NIH Advisory Committee to the Director
March 28, 2017

Summary of HeLa Genome Data Access Request

1. Project #12876, Studies of L1-Mediated Pseudogene Formation in Human HeLa Cells
   University of Michigan

2. Project #12799, Predicting 3D Regulatory Interactions
   J. David Gladstone Institutes

National Institutes of Health
Advisory Committee to the Director
HeLa Genome Data Access Working Group

HeLa Genome Data Access Request: Project 12876

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Studies of L1-Mediated Pseudogene Formation in Human HeLa Cells**
Date Received	12/13/2016
Project Summary (Provided by NIH)	• Long Interspersed Element-1 (L1) retrotransposons are also known as "jumping genes" due to their ability to move about the genome. Much of what is known about L1 "jumping genes" come from using HeLa cells. • The investigator wishes to use the HeLa whole genome sequence data in dbGaP as a reference and to validate how L1 or 'jumping genes' move about the genome to impact its structure and function and contribute to human genetic variation and in some cases, disease.
Institution	University of Michigan
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	We study how Long INterspersed Element-1 (L1) 'jumping genes' (retrotransposons) impact the structure and function of the human genome. Active L1s encode two proteins (ORF1p and ORF2p) that promote the mobility of L1 RNA to new genomic locations by a 'copy and paste' process termed retrotransposition. The L1 proteins also occasionally can act in trans to retrotranspose non-autonomous retrotransposon RNAs (e.g., Alu elements), non-coding RNAs (e.g., U6 small nuclear RNA), and cellular mRNAs. Since the HeLa cell line is routinely used as a model in studies of L1 biology, we respectfully request access to DNA and RNA sequencing data from HeLa cell lines, which were deposited in dbGAP for protected access by members of the scientific community. We will use these data in control and validation experiments to: 1) examine the repertoire of L1-mediated pseudogenes in HeLa genomic DNA; and 2) identify candidate genes that play a role in promoting Alu retrotransposition. As specified in the HeLa Genome Data Use Agreement (HGDUA), only the individuals listed in this application will have access to HeLa data, the data will be stored on secure servers, and we will provide advance notice to the National Institutes of Health (NIH) if others within our laboratories request access to the HeLa data. Notably, our colleague, Dr. Kitzman, co-led a project that generated a haplotype-resolved DNA sequence and RNA sequence datasets from HeLa cells; he is well versed with the NIH compliance policies regarding the handling of HeLa cell data. We do not anticipate that our studies will generate intellectual property (IP) or lead to commercial product development. If these expectations change, we will immediately notify the NIH under the terms stated in the HGDUA. We plan to disseminate the results of our studies at scientific meetings and through publication in peer-reviewed journal(s) and will include the proper

HeLa Genome Data Access Request: Project 12876

	acknowledgement stated in the HGDUA in any publications and presentations. We thank the Lacks family and NIH HeLa Genome Data Use Advisory Board for considering our request to use HeLa cell DNA and RNA sequence data in our studies
Non-Technical Summary (Provided by Requestor)	Sequences derived from 'jumping genes' (i.e., transposable elements) comprise approximately 45% of human genomic DNA and have had a profound effect on the structure and function of our genomes. Indeed, some 'jumping genes' (notably Long INterspersed Element-1 (LINE-1 or L1) retrotransposons) remain active and their mobility can contribute to human genetic variation, and, on occasion, sporadic forms of human disease. However, relatively little is known about the detailed molecular mechanisms of L1 jumping or the cellular factors that promote or restrict L1 jumping. In this proposal, we wish to use the HeLa cell DNA and RNA sequence data in dbGAP to gain insights about how L1-mediated processes impact the structure and evolution of the human genome.

National Institutes of Health
Advisory Committee to the Director
HeLa Genome Data Access Working Group

HeLa Genome Data Access Request: Project 12799

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Predicting 3D Regulatory Interactions**
Date Received	12/20/2016
Project Summary (Provided by NIH)	• The investigator previously published a method, TargetFinder, which predicts 3D interactions between genomic regions from basic genomic data. • The investigator requests to use HeLa whole genome sequence data to validate their method, TargetFinder, in different cell lines
Institution	J. David Gladstone Institutes
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	We recently published a machine-learning method, called TargetFinder, that predicts 3D interactions between enhancers and their target genes from 1D genomic data (sequence, ChIP-seq, DNA-methylation, RNA-seq) [Whalen et al. 2016, Nature Genetics]. The method performs well on held out data within a cell line, but its accuracy is lower when a model trained on one cell line is used to predict on a different cell line. This is an important practical application, because validated interactions (e.g., high resolution Hi-C data) are not yet available in many cell types that do have the 1D data need to make predictions with a trained model. We are evaluating solutions to this problem using data from different cell lines from Rao et al. It would be very helpful to include HeLa. The HeLa Hi-C interaction data will be combined with interaction data, as well as ChIP-seq and RNA-seq, from other ENCODE cell lines. Dissemination: We will publish our research findings in a peer reviewed journal with open access, as well as posting a pre-print on biorxiv. The PI will present research findings at conferences and in seminars. We have no plans to develop commercial products, services, or file IP based on this project. This line of research could not reasonably be expected to result in a commercialized product or service. All software/methods from our lab are open source. We do not expect these plans to change, and we agree to inform the NIH if our plans for IP or commercialization change.
Non-Technical Summary (Provided by Requestor)	We will use validated regulatory interactions from the HeLa cell line to test a method for predicting interactions in cell types that are not as well-characterized.