protocol, treatment protocol, extraction protocol, labeling protocol, hybridization protocol, scanning protocol, data processing and any other pertinent items will be shared and associated with the relevant datasets.

1C.15 Nanostring Projects

To facilitate interpretation of the data, metadata, including sample name, model organism, sample characteristics (e.g., tissue type, genotype, age), growth protocol, treatment protocol, extraction protocol, labeling protocol, hybridization protocol, scanning protocol, data processing and any other pertinent items will be shared and associated with the relevant datasets.

1C.16 Flow Cytometry

To facilitate interpretation of the data, experimental metadata, documentation including reagent information (antibodies, fluorophores used), protocols, information on data collection instrument types will be shared and associated with the relevant datasets.

1C.17 Animal Imaging

To facilitate the interpretation of data, metadata (animal ID, strain, sex, age, treatment, diet, parentage), and references to protocols will be shared.

1C.18 Microscopy Imaging

Metadata terms are specific to each project and include data for species, tissue, cell type, treatment, genotype, antibodies, and fluorescence tags, along with other specific reagents and experimental details.

1C.19 Biostatistics Collaborative Projects

N/A

1C.20 Biostatistics Methods/Software Projects

N/A

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. If so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed. Indicate if none.

2.1 Cryo-EM

Data generated during cryo-EM projects is accessible using a wide variety of software packages. Most are open source or freeware and can be run under all major operating systems. Appropriate tools vary depending on the type of data:

1. Raw data: Many open source packages allow for the manipulation and processing of raw data. Example include Relion, CisTEM, EMAN2, IMOD, Xmipp, Scipion, and Spider. Other popular packages require a license which is free for academic use (i.e. Cryosparc) while a few others require license fees (e.g. Imagic). Any in-house developed software will be made available through github and other software dissemination repositories. Significant computing resources are required to perform calculations on a typical raw data set using any of the tools described above.
2. Maps and models: Software tools necessary to interpret the electron density maps and associated atomic models can be run on a small consumer level laptop or desktop computer. Some applications are even capable of displaying data on cell phones and tablets. Examples of these packages include Chimera, ChimeraX , Coot and Pymol.

2.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

Most metabolomic and lipidomic data are produced using Q-Exactive or Orbitrap mass spectrometers and the Xcalibur or TraceFinder software. Initial data processing is performed using Compound Discoverer. All of these tools are commercial products available from ThermoFisher for a fee. If applicable, additional data processing may be performed using in-house written code (R language) via JupyterNotebooks. The R language and JupyterNotebooks are open-source software tools. Data acquired using the Bruker timsTOF Pro are processed using the commercial tool Metaboscape, available from Bruker for a fee, then further processed in R language via JupyterNotebooks.

2.3 Mass Spectrometry - Quantitative Small Molecule Analysis

Most quantitative and small molecule data are produced using Quantiva, Q-Exactive, or Orbitrap mass spectrometers and the Xcalibur or TraceFinder softwares from ThermoFisher which can be purchased for a fee. Additional datasets are generated using a 6500+ Q-Trap instrument from Sciex and the Analyst software or a Waters Q-ToF G2 and the MassLynx software. Both tools are available for a fee.

2.4 Mass Spectrometry - Protein and Peptide Analysis

Most protein and peptide data are produced using Q-Exactive mass spectrometers and the Xcalibur software. Additional data sets are generated using a MALDI-ToF instrument from Sciex and the 4000 Series Explorer software which is also available for a fee. Finally, some data are produced on a Waters Q-ToF G2 and the MassLynx software which is also available for a fee. Peaklists are generated with Mascot Distiller which is commercial software available for a fee from Matrix Science. Database searching is performed using Spectrum Mill from Agilent (fee), MASCOT from Matrix Science (fee or free over the web at https://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF), Proteome Discoverer from ThermoFisher (fee), Peaks from Bioinformatics Solutions Inc. (fee), or Protein Prospector from the University of California at San Francisco (free and over the web at https://prospector.ucsf.edu/prospector/mshome.htm).

2.5 Molecular Dynamics Simulations

MD trajectory calculations are performed using packages such as Amber (free software, https://ambermd.org/), Gromacs (free software, https://www.gromacs.org/), Tinker (free for non-commercial users, tinkertools.org), and NAMD (free for non-commercial users, http://www.ks.uiuc.edu/Research/namd/). Most analyses are performed through the cpptraj module of Amber and several in-house programs written by the staff in the computational chemistry core. QM/MM calculations are performed using gaussian and Q-chem in combination with Amber. Molecular docking calculations are performed using Autodock (free software, https://autodock.scripps.edu/download-autodock4/), Dock (free software, https://dock.compbio.ucsf.edu/Overview_of_DOCK/index.htm), and Gold. The resulting files can be opened and manipulated using the same programs.

2.6 NMR – Metabolomics

NMR spectrometer data is vendor-specific and typically needs to be Fourier transformed for analysis. Various publicly available products will process this data. Commercial vendors include MNova, Chenomx, Progenesis QI, Compound Discoverer, and Bruker. NIEHS favors Chenomx software (www.chenomx.com).

