NIH Advisory Committee to the Director
December 11, 2020

Summary of HeLa Genome Data Access Requests

1. Project #24301, Detecting mutations contributing to resistance to cancer treatment
   The Bioinformatics CRO, Niceville, FL

2. Project #26116, Nascent transcription at CTCF boundaries regulates insulation and promoter transcription
   National Centre For Biological Sciences, Karnataka, India

3. Project #27115, Identify split reads from whole genome sequencing data of HeLa cell line
   H. Lee Moffitt Cancer Center and Research Institution, Tampa, FL

This page intentionally left blank

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

HeLa Genome Data Access Request: Project 24301

Working Group Finding	Consistent with the Data Use Agreement
Project Title	**Detecting mutations contributing to resistance to cancer treatment**
Date Received	7/17/2020
Project Summary (Provided by NIH)	• The investigator is studying how drug treatment of cancer cells can change the cancer cell's genomic sequence (e.g., mutate) to cause drug resistance, or an inability of cancer cells to respond to drug treatment. • In order to identify which genetic sequences changed, or mutated, as a result of drug resistance, the investigator proposes to compare the genome of drug-treated, drug resistant cancer cells to the untreated cancer cell genomes from the HeLa Cell Genome Sequencing Studies.
Institution	The Bioinformatics CRO, Niceville, FL
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	In our study we have treated HeLa cells with a drug compound, and isolated cells that were sensitive to the treatment, as well as cells that were resistant to 10 uM and 25 uM of the drug. Sensitive and resistant cells have been sequenced with Whole Genome Sequencing (WGS) technology. Our objective is to detect and interpret the differences in the called mutations between resistant and sensitive cells. To this end, we are comparing the variant call sets between the 10 uM resistant vs sensitive, and 25 uM resistant vs sensitive samples, looking for variants that are not present in the sensitive cells but are found in the resistant cells. WGS data pre-processing and variant calling were performed by the somatic variant calling pipelines released by the Broad Institute. Since variant calling is imperfect for each individual sample, we expect our approach to result in a number of false positives for various reasons, one reason being the possibility that real variants are incorrectly undercalled in sensitive cells. If a real variant is correctly called in the resistant cells and incorrectly not called in the sensitive cells, this could be misinterpreted as a mutation potentially relevant to drug resistance. Therefore we would utilize the variant calls from the Whole Genome Sequencing data from HeLa Cell Genome Sequencing Studies for filtering our final variant call set in order to reduce the number of false positives. Variants detected in the HeLa Cell Genome Sequencing Studies data would be filtered out since we can be certain about their presence in both the treatment sensitive and resistant samples. We would not use phenotypic characteristics data from the dataset, only the list of genetic variants. Statements about not seeking IP or commercialization: - We do not have any plans to develop a commercial product or service or file Intellectual Property (IP) based on our findings from the proposed research.

HeLa Genome Data Access Request: Project 24301

	- Our findings cannot reasonably be expected to result in a commercialized product or service. - We are not expecting our plans to change regarding our intention not to seek IP or commercialization. - We agree to inform the NIH if our plans for IP or commercialization change. Please provide us with the contact details of who we would need to notify in case of such changes. Statement on how the research findings from the proposed research will be disseminated: - We intend to publish the findings from this study in a peer reviewed journal. This will be part or a larger study that describes the identification of a new class of cancer chemotherapeutics. We attest that we will acknowledge Henrietta Lacks and her surviving family as the source of the data, as well as the NIH Director, the Advisory Committee to the Director, and the HeLa Genome Data Access Working Group, in any publication and presentation of any of our research findings derived from the HeLa cell line data. Our acknowledgment will follow the example acknowledgement statement we were provided with by the NIH in their revision request.
Non-Technical Summary (Provided by Requestor)	In our research we try to understand why a drug molecule that has proven to be effective against cancer cells becomes ineffective. We compare the mutations (changes in the DNA) in cells for which the drug has the desired effect with those in cells for which the drug fails. We would use data from the HeLa Cell Genome Sequencing Studies to make this comparison more accurate.

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

HeLa Genome Data Access Request: Project 26116

Working Group Finding	Consistent with the Data Use Agreement
Project Title	**Nascent transcription at CTCF boundaries regulates insulation and promoter transcription**
Date Received	8/10/2020
Project Summary (Provided by NIH)	• The investigator proposes to use the HeLa Cell Genome Sequencing Studies to understand the regulation of a specific area in the genome that is highly transcribed, or turned on, in HeLa cells.
Institution	National Centre For Biological Sciences, Karnataka, India
Collaborator(s)	None
Research Use Statement (Provided by Requestor)	NK4a/ARF locus codes for critical cell programming regulators namely p14ARF, p15INK4b, and p16INK4a from two coding genes; CDKN2A (p14ARF and p16INK4a), CDKN2B (p15INK4b). The locus also harbors a long non-coding RNA, CDKN2BAS (Antisense-noncoding-RNA in the INK4a-locus (ANRIL)) at 3' end. Together these proteins inhibit the cyclin dependent kinases (CDKs) to regulate the cell cycle. Due to these inhibitory roles, the locus is methylated or deleted in 90% of the tumors, except for a few types of cancers like breast, prostate, non-small cell lung, Human papillomavirus (HPV) positive cancers such as head/neck, and cervical cancers. Notably, ~90% of the cervical tumors are HPV positive. Apart from cancers, the activation of INK4a/ARF genes is the hallmark of senescence and aging. Hence, this locus is the most reproducible GWAS hotspot associated with various age and lifestyle related diseases such as; Coronary artery disease, Type-2-diabetes, Alzheimer's and Atherosclerosis among others. In spite of the tremendous importance of the locus in disease pathologies, its transcriptional regulation in associated cancers and senescence is poorly understood. Most studies focusing on INK4a/ARF transcriptional regulation have focused on the promoter-driven mechanisms. However, disease-associated variants identified in GWAS studies lie in gene desert region adjacent to CDKN2A/2B genes. Hence, the plausible regulation of the locus through the gene desert cannot be ignored. The deletion of gene desert region in mouse and iPSCs, and their subsequent differentiation into relevant cell type has shown the genome-wide alterations in coding genes associated with coronary artery disease and atherosclerosis suggesting the gene desert regulates the INK4/ARF locus. However, functional regulatory elements behind these roles in the large deleted regions remain unknown. In this proposal, we would like to access genomic data in HeLa to understand the regulation of INK4/ARF locus as the locus is highly transcribed in these cells. Further, the findings from INK4/ARF locus will be tested for their relevance in genome-wide manner.

