# DATA MANAGEMENT AND SHARING PLAN

If any of the proposed research in the application involves the generation of scientific data, this application is subject to the NIH Policy for Data Management and Sharing and requires submission of a Data Management and Sharing Plan. If the proposed research in the application will generate large-scale genomic data, the Genomic Data Sharing Policy also applies and should be addressed in this Plan. Refer to the detailed instructions in the application guide for developing this plan as well as to additional guidance on [sharing.nih.gov.](https://sharing.nih.gov/) The Plan is recommended not to exceed two pages.

Text in italics should be deleted. There is no “form page” for the Data Management and Sharing Plan. The DMS Plan may be provided in the *format*

shown below.

# Element 1: Data Type

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

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Species | Platform/Source | Amount |
| Array-derived genotype data | Human | Illumina | 1,000 research participants (500 cases/controls), prospective enrollment |
| 30x whole-genome sequence data | “ | “ | “ |
| RNA-seq data | “ | “ | “ |
| Single Cell RNA-seq data | “ | 10x/Illumina | “ |
| Hi-C WGS | “ | “ | “ |
| Phenotypic and clinical data | “ | Institutional EHR | “ |
| Demographic data | “ | “ | “ |

# Scientific data that will be preserved and shared, and the rationale for doing so:

Genomic (e.g., sequencing reads and variant call files) and phenotypic/clinical data from this project will be useful to researchers beyond those involved in this project and will therefore be preserved and shared. We will share de-identified patient demographics, genomic and clinical/phenotypic data extracted from medical records that are used to substantiate the findings that we publish. In alignment with NHGRI’s expectation to share comprehensive phenotypic data, we will also select several (5+) other key phenotypic variables extracted from the medical record to provide additional context about the research participants’ health to secondary users to maximize the utility of the shared data.

Data that do not meet quality metrics (e.g., RIN>7, replicate concordance >0.8, FastQC check) will not be preserved and shared. HIPAA identifiers will be preserved at our institution but will not be shared.

# Metadata, other relevant data, and associated documentation:

Metadata – QC metrics, sample id, batch run, assembly, data standards (i.e., data dictionary and ontology), and metadata required for repository submission (e.g., specimen source, instrument platforms)

Associated Documentation – Non-proprietary data collection instruments, methods, and study protocol(s)

# Element 2: Related Tools, Software and/or Code:

All newly developed software and code for processing and analyzing data will be distributed as version controlled, open-source code written in R or Python via GitHub, with detailed user documentation.

# Element 3: Standards:

|  |  |
| --- | --- |
| Data Type | Standard |
| Human array-derived genotype data | VCF |
| 30x whole-genome sequence data | Sequencing data and variant calls will be shared in FASTQ and VCF formats, respectively. |
| RNA-seq data | Data will be QCd and analyzed according to ENCODE Bulk RNA-seq Data Standards. FASTQs, BAM alignment files, and TSV transcript quantifications will be shared. |
| Hi-C WGS | FASTQ |
| Single cell RNAseq data | 10x data will be processed with CellRanger to generate FASTQ files. FASTQ files and the ‘filtered\_feature\_bc\_matrix’ folder will be shared. |
| Demographic, Phenotypic and Clinical Data | * PhenX for surveys
* RxNorm for meds
* PCORnet CDM which is derived from OMOP for EHR data collection for secondary outcomes
* Current Procedural Terminology (CPTs), Logical Observation Identifier Names and Codes (LOINCs) and diagnoses ICD10 codes
 |
| Study protocols | Customized (non-standard) & to be developed |

# Element 4: Data Preservation, Access, and Associated Timelines

1. **Repository where scientific data and metadata will be archived:**

The study will be registered in dbGaP.

Transcription data will be made available through GEO.

Variant data will be submitted to the NIMH Data Archive (NDA) or the Sequence Read Archive (SRA).

Protocols related to donor recruitment, tissue collection/preservation/biobanking, pathology/tissue dissection, whole-genome sequencing, and data processing and analysis will also be openly available on the website protocols.io and/or on the project website at the time of data release.

# How scientific data will be findable and identifiable:

Our dataset will be registered in dbGaP. We will reference the accession number(s) for our dataset(s) in all relevant future publications.

# When and how long the scientific data will be made available:

We will meet the data submission and release timeframes specified by the NIH Genomic Data Sharing and Data Management and Sharing Policies, as described on NIH’s data sharing website and NHGRI’s data sharing policies and expectations webpage. We will generate genomic data in batches of 100 participants. In accordance with NIH and NHGRI’s Expectations for Data Submissions and Release, we will begin submitting genomic data no later than 3 months after data from the first batch is generated and quality measures has been assessed. We will add subsequent batches as they are generated. Genomic data will be released 6 months after they are submitted. Phenotypic and clinical data, metadata, and associated documentation will be submitted along with the genomic data files, and the dataset will be released in full by the time any results supported in whole or in part by this award are posted to a preprint or submitted to a journal. In the event that we do not publish on these data or a portion of the data, they will be released before the end of this award.

# Element 5: Access, Distribution, or Reuse Considerations

1. **Factors affecting subsequent access, distribution, or reuse of scientific data:**

Research participants will be consented for data sharing of their individual genomic and clinical data via controlled access. Our institution will provide an Institutional Certification upon registering the study in dbGAP. Participants will be consented in a manner that allows for any research question to be explored (i.e., the General Research Uses (GRU) data use limitation). Genomic Summary Results from this study can be shared through unrestricted access.

# Whether access to scientific data will be controlled:

Individual-level genomic and clinical data will be shared via controlled-access policies defined by the SRA or NDA.

Protections for privacy, rights, and confidentiality of human research participants:

Data will be de-identified according to HIPAA and the Common Rule. Participants will have the opportunity [to opt-out of such sharing] or to withdraw their data from the database by contacting the study team or the university’s research administration office. We will track these preferences closely and respect individual participant wishes.

Upon receipt of an NIH Award, the data for this study will be protected by a Certificate of Confidentiality.

# Element 6: Oversight of Data Management and Sharing:

The Office of Sponsored Programs at Emory University which will be administering this award has created a data management and sharing plan compliance system as part of their process for submitting the annual NIH progress report. That Office along with the PI will be monitoring the submission of data to the GEO and SRA/NDA databases.