2.7 NMR – Structural Biology

NMR spectrometer data is vendor-specific and typically needs to be Fourier transformed for analysis to perform chemical shift assignment. The following publicly available programs are typically used: NMRPIPE, NMRDRAW, Sparky, NMRViewJ, and XEASY. Software for purchase includes NMRFx, and Topspin. NMR structure calculations can be performed on personal computers with publicly available programs like XPLOR-NIH or CYANA. Public web servers can also perform these calculations from various stages of analysis and include: https://nmrfam.wisc.edu/software/, and https://nmrtist.org.

2.8 X-ray Crystallography

Most .mtz reflection files are produced using the tools HKL2000 or HKL3000 which can be purchased for a fee. Some of this data is generated using XDS which is free to academic users. The structural refinement process uses PHENIX software, which is also free to academic users.

2.9 Neurobehavioral Studies

Any specialized analysis tools that are developed in-house will be made available free-of-charge via the NIEHS Github or other open source-code repositories. Commercially available analysis tools will be described in detail so that other researchers can purchase them if necessary and reproduce the steps we followed. For behavioral analysis data, these include: Ethovision (Commercial) , FiPhA (open source), Deep Lab Cut (open source); for electrophysiological and optical analysis: Plexon Software (commercial) Inscopix Software (commercial), CaImAn (open source); for physiological telemetry: Ponemah (commercial).

2.10 Whole-Genome DNA Sequencing of Human Subjects

Raw sequencing data may be accessed and converted to text-based FASTQ or binary BAM/CRAM using the SRA Toolkit, available from the NCBI Sequence Read Archive. The VCF and GVCF file formats utilized in these projects are text-based, but may be read and manipulated with a host of compatible analytical tools including the Genome Analysis Tool Kit (GATK), Picard tools, and VCF tools. These files may also be examined visually using the UCSC Genome Browser, IGV, or other genome browser applications. All described tools are freely available and open-source.

2.11 Single-Cell RNA Sequencing

Seurat (https://satijalab.org/seurat/) and other scRNA-seq analytical tools such as Scanpy (https://scanpy.readthedocs.io/en/stable/) are freely available software. Software and version used to analyze each dataset will be submitted as part of the metadata described previously. R (https://www.r-project.org/) code used to load and analyze data will be made available via github (https://github.com/). Seurat and R versions will be indicated in all submitted code.

2.12 Illumina RNA Sequencing

Raw sequencing data will be accessible and convertible to text-based FASTQ format using the SRA Toolkit, available from the NCBI Sequence Read Archive. Any bigWig format coverage tracks can be converted to text using utilities available from the UCSC Genome Browser, and the tracks may be viewed directly on a web-based or stand-alone genome browser application (UCSC, IGV). All described tools are freely available and open-source.

2.13 Nanopore RNA Sequencing

Raw data will be base-called using Guppy (https://timkahlke.github.io/LongRead_tutorials/BS_G.html), mapped to an appropriate transcriptome with Minimap2 (https://github.com/lh3/minimap2), used to determine the poly(A) tail length with Nanopolish (https://github.com/jts/nanopolish) and the subsequent processed dataset used for statistical analysis. All described software is freely available for download.

2.14 Microarray Projects

Affymetrix data will be made available in CHP and CEL formats, Agilent in TXT format, and Illumina in IDAT format. Processed data for all platforms will be stored in TXT or XLSX formats. CHP/CEL and IDAT require the use of specialized tools, such as Omicsoft Array Studio (CHP/CEL, available for a fee) and Illumina Genome Studio (IDAT, available from the manufacturer). Open-source R packages such as affy and illuminio are available for reading CHP/CEL and IDAT files, respectively. Subsequent statistical analysis can be completed using a variety of tools include Array Studio and the R package limma. Customized methods may be employed in the analysis of Illumina data which will be specified in the associated manuscript and any code made available upon publication.

2.15 Nanostring Projects

Nanostring data will be made available in RCC, TXT, and XLSX file formats. RCC files require the use of specialized tools, such as nSolver, available from the manufacturer, or the open-source R package nanostringr, to be accessed and manipulated.

2.16 Flow Cytometry

Summarized and representative data will be available as spreadsheets and images that do not require specialized tools. Raw .fcs data files can be accessed and analyzed using either instrument-specific manufacturer acquisition software, or a secondary analysis program such as FCSExpress,(https://denovosoftware.com/), both fee-based, or free software such as FlowJo (https://www.flowjo.com/) or FACS analysis libraries in R and/or Python. The exact versions of the software and the libraries used in each project will be listed along with the shared data. Code that was used to analyze specific datasets will be made available upon request or shared through GitHub with links provided in the publication.

2.17 Animal Imaging

No specialized tools will be needed to access or manipulate the shared data (spreadsheets and images).

For the Faxitron Bioptics Ultrafocus DXA, the raw data can be only viewed and analyzed using proprietary Bioptics Vision software that is tied to the instrument itself. For the FUJIFILM VisualSonics Vevo Ultrasound Imager, the raw data can be viewed and analyzed using proprietary Visualsonics Version 3.27.15251 on the instrument itself or on a desktop equipped with a fee-based license for this software.

