OMB No. 0925-0001 and 0925-0002 *(Rev. 07/2022 Approved Through TBD)*

Sample DMS Plan – Technology Development Project

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. 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

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

Type	Source	Amount
Nanopore sequence data	[Small Business X Technology]	20 human cell lines, obtained from BIOBANK X
30x whole-genome sequence data	"	"
RNA-seq data	"	"

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

Some of the nanopore sequence data generated over the course of this technology development project will be preliminary data that doesn't meet the quality metrics that warrant broad data sharing. As the technology matures, and the quality of the sequencing reads improves, we anticipate generating some high-quality genomic data (sequencing reads, base modification calls, and variant call files) that would be useful to researchers (e.g., as a reference for these newer file types) beyond those involved in this project. These files will therefore be preserved and shared. Because of the size of nanopore sequencing files, we will share compressed file types.

30X whole-genome and RNA-seq data that are generated as controls for our tech dev project will also be shared.

C. Metadata, other relevant data, and associated documentation:

Metadata: QC metrics for the genomic data types, data standards used, and metadata required for AnVIL submission
Associated Documentation: 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/Format
Nanopore sequence data	FAST5
30x whole-genome sequence data	Sequencing data and variant calls will be shared in CRAM 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.
Study protocols	Customized (non-standard) & to be developed

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:

The NHGRI Analysis, Visualization, and Informatics Lab-Space (AnVIL).

B. How scientific data will be findable and identifiable:

Our dataset will be registered in dbGaP and assigned a phsID. Data will be findable and identifiable using via the standard data indexing tools in AnVIL (currently the AnVIL catalog). We will reference the accession number(s) for our dataset(s) in all relevant future publications.

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

We will meet the data submission and release timeframes specified by the NIH GDS and DMS Policies, as described on NIH's data sharing website and NHGRI's data sharing policies and expectations webpage. We will submit genomic data no later than 3 months after observing that quality measures have been met. Genomic data will be released 6 months after they are submitted.

Currently, AnVIL has no process for deleting or retiring data sets; data will be available for as long as AnVIL/NHGRI preserves the dataset.

Element 5: Access, Distribution, or Reuse Considerations

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

We will be using BIOBANK X cell lines that are consented for unrestricted data sharing. Our institution will provide an Institutional Certification upon registering the study in dbGaP, indicating that both individual-level genomic data and Genomic Summary Results from this study can be shared through unrestricted access.

B. Whether access to scientific data will be controlled:

No, we are using human samples for which genomic data can be shared via unrestricted access.

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

Only genomic data will be shared; we are not obtaining demographic or phenotypic information from BIOBANK X.

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 study PI will be overseeing execution of this Data Management and Sharing Plan. X PI will be assessing quality metrics and will determine when data are of a sufficient quality to be shared broadly via the AnVIL. Progress on data sharing will be reported in the Research Performance Progress Report. Given this is a technology development project, we anticipate that this Plan may need to be updated as the project progresses.