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. Sequencing will be performed using the PacBio platform, which offers high-accuracy long reads suitable for resolving complex genomic regions. The total expected data output is approximately 700 megabytes (MB), capturing 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 the scientific data (and accompanying metadata described in 1.C.) will be preserved and shared. These data will complement existing whole-genome data that are available from the HeLa cell line maintained in dbGaP.

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, relevant metadata pertaining to the sample (obtained from ATCC), and other metadata required for dbGaP deposition. Associated Documentation: Study methods and protocols.

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.

We are not generating any new tools, software and/or code for processing or analyzing these HiFi WGS data (i.e., FASTQ files); we will make use of open-source tools that are freely available to the scientific community.

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.

HiFi long-read whole-genome/ sequencing data FASTQ

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 metadata will be deposited to dbGaP.

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 and deposited in dbGaP and assigned a phsID. Data will be findable and identifiable via the standard data indexing tools in dbGaP/NCBI. 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 Genomic Data Sharing and Data Management and Sharing Policies, as described on NIH's data sharing website and NHGRI's data sharing policies and expectations webpage. In accordance with the NIH Genomic Data Sharing Policy and NHGRI's Expectations for Data Submissions and Release timelines for level 2 data, we will begin submitting genomic data no later than 3 months after data is generated and quality measures has been assessed. Genomic data will be released 6 months after they are submitted to dbGaP in accordance with the level 2 data release timeline. Metadata and associated documentation will be submitted along with the genomic data files, and the dataset will be released in full by the time any results, supported in whole or in part by this award, at time of associated publication. If we do not publish on these data or a portion of the data, they will be released by the end of this award. Level 3-4 data will not be generated for this project. Currently, dbGaP has no process for deleting or retiring data sets; data will be available for as long as dbGaP 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.

The NIH-Lacks Family Agreement (Agreement) expects NIH-funded investigators who generate HeLa cell whole genome sequence data (DNA or RNA) to submit those data to NIH through dbGaP to the HeLa cell Genome Sequencing Studies. The data are distributed according to the HeLa Genome Data Use Agreement under Health, Medical, and Biomedical (HMB) purposes. Through the NIH-Lacks Family Agreement, an

Institutional Certification with the designated HMB data use limitation has been completed for the HeLa cell Genome Sequencing Studies and is held by dbGaP. No additional Institutional Certification is needed.

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, data will be controlled. Access requests for HeLa Cell Genome Sequences Studies are evaluated by the Advisory Committee to the Director (ACD) HeLa Genome Data Access Working Group to assess whether terms of use align with the HeLa Genome Data Use Agreement. The requests are then reviewed by the ACD to recommend to the NIH Director to approve or disapprove. The NIH Director makes the final access decision.

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).

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. The Study PI will be assessing quality metrics and will deposit all scientific data and metadata according to the timelines provided above. Progress on data sharing will be reported in the Research Performance Progress Report.