2.18 Microscopy Imaging

Data can be analyzed using open-source tools such as ImageJ and FIJI, and commercial software such as Metamorph, Imaris, Arivis, Aivia, Zeiss Zen, and Huygens. All data generated will be able to be opened by ImageJ/FIJI through the Bioformats Importer Plugin.

2.19 Biostatistics Collaborative Projects

N/A

2.20 Biostatistics Methods/Software Projects

N/A

Element 3: Data Standards

For DMS plans associated with clinical protocols, answer Element 3A and 3B. Other plans should answer only Element 3B.

3A. Data Standards for Clinical Protocols - Common Data Elements (CDEs): *Describe what Common Data Elements (CDEs) will be used. Justify if CDEs are not used.*

3A.1 Cryo-EM

N/A

3A.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

N/A

3A.3 Mass Spectrometry - Quantitative Small Molecule Analysis

N/A

3A.4 Mass Spectrometry - Protein and Peptide Analysis

N/A

3A.5 Molecular Dynamics Simulations

N/A

3A.6 NMR – Metabolomics

N/A

3A.7 NMR – Structural Biology

N/A

3A.8 X-ray Crystallography

N/A

3A.9 Neurobehavioral Studies

N/A

3A.10 Whole-Genome DNA Sequencing of Human Subjects

The previously described preservation strategy complies with CDE standards implemented in dbGaP, the premier repository for controlled-access human DNA sequencing data. All phenotype information submitted to dbGaP will comply with the standard for Common Metadata Elements for Cataloging Biomedical Datasets.

3A.11 Single-Cell RNA Sequencing

N/A

3A.12 Illumina RNA Sequencing

N/A

3A.13 Nanopore RNA Sequencing

N/A

3A.14 Microarray Projects

N/A

3A.15 Nanostring Projects

N/A

3A.16 Flow Cytometry

N/A

3A.17 Animal Imaging

N/A

3A.18 Microscopy Imaging

N/A

3A.19 Biostatistics Collaborative Projects

N/A

3A.20 Biostatistics Methods/Software Projects

N/A

3B. Data Standards for All Plans: *State what additional common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources; 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 proposed research. Plans not associated with a clinical protocol should describe the use of Common Data Elements here, if applicable. Indicate if no consensus standards exist.*

3B.1 Cryo-EM

Data will be shared according to the standards established by the EMDB, EMPIAR, and PDB repositories.

3B.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

The proposed minimum reporting standards for chemical analysis from the Chemical Analysis Working Group (CAWG) of the Metabolomics Standards Initiative (MSI) (https://doi.org/10.1007%2Fs11306-007-0082-2) will be followed when possible.

3B.3 Mass Spectrometry - Quantitative Small Molecule Analysis

The proposed minimum reporting standards for chemical analysis from the Chemical Analysis Working Group (CAWG) of the Metabolomics Standards Initiative (MSI) (https://doi.org/10.1007%2Fs11306-007-0082-2) will be followed when possible.

3B.4 Mass Spectrometry - Protein and Peptide Analysis

Recommendations from the Human Proteome Organization Proteomics Standards Initiative will be adhered to when possible.

3B.5 Molecular Dynamics Simulations

Raw data generated within the molecular modeling core mainly come from commercial programs that have proprietary file standards. Most files exclusively contain the position information (coordinates) of atoms in the system and many million frames are used to collect such data that is written in binary format. Since no direct searchable information exists within these data files, data standards cannot be applicable for the types of data generated within the molecular modeling core and the vendors who provide the program modules control such standards.

3B.6 NMR – Metabolomics

Recommendations from the NMR interest group of MANA will be followed rigorously.

3B.7 NMR – Structural Biology

All coordinate files containing metadata are deposited in the PDB in the required .cif format which is easily searchable at https://www.rcsb.org/. PDB uses NCBI annotations for structure genomic location and for gene, protein, and species names. Note that PDB and BMRB and currently working on new NMR data reporting standards for structural biology. Recommendations from these international bodies will be followed rigorously.

3B.8 X-ray Crystallography

All coordinate files containing the metadata will be deposited in the PDB in the required .cif format which is easily searchable at https://www.rcsb.org/. PDB uses NCBI annotations for structure genomic location as well as gene, protein, and species names.

3B.9 Neurobehavioral Studies

Formal standards for neurobehavioral data have not yet been widely adopted. However, our data and other materials will be structured and described according to emerging best practices. Data will be stored in common and open formats, such as nwb (the open neurodata without borders standard), and open source binaries (.pickle). Information needed to make use of this data (e.g. the meaning of variable names, code, information about missing data, other metadata etc) will be recorded in repositories and readme files that will be accessible to the research team and will subsequently be shared alongside final datasets. Information about our research process, including the details of our analysis pipeline will be maintained contemporaneously, using detailed protocols and electronic lab notebooks. This information will be accessible to all members of the research team and will be shared alongside our data.

3B.10 Whole-Genome DNA Sequencing of Human Subjects

Submission of preserved data to dbGaP via their standardized forms ensures all common data elements will be well-organized, and highly discoverable/accessible.

