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 high-fidelity (HiFi) long-read whole-genome sequencing data from a single human HeLa cell line, obtained from ATCC. The sequencing will be performed using the PacBio platform. The expected data output is approximately 700 megabytes (MB), encompassing raw sequence reads (subreads/bam files), base-called reads (fastq files), and assembled genome data (fasta files). Associated quality metrics and summary files will also be generated.

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 sequencing data — including raw HiFi reads, base-called reads, and the assembled genome will be preserved and shared. Additionally, associated quality control metrics and summary documentation will be included. Sharing these data maximizes transparency, enables reproducibility, and provides resources for the genomics research community to further investigate complex genomic regions using long-read sequencing.

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 and documentation will be made accessible:

- Sample information (source, cell line identifier, ATCC accession, passage number, extraction date, etc.)
- Sequencing protocol and platform details (PacBio model, chemistry version, run parameters)
- Library preparation methods and protocols
- Data processing pipeline descriptions (software versions, parameters used)
- Data quality control metrics (read length distributions, error rates, coverage statistics)
- Study protocol and consent process documentation

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.

While the primary data files (fastq, bam, fasta) are standard formats accessible with open-source tools, specialized software may be needed for in-depth analysis. Examples include PacBio SMRT Link, samtools, minimap2, and IGV for visualization. These tools are freely available for academic use; download links and usage instructions will be provided in the associated documentation. Any custom scripts used for data processing will be shared via a public GitHub repository.

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.

Standard genomics data and metadata formats will be used to ensure interoperability:

- Sequencing data: FASTQ (base-called reads), BAM (raw reads), FASTA (assembled genome)

- Metadata: Minimum Information About a Sequencing Experiment (MINISEQE) guidelines, following the BioSample and BioProject standards used by NCBI

- Quality metrics: Standard QC files (e.g., MultiQC reports), and documentation in PDF or Markdown

All files will be named and organized according to accepted community conventions to facilitate reuse.

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.

All scientific data and associated metadata will be deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and GenBank, as appropriate. Metadata will be linked through NCBI BioSample and BioProject records.

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.

Scientific data will be assigned persistent unique identifiers (accession numbers) by the NCBI repositories (BioProject, BioSample, SRA, and GenBank). These identifiers will be referenced in publications and project documentation, ensuring data are discoverable via standard indexing tools such as PubMed and the NCBI search portal.

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 made available no later than the time of first publication of results or at the end of the funding period, whichever occurs first. Data will remain publicly accessible in the NCBI repositories for a minimum of 10 years from the date of deposit, 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.*

All research participants (cell line donors) have consented for broad data sharing. Because the data derive from an established and widely distributed cell line (HeLa), and do not include personally identifiable information, there are no anticipated restrictions on access, distribution, or reuse beyond those imposed by repository and NIH policies. Users will be encouraged to cite the data and acknowledge the source in any derivative work.

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 be open and unrestricted. Data will be deposited in open-access repositories (NCBI SRA/GenBank), and no additional approval will be required for data 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).

The data are derived from an immortalized cell line (HeLa) widely used in research and do not contain any new personally identifiable information. All data will be de-identified, and there are no direct privacy risks to individuals. The study will comply with all applicable institutional, NIH, and legal guidelines regarding the use and sharing of human-derived cell lines.

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's Data Steward and the Principal Investigator (PI). The Data Steward (a designated staff member with training in data management best practices) will monitor data collection, documentation, and deposition timelines on a quarterly basis. The PI will review compliance annually and prior to major deliverables (e.g., publications, grant reporting). The Institutional Data Management Office will provide additional oversight as required by institutional policy.