Plan Overview

A Data Management Plan created using DMPTool

Title: Single Cell Transcriptomics in AD

Creator: Jacqueline Kaczaral - **ORCID:** 0000-0001-7684-3846

Affiliation: Washington University in St. Louis (wustl.edu)

DMP ID: https://doi.org/10.48321/D1H311

Funder: National Institutes of Health (nih.gov)

Template: NIH-GEN DMSP (Forthcoming 2023)

Project abstract:

Alzheimer's Disease (AD) is the result of complex interactions among genetic factors that cause pleiotropic changes in molecular networks linking a host of biological processes in multiple cell-types of the brain. Genetic mutations *APP*, *PSEN1* and *PSEN2* with autosomal dominant inheritance (ADAD) involve the amyloid cascade hypothesis in the etiology and pathogenesis of AD[1]. The strongest common risk factor for AD, *APOE* ε4, implicates cholesterol metabolism, while *TREM2* is involved in the immune system

Start date: 10-23-2022

Last modified: 10-21-2022

Single Cell Transcriptomics in AD

Data Type

Types and amount of scientific data expected to be generated in the project: ***Summarize the types and estimated amount of scientific data expected to be generated in the project.***

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

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.

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.***

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.

A brief listing of the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

To facilitate interpretation of 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

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed

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.

If applicable, specify how needed tools can be accessed, (e.g., open source and freely available, generally available for a fee in the marketplace, available only from the research team) and, if known, whether such tools are likely to remain available for as long as the scientific data remain available.

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

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist

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 Associated Timelines

Repository where scientific data and metadata will be archived: Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived; see Selecting a Data Repository**)**

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.

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.

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.

When and how long the scientific data will be made available: Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

The 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, or Reuse Considerations

Factors affecting subsequent access, distribution, or reuse of scientific data: NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See Frequently Asked Questions **for examples of justifiable reasons for limiting sharing of data.**

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.

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

All 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.

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).

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

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

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.