3B.11 Single-Cell RNA Sequencing

To facilitate their efficient use, all data and materials will be structured and described using the FASTQ and CellRanger output file standards. FASTQ files and CellRanger mapped matrix files will be deposited to GEO and SRA. All three CellRanger outputs will be deposited ("barcodes.tsv.gz","features.tsv.gz", and "matrix.mtx.gz") to ensure the public has multiple entry points to different analysis tools. Details of the pipeline used to analyze raw data and Seurat filtered matrices to define clusters and marker genes will be maintained contemporaneously, using protocols and code described in the metadata. This information will be shared alongside our data via github.

3B.12 Illumina RNA Sequencing

The described preservation strategy complies with the MINSEQE specification (FGED Society - MINSEQE) and contains all common data elements described therein. Submission of preserved data to the Gene Expression Omnibus via their standardized forms ensures all common data elements will be well-organized, and highly discoverable/accessible.

3B.13 Nanopore RNA Sequencing

To facilitate their efficient use, all of the data and materials will be structured and described using the following standards: FAST5 files, FASTQ files, BAM files. Raw read data will be structured and described using the FAST5 standard, which has been widely adopted by the long read community. Base-called read and mapped read data will be structured and described using the FASTQ and BAM standards, which have been widely adopted in the sequencing community. Formal standards for differential gene expression, read count and poly(A) tail length data have not yet been widely adopted. However, data and other materials will be structured and described according to community best practices. Data will be stored in common and open formats, such as tabulated tables or Excel spreadsheets. Information needed to make use of this data such as sample, condition, etc. will be recorded in associated metadata or in an additional tab in the Excel spreadsheet that will be accessible to the research team and will subsequently be shared alongside final datasets. Information about our research process, including the details of our analysis pipeline will be maintained contemporaneously, using protocols described in the metadata. This information will be accessible to all members of the research team and will be shared alongside our data. The data will be submitted to GEO; GEO submission procedures closely follow the MIAME and MINSEQE standards.

3B.14 Microarray Projects

Whenever possible, Gene Expression Omnibus data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources. Microarray data will be structured and described using the Gene Expression Omnibus standard, which has been widely adopted in the microarray, next-gen sequencing, and Nanostring communities. All data will be deposited into the Gene Expression Omnibus.

3B.15 Nanostring Projects

Whenever possible, Gene Expression Omnibus data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources. Nanostring data will be structured and described using the Gene Expression Omnibus standard, which has been widely adopted in the microarray, next-gen sequencing, and Nanostring communities. All data will be deposited into the Gene Expression Omnibus.

3B.16 Flow Cytometry

Several standards for flow cytometry data have been suggested including MIFlowCyt: the minimum information about a Flow Cytometry Experiment, DOI: 10.1002/cyto.a.20623. However, no format has been generally accepted, except for the open .fcs file format that is broadly used. For the ease of use the files will be named to include sample name and treatment and will be accompanied by a data dictionary listing pertinent metadata including, but not limited to file name, sample name, treatment, dose, duration or patient/animal condition, fluorophore and antibody labels, and antibody clones and isotype control clones.

3B.17 Animal Imaging

Formal standards for both data types have not yet been widely adopted. However, our data and other materials will be structured and described according to best practices including (but not limited to) experiment specific information such as dates, subject identification, treatments, measurement parameters, measurement types, genotypes, strains, ages, parental derivation, and sexes.

3B.18 Microscopy Imaging

Formal standards for imaging data have not yet been widely adopted. However, our data and other materials will be structured and described according to best practices. Imaging data will be stored in common and open formats compatible with ImageJ. Information needed to make use of this data will be recorded in lab notebooks that will be accessible to the research team and will subsequently be shared. Information about the research process, including the details of analysis pipelines will be maintained contemporaneously, using lab notebooks. This information will be accessible to all members of the research team and will be shared alongside the data.

3B.19 Biostatistics Collaborative Projects

N/A

3B.20 Biostatistics Methods/Software Projects

N/A

Element 4: Data Preservation, Access, and Associated Timelines

4A. 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. Note that submission of a study to ClinicalTrials.gov meets the requirements of FDAAA but does not fulfill the requirements of the Data Management and Sharing Policy.*

4A.1 Cryo-EM

Raw data for selected datasets deemed useful for method development or required by scientific journals will be made available through deposition with the Electron Microscopy Public Image Archive (EMPIAR, https://www.ebi.ac.uk/empiar/). Maps will be made available via deposition at the Electron Microscopy Data Bank (EMDB, https://www.ebi.ac.uk/emdb/). Atomic Models will be made available via deposition in the Protein Data Bank (PDB, https://www.rcsb.org/).

4A.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

All raw mass spectrometry data, appropriate mid-level processed data, tabular analysis, and associated metadata will be submitted to Metabolomics Workbench (https://www.metabolomicsworkbench.org).

4A.3 Mass Spectrometry - Quantitative Small Molecule Analysis

All raw mass spectrometry data, appropriate mid-level processed data, tabular analysis, and associated metadata will be submitted to Metabolomics Workbench (https://www.metabolomicsworkbench.org).

4A.4 Mass Spectrometry - Protein and Peptide Analysis

All raw data, appropriate mid-level processed data, final results and metadata for a published project will be deposited at the PRIDE data repository (https://www.ebi.ac.uk/pride/).

