

Data Management and Sharing Plan

Element 1: Data Type:

A. 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.

The proposed project will generate high-throughput sequencing data from 20 human cell lines, obtained from BIOBANK X, using technologies developed by Small Business X Technology. The types and expected quantities of data include:

Nanopore sequence data: Long-read sequencing data will be produced for all 20 human cell lines using Oxford Nanopore technology, enabling structural variant detection and improved genome assembly.

30x whole-genome sequencing (WGS) data: High-coverage short-read sequencing will be performed on the same 20 cell lines to generate comprehensive genomic variant data.

RNA sequencing (RNA-seq) data: Transcriptomic profiling will be conducted on all cell lines to capture gene expression dynamics and support integrative omics analyses.

B. 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.

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 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:

Briefly list 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.

Metadata: QC metrics for 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:

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.

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:

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.

All data generated in this study will follow established community standards and formats where applicable to ensure interoperability and support data sharing: Nanopore sequence data will be provided in the FAST5 format, which includes raw signal-level data suitable for downstream base calling and quality control. 30x whole-genome sequencing (WGS) data will be processed and shared in: CRAM format for compressed sequencing reads VCF format for variant calls RNA sequencing (RNA-seq) data will be analyzed following the ENCODE Bulk RNA-seq Data Standards, including quality control and standardized workflows. The following file types will be shared: FASTQ files (raw reads); BAM files (aligned reads); TSV files (transcript quantifications); Study protocols will be documented in a customized, non-standard format developed specifically for this project, to support transparency and reproducibility.

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

A. 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.

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

B. 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.

Our dataset will be registered in dbGaP and assigned to a phsID. Data will be findable and identifiable using 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:

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

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.

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:

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

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:

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).

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:

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

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 sufficient quality to be shared broadly via the AnVIL. Progress in 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.