University of Idaho
Library

Elements of the DMSP

December 8, 2022

Six Elements

- Data Type
- Related Tools, Software, and/or Code
- Standards
- Data Preservation, Access, and Associated Timelines
- Access, Distribution, or Reuse Considerations
- Oversight of Data Management and Sharing

Definition of Data

Scientific data is the recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications.

Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.

Data Type

Summary of the types and estimated amount of data to be generated or used in the research

- Modality: imaging, genomic, mobile, survey, etc.
- Aggregation: individual, aggregated, summarized
- Degree of Processing: how raw or processed the data will be

Description of which data will be preserved and shared.

> “NIH does not anticipate that researchers will preserve and share **all** scientific data generated in a study. Researchers should decide which scientific data to preserve and share based on ethical, legal, and technical factors that may affect the extent to which scientific data are preserved and shared. Provide the rationale for these decisions.”

Data Type Example, part 1

“This project will produce sequencing, snRNAseq (transcriptomic), snATACseq (epigenetic), and WGS data generated/obtained from 10x snRNAseq via the Chromium or Visium platform on Illumnia devices from patients from the Knight ADRC.

Data will be collected from 70 research specimens, generating 280 datasets totaling approximately 8400 Gb in size (8.4 Tb).

The data files will be used or produced in the course of the project includes: comma and tab separated files (csv or tsv), fastq sequencing files, expression matrixes and barcodes (gz), and R code (R).

Raw data will be transformed by our snRNAseq pipeline and the subsequent processed dataset used for statistical analysis and machine learning. To protect research participant and family member identities, only the de-identified individual data will be made available for sharing.”

Data Type Example, part 2

“Based on technical considerations, the following data produced in the course of the project will be preserved and shared: Raw sequencing files, data that has been validated for quality, all processed data generated from the raw sequencing files, and the associated code used to process the files.”

Data Type Example, part 3

“To facilitate interpretation of the data, clinical metadata, biospecimen metadata, assay metadata, code and readmes will be shared and associated with the relevant datasets.

The clinical metadata will include persistent unique identifies, information on individual donors, such as sex, ethnicity, age at death, post mortem interval, clinically significant measures (cognitive assessment scores, NIA-Reagan score, braak stage, CDR, case/control status, APOE genotype, years education, CERAD score) and other any other relevant neuropathology data.

Biospecimen metadata for brain samples includes: specimen IDs, tissue source, Brodmann area, and sample status information.

The assay metadata includes information about the platform, libraries generated, assay, sequencing batch, and valid barcode reads. More specific metadata for the assays includes information gathered during the quality control (QC) process, including information on percent ribosomal or mitochondrial, Seurat score, clusters and sub-clusters of cells.”

Related Tools, Software and/or Code

Whether specialized software is needed to access or manipulate the shared data to support replication or reuse and the name of this software.

If applicable, provide:

