# 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. The types and expected quantities of data include:
- Nanopore sequence data for all 20 human cell lines,
- 30x whole-genome sequencing (WGS) data for the same 20 cell lines, and
- RNA sequencing (RNA-seq) data for transcriptomic profiling on all cell lines.

The estimated amount of data is substantial due to the high-throughput nature of these technologies.

### 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.*

All generated data types (Nanopore sequence data, 30x WGS data, and RNA-seq data) from the project will be preserved and shared. The rationale for sharing these data is to facilitate open science principles, promote reproducibility of findings, enable meta-analyses across similar datasets, and accelerate genomic research by providing valuable resources to the scientific community.

### 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.*

The following metadata, other relevant data, and associated documentation will be made accessible:

- Sample information (e.g., cell line identifiers, source),
- Sequencing protocol details,
- Data processing pipelines used for each type of sequencing data,
- Quality control metrics for the sequencing runs,
- Study protocols,
- Informed consent documents that outline broad data sharing agreements with participants,
- and any other relevant documentation necessary to interpret and work with the shared data.

## 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, software, and/or code will be needed to access or manipulate the shared scientific data. These include but are not limited to:
- Oxford Nanopore Technologies' software for nanopore sequence data analysis,
- BWA and SAMtools for short-read alignment and processing,

- HISAT2 and StringTie for RNA-seq data analysis.

Access information for these tools will be provided in the project's documentation, including links to publicly available repositories or instructions on how to obtain necessary licenses.

**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 scientific data and associated metadata will apply common data standards to enable interoperability of datasets and resources. Specifically, we will adhere to:

- FASTQ format for raw sequencing reads,

- BAM format for aligned sequencing data,

- GFF/GTF formats for genomic feature annotations,

- and the Human Genome Variation Society (HGVS) nomenclature for variant descriptions.

These standards are widely adopted in the genomics community, facilitating the integration of our dataset with existing resources.

**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 scientific data and metadata arising from this project will be archived at the National Center for Biotechnology Information (NCBI) through databases such as the Sequence Read Archive (SRA) for sequencing data and potentially dbGaP if applicable, given the involvement of human subjects and genomic data.

**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.*

The scientific data will be made findable and identifiable via persistent unique identifiers (e.g., SRA accession numbers, DOIs for associated publications), as well as through standard indexing tools available at NCBI and other partner databases.

**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.*

The scientific data will be made available to other users no later than the time of publication of the main findings or upon completion of the project, whichever comes first. The data are expected to remain

accessible for at least 10 years following the close of the project to maximize their utility and impact on the research community.

**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.*

Given that all research participants will provide broad consent for data sharing, factors affecting access are minimal. However, we will adhere to all applicable regulations and guidelines related to human subjects research, including ensuring that shared data do not compromise participant privacy.

**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).*

Access to the scientific data will generally not be controlled, as the goal is to share these resources openly with the scientific community. However, access may require registration or agreement to terms of use at the repository where the data are archived, in line with standard practices for such databases.

**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).*

To protect the privacy, rights, and confidentiality of human research participants, all shared data will be de-identified according to HIPAA guidelines or equivalent standards. Additionally, the study protocol has been reviewed and approved by an Institutional Review Board (IRB) to ensure compliance with ethical and regulatory requirements for human subjects research.

**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).*

Not explicitly outlined in previous sections but integral to our approach is the commitment to preserve data integrity and facilitate sharing. This includes:

- Regularly backing up all project data,

- Using version control systems (e.g., Git) for tracking changes in analysis scripts and documentation,

- Documenting data processing and analysis workflows thoroughly, and

- Providing clear instructions for accessing and utilizing the shared datasets.