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 project will generate three main types of high-throughput sequencing data from 20 human cell lines obtained from BIOBANK X:
- Nanopore sequence data: Long-read sequencing data (~100 Gb per cell line, ~2 Tb total) produced using Oxford Nanopore technology, enabling structural variant detection and high-quality genome assembly.
- 30x Whole-Genome Sequencing (WGS) data: High-coverage short-read sequencing (~120 Gb per cell line, ~2.4 Tb total) for comprehensive genomic variant analysis.
- RNA sequencing (RNA-seq) data: Transcriptomic data (~8 Gb per cell line, ~160 Gb total) to capture gene expression profiles.

The total estimated raw data output is approximately 4.6 Tb, excluding processed and intermediate files.

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.

The following scientific data will be preserved and shared:
- Raw sequencing data: FASTQ files from Nanopore, WGS, and RNA-seq runs.
- Processed data: Aligned BAM/CRAM files, variant call files (VCF), and gene expression matrices.

Sharing these data types aligns with NHGRI and NIH guidelines for maximizing utility and reproducibility. Preserving both raw and processed data ensures that future users can verify results, apply new analytical methods, and integrate findings with other datasets.

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.

- Sample-level metadata (e.g., cell line identifiers, BIOBANK X accession numbers, passage number, tissue/cell type, demographic summaries where permitted)
- Experimental protocols (including library preparation, sequencing parameters, and quality control metrics)
- Data processing and analysis pipelines (including software versions and parameter settings)
- Data dictionaries and README files to facilitate data interpretation

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.

Specialized tools and software will be required for access and manipulation of the data:

- Nanopore data: Oxford Nanopore’s Guppy basecaller, MinKNOW software, and compatible analysis tools (available at [Oxford Nanopore Community](https://community.nanoporetech.com/))

- Short-read data: Standard genomics tools such as BWA (for alignment), GATK (for variant calling), and STAR (for RNA-seq alignment), all of which are open-source and available via [Bioconda](https://bioconda.github.io/) or [GitHub](https://github.com/)

- Custom scripts: Any custom code or workflows developed for data processing and analysis will be shared in a public GitHub repository, with documentation provided.

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.

The following community-accepted data and metadata standards will be applied:

- Raw sequence data: FASTQ format (standard for high-throughput sequencing data)

- Aligned sequence data: BAM/CRAM format (SAM/BAM Specification)

- Variant calls: VCF format (Variant Call Format Specification)

- RNA-seq quantification: Standard tab-delimited gene expression matrices (e.g., TPM/FPKM tables)

- Metadata: Minimum Information About a Sequencing Experiment (MINSEQE) guidelines and NCBI BioSample metadata standards will be followed

- Protocols: Described using protocols.io links or equivalent

These standards will ensure interoperability, reproducibility, and reusability of the datasets.

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.

-Raw and processed genomic data: Database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap/)
-RNA-seq data: Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) or Sequence Read Archive (SRA)
-Metadata and protocols: Associated with the primary dbGaP/GEO/SRA submissions and linked in publications

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.

Each dataset will be assigned persistent unique identifiers (e.g., dbGaP accession numbers, GEO Series accession numbers, BioSample IDs). These identifiers will be referenced in publications and project webpages, enabling data discovery via repository search tools and standard indexing services.

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.

Data will be submitted to repositories and made available no later than the time of first publication of results or at the end of the grant period, whichever occurs first. Data will remain available in the repositories for at least 10 years, in accordance with NIH and repository policies.

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.

Data will be shared consistent with the broad consent provided by participants. Potential factors affecting reuse include:

- Data use limitations as specified in consent forms (e.g., non-commercial use, general biomedical research)
- Protection of participant privacy and compliance with applicable laws and regulations (e.g., HIPAA, GINA)
- Controlled access requirements for sensitive human genomic data

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).

Yes, access to human genomic data (raw and processed) will be controlled via dbGaP's Data Access Committees. Qualified investigators must apply and agree to data use limitations.

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).

All data will be de-identified prior to submission, in compliance with NIH Genomic Data Sharing Policy and institutional IRB requirements. Data will be assigned random study IDs with all direct identifiers removed. A Certificate of Confidentiality will be obtained. Access to controlled data will be limited to approved users, and all users must agree to terms that prohibit re-identification and unauthorized data sharing.

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).

Compliance with this Data Management and Sharing Plan will be overseen by the Project Principal Investigator (PI) and the institutional Data Steward. The PI will ensure timely data submission, accurate metadata, and adherence to sharing timelines. The Data Steward will monitor compliance at quarterly project meetings and prior to manuscript submission. Oversight will include regular audits of data integrity and documentation. Any issues will be promptly reported to the NHGRI Program Officer.