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 aims to generate high-fidelity (HiFi) long-read whole-genome sequencing data from a single human HeLa cell line using the PacBio platform. The total expected data output is approximately 700 megabytes (MB), which will capture the complete genome of the HeLa cell line with high-resolution coverage.

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 HiFi long-read whole-genome sequencing data will be preserved and shared. The rationale for sharing this data is to contribute to the scientific community's understanding of the HeLa cell line genome, facilitate future research, and enable the validation of findings by other investigators. Sharing this high-quality genomic data can also promote collaborations and accelerate discoveries in genomics and related fields.

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 metadata to be shared will include sequencing parameters, sample preparation methods, and data processing pipelines used. Other relevant data will encompass quality control metrics for the sequencing runs and the bioinformatic tools used for data analysis. Associated documentation will include study protocols, data collection instruments, and detailed descriptions of the computational workflows employed for data generation and analysis.

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 needed to access or manipulate the shared scientific data include PacBio's SMRT Link for primary data analysis, and potentially other bioinformatic tools like Genome Analysis Toolkit (GATK) for secondary data analysis. These tools can be accessed through their respective official websites or through institutional 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 common data standards to be applied to the scientific data and associated metadata will include those specified by the Genomic Data Commons (GDC) for genomic data, and the MINSEQE guidelines for reporting high-throughput sequencing experiments. These standards will enable interoperability of datasets and resources, facilitating their integration into larger studies or meta-analyses.

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 the project will be archived in the National Center for Biotechnology Information (NCBI) database, specifically through the Sequence Read Archive (SRA) for raw sequencing data and potentially the Genomic Data Commons (GDC) for analyzed 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 such as accession numbers provided by the NCBI or DOI numbers for publications associated with the dataset. Standard indexing tools and metadata tags will also be used to facilitate discovery through database searches.

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 no later than the time of publication of the main findings from the project or at the end of the performance period, whichever comes first. The data are expected to remain available indefinitely, as per the policies of the chosen data repositories.

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 consent for broad data sharing, no significant limitations are anticipated regarding access, distribution, or reuse of the scientific data beyond standard considerations for protecting participant privacy and confidentiality.

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 not be controlled; it will be made openly available through the designated public repositories without the need for approval or access restrictions.

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).

Although the project involves sequencing a well-characterized HeLa cell line (which is not considered human subjects research in the traditional sense since it's an immortalized cell line), any future studies involving primary human samples will adhere to rigorous standards for de-identification and protection of privacy, rights, and confidentiality. For this specific proposal, given the use of a cell line, these considerations are minimized but will be addressed as per institutional guidelines and regulatory requirements.

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 monitored and managed by the Principal Investigator (PI) in collaboration with the institution's research data management office. Regular checks will be conducted to ensure that all generated data are properly documented, stored, and shared according to the plan outlined here. The frequency of oversight will be at least quarterly during active data generation phases and annually thereafter to ensure ongoing compliance and address any emerging issues related to data sharing and management.