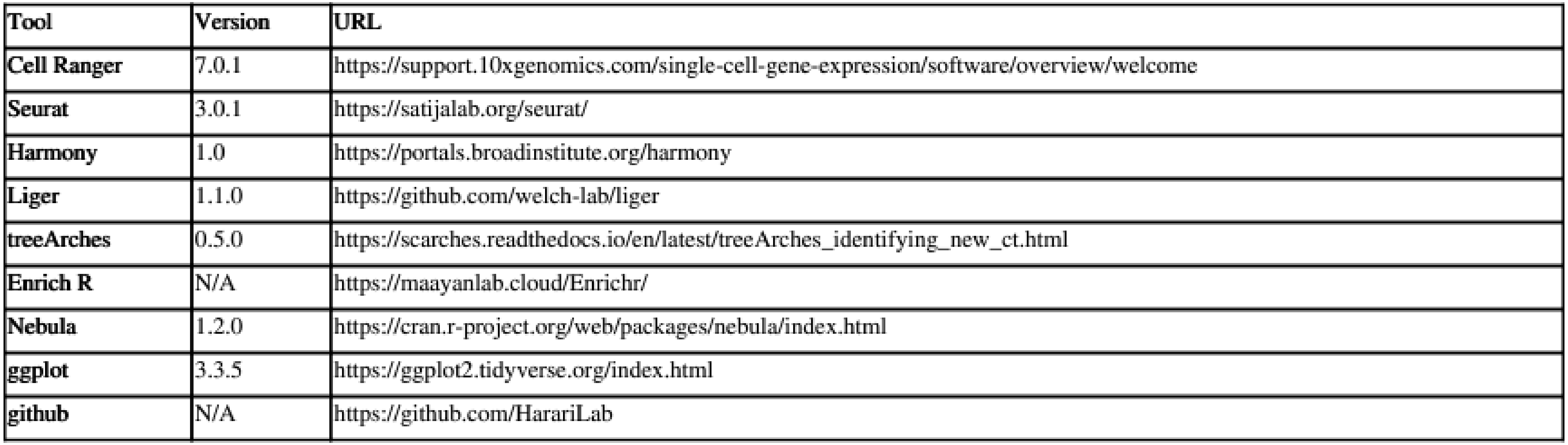
- Name of the tool
- How it can be accessed
- If known, how long will this likely be available

Related Tools… Example, part 1

“Raw sequencing files in fastq format will be made available and may require the use of programs or scripts including 10x Cell Ranger and Seurat to be manipulated. Metadata and processed sequencing data will be made available in csv format and will not require the use of specialized tools to be accessed or manipulated. Downstream analysis and visualization will be made available in csv format and images that require no specialized tools to access. Our complete analysis pipeline will be available on github.”

Related Tools… Example, part 2

“All tools are expected to remain publically available as long as the data remains available. The following tools are all available free of charge. Cell Ranger is proprietary software that is licenced by 10x genomics, all others are open source.”



Tool	**Version**	**URL**
Cell Ranger	7.0.1	https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome
Seurat	3.0.1	https://satijalab.org/seurat/
Harmony	1.0	https://portals.broadinstitute.org/harmony
Liger	1.1.0	https://github.com/welch-lab/liger
treeArches	0.5.0	https://scarches.readthedocs.io/en/latest/treeArches_identifying_new_ct.html
Enrich R	N/A	https://maayanlab.cloud/Enrichr/
Nebula	1.2.0	https://cran.r-project.org/web/packages/nebula/index.html
ggplot	3.3.5	https://ggplot2.tidyverse.org/index.html
github	N/A	https://github.com/HarariLab

Standards

What standards will be applied to the data and metadata? Standards include, but are not limited to:

- Data formats
- Data dictionaries
- Data identifiers
- Other types of documentation

If no “consensus” standards exist, make this clear.

Standards

“The FAIR Data sharing protocols will be applied, so that the data will be Findable, Accessible, Interoperable, and Re-usable. Our sequencing data will be structured and described using the following description of standards:

All shared studies contain 1) The description of the biological system, samples, and the experimental variables being studied, 2) The sequence read data for each assay, 3) The ‘final’ processed (or summary) data for the set of assays in the study, 4) General information about the experiment and sample-data relationships, and 5) Essential experimental and data processing protocols, 6) Metadata appropriate to the datasets so that they can be linked.

Similar protocols will be followed for all proteomic, epigenetic, and genomic data that is generated in the course of this project. The data formats of fastq files, csv or tsv files, and R code are standard across data repositories that store sequencing data.”

Data Preservation, Access, and Timelines

Where the data/metadata will be archived (e.g. repository name)

How the data will be findable and identifiable (e.g. use of a DOI number or other identifier)

When the data will be available

- NIH preference: ASAP, or no later than:
  a) time of publication of an associated paper, or
  b) the end of the period of performance

For how long the data will be available

- Researchers' best estimate based on anticipated value, identify differences if subsets of the data have differing timelines

Data Preservation... Example

“All dataset(s) that can be shared will be deposited in the National Institute on Aging Genetics of Alzheimer's Disease Data Storage (NIAGADS) repository.