4A.5 Molecular Dynamics Simulations

All data generated for a published project will be stored locally at NIEHS. Per journal request, starting and final coordinate files may be shared via repositories maintained by the journal.

4A.6 NMR – Metabolomics

All NMR data, tabular analyses, and metadata will be deposited at Metabolomics Workbench, unless NIH establishes a standard alternative. Raw data and other metadata will be shared upon request.

4A.7 NMR – Structural Biology

All structural data and metadata for published projects will be deposited at the PDB https://www.rcsb.org/ and accompanying NMR chemical shift data and experiment lists will be found in the linked BMRB (www.bmrb.io). Raw data will be shared upon request.

4A.8 X-ray Crystallography

All structural data and metadata for published projects will be deposited at the PDB https://www.rcsb.org/. Any other data outside the structure determination that maybe generated and reported such as enzymatic activity data, will be deposited in the Figshare repository (https://figshare.com/).

4A.9 Neurobehavioral Studies

Behavioral data will be shared via the Mouse Phenome Database repository (phenome.jax.org) which provides metadata, persistent identifiers (i.e., DOIs), and long-term access. This database is maintained by Jackson Laboratory and funded by NIH. If the data is sufficiently relevant to chemical exposures it may also be stored on the NIEHS Chemical Effects in Biological Systems (CEBS) database (cebs.niehs.nih.gov). In vivo electrophysiological and optical recording data will be shared via the Distributed Archives for Neurophysiology Data Integration (dandiarchive.org). This database is managed via a distributed group of research institutions and academics and is funded by the BRAIN Initiative and NIMH. If the data is sufficiently relevant to chemical exposures it may also be stored in CEBS. In vivo physiology telemetry data will be shared via Physionet (physionet.org) and/or CEBS. Physionet is managed by the MIT Laboratory for Computational Physiology and funded by NIH.

4A.10 Whole-Genome DNA Sequencing of Human Subjects

Preserved data will be archived in the NCBI Database of Genotypes and Phenotypes (dbGaP).

4A.11 Single-Cell RNA Sequencing

All dataset(s) that can be shared will be deposited at the NIH GEO: https://www.ncbi.nlm.nih.gov/geo/ and SRA: https://www.ncbi.nlm.nih.gov/sra repositories.

4A.12 Illumina RNA Sequencing

Preserved data will be archived in the NCBI Gene Expression Omnibus and Sequence Read Archive repositories.

4A.13 Nanopore RNA Sequencing

Datasets resulting from this research will be shared via the NIH-supported GEO repository https://www.ncbi.nlm.nih.gov/geo/.

4A.14 Microarray Projects

All dataset(s) that can be shared will be deposited in Gene Expression Omnibus.

4A.15 Nanostring Projects

All dataset(s) that can be shared will be deposited in Gene Expression Omnibus.

4A.16 Flow Cytometry

It is not a common practice in the flow cytometry field to share raw data. The summarized data, statistical analysis and representative images will be shared in the associated publication, the raw data will be shared upon request. However, there are examples of repositories that can be used to share these data such as Flow Cytometry File Repository (https://flowrepository.org/) that is supported by the International Society for Advancement of Cytometry (ISAC) and provides persistent identifiers for the data.

4A.17 Animal Imaging

The measurement data, statistical analysis, and representative images will be shared as part of a publication in the main publication or in supplemental material. Additional study data will be shared via the Figshare repository.

4A.18 Microscopy Imaging

There are currently no large, open, public repositories for imaging data. Therefore, data can be only shared using a generalist repository such as Dryad, which provides metadata, persistent identifiers (i.e., DOIs), and long-term access. Dryad is the institutional data repository supported by the University of California and all data is shared under a CC0 waiver, which makes the dataset(s) publicly available. Dryad datasets are backed up to Merritt, the UC's CoreTrustSeal-certified digital repository, for long-term storage and accessibility. Procedures in place to ensure dataset preservation include storage of data files in multiple geographic locations, regular audits for fixity and authenticity, and succession plans in the event of repository closure. Imaging data will be stored internally at NIEHS and finalized data for publication will be uploaded and stored according to the journal's criteria.

4A.19 Biostatistics Collaborative Projects

N/A

4A.20 Biostatistics Methods/Software Projects

N/A

4B. 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.*

4B.1 Cryo-EM

Each structure submitted to EMDB, EMPIAR, and PDB is assigned a public accession ID that serves as a permanent link to the data. The data are searchable on multiple parameters such as author, protein, ligand, conditions, sequence, etc. There are also cross-references between the submissions in these databases.

4B.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

All datasets at the Metabolomics Workbench data repository receive a unique web accessible DOI. This website allows the public to search the metadata for many different items based on given search parameters, including authors, molecules, conditions, etc.

4B.3 Mass Spectrometry - Quantitative Small Molecule Analysis

All datasets at the Metabolomics Workbench data repository receive a unique web accessible DOI. This website allows the public to search the metadata for many different items based on given search parameters, including authors, molecules, conditions, etc.

4B.4 Mass Spectrometry - Protein and Peptide Analysis

Each dataset submitted to PRIDE is assigned a nine-character alphanumeric identifier called the PRIDE dataset identifier that serves as a permanent link to the data. This website allows the public to search the metadata for many different items based on given search parameters, including authors, proteins, ligands, conditions, sequence etc.