The NIAGADS provides metadata, persistent identifiers (accession numbers), and long-term access. This repository is supported by the NIA and datasets are available through a request process for qualified investigators, and requires signatures on several sharing agreements, an intended use statement, and verification of IRB approval.

The data will be made available as soon as possible or at the start of the publication process, whichever comes first. Except for the case that data need to be removed (i.e. a participant withdraws from a study and requests their data be destroyed), NIAGADS data are managed indefinitely and available for data request. If NIAGADS funding is discontinued, NIAGADS will host the data and website for one more year before arranging for the data to be hosted at other qualified access repositories such as dbGaP.”

Access, Distribution, and Reuse Considerations

Describe and applicable factors affecting access, distribution or reuse of data related to:

- Informed consent
- Privacy and confidentiality (i.e., de-identification, Certificates of Confidentiality, and other protective measures) consistent with applicable federal, Tribal, state, and local laws, regulations, and policies

Whether access will be controlled (available after approval)

Any other conditions that may limit sharing

Access, Distribution, and Reuse Example

“Following all federal, Tribal and state laws, all data from donors that do not allow for sharing will be excluded from shared datasets. Participants have been signing consent forms since 1993 and the wording has evolved over time, and it was not until spring of 2022 that language discussing broad sharing was included. Most participants allow for sharing for study of neurodegenerative diseases, with some allowing for sharing only for academic research use. Those allowing for partial sharing will be shared with NIAGADS with the conditions specified in the consent documentation.

All data will be shared in the controlled access data repository, NIAGADS. The access to this repository is limited to qualified investigators with a legitimate research interest, and is approved by an independent committee of researchers (the Data Use Committee) designated by NIAGADS.

In order to ensure participant consent for data sharing, IRB documentation and informed consent documents will include language describing plans for data management and sharing data, describing the motivation for sharing, and explaining that personal identifying information will be removed. To protect participant and family member privacy and confidentiality, shared data will be de-identified according to all federal and state guidelines and following the safe-harbor method. That method specifies that many identifiers are removed from data to be considered de-identified, including, but not limited to: names, all geographic subdivisions smaller than state, dates (except year), ages over 89 (listed as 90+ in all datasets), identifiable electronic numbers, biometric identifiers, various ID numbers (SSN, etc), and other possible identifiers. Only the minimum of PHI will be collected for the purposes of the study, and all team members are HIPAA trained.”

Oversight of Data Management and Sharing

Indicate how compliance with the Plan will be monitored and managed, frequency of oversight, and by whom (e.g., titles, roles).

Oversight Example

“Institutional support is provided via the Becker Medical Library and Office of the Vice Chancellor of Research at Washington University. Chris Sorensen, Senior Support scientist will provide support from the Becker Medical Library, and Cathy Alvey, Senior Grant Specialitst, will provide support from the Office of Sponsored research.

The following individual, Oscar Harari, will ultimately be responsible for data collection, management, storage, retention, and dissemination of project data, including updating and revising the Data Management and Sharing Plan when necessary, and will report on data sharing and compliance in the annual project progress reports. Oscar Harari is the Principal Investigator of the project, an Associate Professor of Psychiatry at Washington University in St. Louis. His email is harario@wustl.edu.

Jacqueline Kaczaral, Research Project Coordinator in Dr. Harari's lab, will also maintain the Data Management and Sharing Plan, and coordinate permissions with data repositories.”

References

NIH. 2020. *Supplemental Information to the NIH Policy for Data Management and Sharing: Elements of an NIH Data Management and Sharing Plan*. Notice: NOT-OD-21-014. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-014.html

Examples from:

- Kaczaral, J. 2022. *Single Cell Transcriptomics in AD.* [draft DMSP]. DMPTool. https://doi.org/10.48321/D1H311