4B.5 Molecular Dynamics Simulations

The only publicly available molecular dynamics simulation data will be associated with publications and findable via Pubmed or Pubmed Central, both of which provide a persistent identifier, PMID or PMCID, respectively. All raw data will be stored and managed locally at NIEHS by the staff in the computational chemistry and molecular modeling core.

4B.6 NMR – Metabolomics

Metabolomics Workbench is searchable via metadata. Deposited data sets will receive a unique web accessible DOI.

4B.7 NMR – Structural Biology

All structural data will be searchable in the PDB https://www.rcsb.org/. This website can search the metadata for many different items based on given search parameters, including authors, proteins, ligands, conditions, sequence etc. Each structure submitted to PDB is assigned a 4-character alphanumeric identifier called the PDB identifier or PDB ID accession number that serves as a permanent link to the data. The PDB also contains a link to the publication in PUBMED.

4B.8 X-ray Crystallography

All structural data is searchable in the PDB (https://www.rcsb.org/). This website can search the metadata for many different items based on given search parameters, including authors, proteins, ligands, conditions, sequence etc. Each structure submitted to PDB is assigned a 4-character alphanumeric identifier called the PDB identifier or PDB ID accession number that serves as a permanent link to the data. Each file or file collection submitted to Figshare.com is assigned a unique digital identifier, DOI.

4B.9 Neurobehavioral Studies

The tools described in Sections 3B and 4A will provide sufficient metadata and unique identifiers for each study and subject to ensure data is findable and identifiable via DOIs.

4B.10 Whole-Genome DNA Sequencing of Human Subjects

Preserved data will be discoverable within the dbGaP repository and accession numbers will be provided within all published manuscripts to facilitate direct access to interested readers.

4B.11 Single-Cell RNA Sequencing

GEO submission requires metadata including experimental design, sequencing platform and associated publication. GEO links SRA data via the associated Accession number and this unique number will be described in the publication. The data will be searchable directly through GEO which allows searching of the metadata for key words or accession number.

4B.12 Illumina RNA Sequencing

Preserved data will be discoverable within the GEO and SRA repositories using key-word based search engines and accession numbers will be provided within all published manuscripts to facilitate direct access to interested readers.

4B.13 Nanopore RNA Sequencing

The GEO issues accession numbers to submitted files and datasets to be used as data reference or citation. GEO supports metadata-based searching.

4B.14 Microarray Projects

The Gene Expression Omnibus (GEO) provides metadata, persistent identifiers for data/file types (microarray file formats, CEL, CHP, IDAT), and long-term access. This repository is supported by NIH and dataset(s) are publicly available with a GEO accession number using the GEO website (https://www.ncbi.nlm.nih.gov/geo/).

4B.15 Nanostring Projects

The Gene Expression Omnibus (GEO) provides metadata, persistent identifiers for data/file types (Nanostring file formats, RCC, TXT, and XLSX), and long-term access. This repository is supported by NIH and dataset(s) are publicly available with a GEO accession number using the GEO website (https://www.ncbi.nlm.nih.gov/geo/).

4B.16 Flow Cytometry

Data will be preserved and shared locally at the NIEHS. Folder structure and ELN records will enable data findability.

4B.17 Animal Imaging

Data shared via publication will be findable by Pubmed ID using the persistent PMID identifier. Data shared via Figshare will be accessible using a DOI assigned by the repository. Both PubMed and Figshare are searchable using keywords and structured search.

4B.18 Microscopy Imaging

Dryad provides metadata-based search and persistent identifiers (i.e., DOIs). Internal image data storage will be documented in lab notebooks and linked to the experiments that generated the data. The published data will be stored and tagged according to the journal’s criteria.

4B.19 Biostatistics Collaborative Projects

N/A

4B.20 Biostatistics Methods/Software Projects

N/A

4C. When and how long the scientific data will be made available: *Describe when the scientific data will be made available to other users, and if known, for how long data will be available. Scientific data must be made available no later than the time of an associated publication (when the publication first appears, either online or in print), or by the end of the project/protocol, whichever comes first.*

4C.1 Cryo-EM

All data required to reproduce and validate experimental results associated with published findings will be preserved locally at NIEHS for at least seven years. Shared data will be made available at the time of associated publication and will remain available indefinitely.

4C.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

All raw mass spectrometry data will be preserved at NIEHS indefinitely. Shared data will be made available upon acceptance of the manuscript and remain available indefinitely.

4C.3 Mass Spectrometry - Quantitative Small Molecule Analysis

All raw mass spectrometry data will be preserved at NIEHS indefinitely. Shared data will be made available upon acceptance of the manuscript and remain available indefinitely.

4C.4 Mass Spectrometry - Protein and Peptide Analysis

All raw mass spectrometry data will be preserved at NIEHS indefinitely. Shared data will be made available upon acceptance of the manuscript and remain available indefinitely.

4C.5 Molecular Dynamics Simulations

Data will be made available upon acceptance of the manuscript and remain available indefinitely.

4C.6 NMR – Metabolomics

Data will be made available upon acceptance of the manuscript and remain available indefinitely.

4C.7 NMR – Structural Biology

Data will be made available upon acceptance of the manuscript and remain available indefinitely.

4C.8 X-ray Crystallography

Data will be made available upon acceptance of the manuscript and remain available indefinitely.

4C.9 Neurobehavioral Studies

Data will be made available as soon as possible or at the time of associated publication, whichever comes first.

4C.10 Whole-Genome DNA Sequencing of Human Subjects

Data will be made available to journals and peer reviewers during the manuscript submission and revision process, then to the general public on the date of publication. Projects that meet the definition of a large-scale genomic data set will be made publicly available on the timelines specified by the NIH Genomic Data Sharing Policy (https://osp.od.nih.gov/wp-content/uploads/Supplemental_Info_GDS_Policy.pdf).

4C.11 Single-Cell RNA Sequencing

Data will be made available as soon as possible or at the time of associated publication, whichever comes first.

4C.12 Illumina RNA Sequencing

Data will be made available to journals and peer reviewers during the manuscript submission and revision process, then to the general public on the date of publication. Projects that meet the definition of a large-scale genomic data set will be made publicly available on the timelines specified by the NIH Genomic Data Sharing Policy (https://osp.od.nih.gov/wp-content/uploads/Supplemental_Info_GDS_Policy.pdf).

4C.13 Nanopore RNA Sequencing

Data will be made available as soon as possible or at the time of associated publication, whichever comes first.

4C.14 Microarray Projects

Data will be made available as soon as possible or at the time of associated publication, whichever comes first.

4C.15 Nanostring Projects

Data will be made available as soon as possible or at the time of associated publication, whichever comes first.

4C.16 Flow Cytometry

Summary data and statistical analyses will be made available at the time of associated publication. Raw data will be preserved locally at least for 7 years.

4C.17 Animal Imaging

Raw image data will be preserved locally at NIEHS for at least 7 years. Tabular data will be made available at the time of associated publication and will be available indefinitely.

4C.18 Microscopy Imaging

Representative data will be made available at the time of associated publication and will stay available indefinitely. Full datasets will be preserved locally for at least 7 years.

4C.19 Biostatistics Collaborative Projects

N/A

4C.20 Biostatistics Methods/Software Projects

N/A

Element 5: Access, Distribution, or Reuse Considerations

5A. 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.*

5A.1 Cryo-EM

All data contained in the EMDB, EMPIAR, and PDB archive are available under the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. There are no limitations for subsequent access, distribution, or reuse.

5A.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

The content of the Metabolomics Workbench is protected by international copyright, trademark and other laws. The public may download articles and web pages from this site for their personal, non-commercial use only, provided that they keep intact all authorship, copyright and other proprietary notices. If one uses the Metabolomics Workbench, they accept these terms. The NMDR reserves the right to modify these terms at any time.

5A.3 Mass Spectrometry - Quantitative Small Molecule Analysis

The content of the Metabolomics Workbench is protected by international copyright, trademark and other laws. The public may download articles and web pages from this site for their personal, non-commercial use only, provided that they keep intact all authorship, copyright and other proprietary notices. If one uses the Metabolomics Workbench, they accept these terms. The NMDR reserves the right to modify these terms at any time.

5A.4 Mass Spectrometry - Protein and Peptide Analysis

All data contained in the PRIDE archive are available under the terms of the Creative Commons Public Domain (CC0) License. There are no limitations for subsequent access, distribution, or reuse. Following good scientific practices, users that reuse the deposited datasets in PRIDE are recommended to give appropriate credit to the original authors/submitters by citing the original dataset or the publication including the dataset.

5A.5 Molecular Dynamics Simulations

There are no limitations for subsequent access, distribution, or reuse.

5A.6 NMR – Metabolomics

The content of the Metabolomics Workbench is protected by international copyright, trademark and other laws. The public may download articles and web pages from this site for their personal, non-commercial use only, provided that they keep intact all authorship, copyright and other proprietary notices. If one uses the Metabolomics Workbench, they accept these terms. The NMDR reserves the right to modify these terms at any time.

5A.7 NMR – Structural Biology

All data contained in the PDB archive are available under the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. There are no limitations for subsequent access, distribution, or reuse.

5A.8 X-ray Crystallography

All data contained in the PDB archive are available under the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. There are no limitations for subsequent access, distribution, or reuse. All data submitted to Figsare.com will be published under CC-BY 4 license that doesn’t limit data reuse but requires source attribution.

5A.9 Neurobehavioral Studies

Technical limitations of public repositories limit the sharing of raw data files due to their large size. All processed data will be made available to the fullest extent possible.

5A.10 Whole-Genome DNA Sequencing of Human Subjects

Access to data and the scope of its reuse by third parties will be limited by the degree of consent granted by each individual participant.

5A.11 Single-Cell RNA Sequencing

The data does not contain any sensitive information and will be shared without restrictions.

5A.12 Illumina RNA Sequencing

No limitations on access, distribution, or reuse are anticipated for data generated as part of such projects.

5A.13 Nanopore RNA Sequencing

The data doesn’t include any sensitive or protected information and will be shared with no restrictions or use limitations.

5A.14 Microarray Projects

N/A

5A.15 Nanostring Projects

N/A

5A.16 Flow Cytometry

There are no limitations for subsequent access, distribution, or reuse of these data.

5A.17 Animal Imaging

There are no factors limiting data access, distribution, or reuse.

5A.18 Microscopy Imaging

Data access, distribution and reuse will require attribution in accordance with journal policy.

5A.19 Biostatistics Collaborative Projects

N/A

5A.20 Biostatistics Methods/Software Projects

N/A

5B. Will access to scientific data be controlled: *State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).*

5B.1 Cryo-EM

Access will be controlled until publication of the submitted manuscript reporting the data. At that point it will be released for public access.

5B.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

Access will be controlled until publication of the submitted manuscript reporting the data. At that point it will be released for public access.

5B.3 Mass Spectrometry - Quantitative Small Molecule Analysis

Access will be controlled until publication of the submitted manuscript reporting the data. At that point it will be released for public access.

5B.4 Mass Spectrometry - Protein and Peptide Analysis

Access will be controlled until publication of the submitted manuscript reporting the data. At that point it will be released for public access.

5B.5 Molecular Dynamics Simulations

Upon manuscript publication, the data will be released for public access, upon request.

5B.6 NMR – Metabolomics

After publication, the data will be released for public access with no access control.

5B.7 NMR – Structural Biology

After publication, the data will be released for public access with no access control.

5B.8 X-ray Crystallography

Access will be controlled until publication of the submitted manuscript reporting the data. At that point it will be released for public access.

5B.9 Neurobehavioral Studies

Data will be made available as soon as possible or at the time of associated publication, whichever comes first.

5B.10 Whole-Genome DNA Sequencing of Human Subjects

Access to preserved data will be controlled and mediated through dbGaP in accordance with documented consent granted by each project participant.

5B.11 Single-Cell RNA Sequencing

Access to the data will not be controlled.

5B.12 Illumina RNA Sequencing

Access to preserved data will not be controlled.

5B.13 Nanopore RNA Sequencing

The data do not require access control.

5B.14 Microarray Projects

Access to the data will not be controlled.

5B.15 Nanostring Projects

Access to the data will not be controlled.

5B.16 Flow Cytometry

Access will not be controlled. Personal identification will not be possible due to sharing of only de-identified data.

5B.17 Animal Imaging

The access to scientific data will not be controlled.

5B.18 Microscopy Imaging

Data access will not be controlled.

5B.19 Biostatistics Collaborative Projects

N/A

5B.20 Biostatistics Methods/Software Projects

N/A

5C. 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).

5C.1 Cryo-EM

N/A

5C.2 Mass Spectrometry - Metabolomic and Lipidomic Analysis

The collective mass spectrometry groups at the NIEHS use only de-identified data from scientific collaborators. PHI or PII cannot be recovered or inferred from the raw data. Compliance with the Institutional Review Boards, and PII is the primary responsibility of the principal investigator.

5C.3 Mass Spectrometry - Quantitative Small Molecule Analysis

The collective mass spectrometry groups at the NIEHS use only de-identified data from scientific collaborators. PHI or PII cannot be recovered or inferred from the raw data. Compliance with the Institutional Review Boards, and PII is the primary responsibility of the principal investigator.

5C.4 Mass Spectrometry - Protein and Peptide Analysis

The collective mass spectrometry groups at the NIEHS use only de-identified data from scientific collaborators. PHI or PII cannot be recovered or inferred from the raw data. Compliance with the Institutional Review Boards, and PII is the primary responsibility of the principal investigator.

5C.5 Molecular Dynamics Simulations

N/A

5C.6 NMR – Metabolomics

The NMR core facility uses only de-identified data from scientific collaborators. Compliance with the Institutional Review Boards, and PII is the primary responsibility of the principal investigator.

5C.7 NMR – Structural Biology
N/A

5C.8 X-ray Crystallography
N/A

5C.9 Neurobehavioral Studies
N/A

5C.10 Whole-Genome DNA Sequencing of Human Subjects
No personally identifiable information will be shared or provided to public data repositories for any project participant. Participant privacy and confidentiality will be protected through de-identification of all generated or collected data and the use of unique randomized alphanumeric sample identifiers.

5C.11 Single-Cell RNA Sequencing
N/A

5C.12 Illumina RNA Sequencing
These projects include no human research participants, as such, no protections for privacy, rights, or confidentiality are applied to the preserved data.

5C.13 Nanopore RNA Sequencing
N/A

5C.14 Microarray Projects
N/A

5C.15 Nanostring Projects
N/A

5C.16 Flow Cytometry
For patient data, all labels had been deidentified and the data do not contain intrinsic features enabling reidentification.

5C.17 Animal Imaging
N/A

5C.18 Microscopy Imaging
N/A

5C.19 Biostatistics Collaborative Projects
N/A

5C.20 Biostatistics Methods/Software Projects
N/A

Element 6: Oversight of Data Management and Sharing
Compliance with this plan will be monitored and managed by the Scientific Director (or designee). (No additional information is required).

Element 7: Other Elements (if applicable)
Include IC-specific or other additional information here. Indicate if not applicable.
N/A