HeLa Genome Data Access Request: Project 26116

	The goal of access to the data requested: We have previously shown that INK4a/ARF gene desert harbors several enhancers with the unexplored potential of regulating this locus in HeLa (Harismendy et al., 2011). How these multiple enhancers, that also overlap with GWAS variants, transcriptionally regulate this multigene locus is unknown, an important step towards understanding the biological relevance of these variants. Further, the 3' end of INK4/ARF TAD (Topologically associated domains) harbors multiple CTCF sites. We have observed that transcriptional potential of INK4/ARF locus is greatly altered upon deletions of these CTCF sites in HeLa cells. Moreover, the enhancers within the gene desert also show loss of activity under these conditions. These data suggest that transcriptional potential of TAD and its enhancer strength is regulated by the boundary. We would like to extend this analysis to genome-wide TADs in HeLa obtained from Hi-C and will compare these findings with; RNA-expression from bulk and single cell RNA-seq. Further, to test the relevance of these observations in non-cancerous cells, we will perform similar analysis and comparisons of HeLa data with IMR90 fibroblasts (young and senescent). The knowledge gained will form the basis of designing therapies to modulate the expression of INK4a/ARF proteins in aging related pathologies and cancers. We do not anticipate any intellectual property (IP) or commercial products out of this finding from HeLa cell lines. If there are any changes to our plan, we agree to notify the National Institutes of Health (NIH) under the terms of the HeLa Genome Data Use Agreement. We will publish our findings in the peer-review journals and present them in scientific conferences, with proper acknowledgment to the data source.
Non-Technical Summary (Provided by Requestor)	Contrary to the most cancers, this locus is highly activated in HPV-positive tumors such as cervical cancers and head/neck carcinomas. Thus, these genes are highly expressed in HeLa due to the presence of HPV. In this proposal we would like to understand how the locus is activated in HeLa. Particularly, we will be focusing on several DNA regulatory elements (enhancers and CTCF-sites) present adjacent to this locus. The genomic data access in HeLa will be crucial in relating the observation we have made so far or will be making with; genome-wide enhancers, CTCF-sites, chromatin architecture and transcription. The understanding of transcriptional regulation of this locus will help in developing therapies to target these genes in cervical tumors, aging and age related diseases.

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

HeLa Genome Data Access Request: Project 27115

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Identify split reads from whole genome sequencing data of HeLa cell line**
Date Received	10/20/2020
Project Summary (Provided by NIH)	• The investigator predicts that in HeLa cells, single stranded DNA uses U-turn like molecules to duplicate into double stranded DNA. The investigator proposes to use the HeLa Cell Genome Sequencing Studies to analyze if these U-turn like molecules play a role in DNA duplication by comparing the HeLa genome sequences to previously published duplication sequences. • The comparison between the two pair of alignments may uncover the use of U-turn like molecules in DNA duplication in HeLa cells.
Institution	H. Lee Moffitt Cancer Center and Research Institution, Tampa, FL
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	The HeLa cell line was used to decide the replication directionality (RD) with a pipeline named okazaki-seq in a previous publication. Recently, our group found a U-turn like DNA molecules in most sequencing datasets. To validate if these U-turn like molecules play a role in DNA replication, we thus request the whole genome sequencing data of HeLa cells and plan to compare the U-turn DNA loci to previous published RD transition region. We will use the HeLa cell data to identify distinct split DNA molecules in human genome. We will re-align these whole genome sequencing reads and identify the relations between such DNA molecules and human genome replication. We will require longer read length HeLa cell whole genome sequencing datasets and discover reads or read pairs harboring such structure variations. We validate our findings by comparing to published HeLa cell okazaki-seq datasets. We expect to understand further about the DNA replication nature from this analysis. We will not combine the HeLa cell sequencing data with other datasets. We have reviewed and agree with the principles for responsible research use and data handling of the genomic datasets as defined in the NIH Data Sharing Policy, and as detailed in this agreement, and in the dbGaP Approved User Code of Conduct. We will ensure that all uses of the data are consistent with federal, state, and local laws and regulations and any relevant institutional policies. We also certify that we are in good standing with the institution and relevant funding agencies and are eligible to conduct independent researches. We will submit annual data use reports to the HeLa Genome Data Access Working Group describing the research use of the data. Additional agreements were addressed in below: o All our statements are true and applicable to the access and use of all versions of these datasets.

HeLa Genome Data Access Request: Project 27115

	o We agree that information about the PI and the approved research will be posted on an NIH web site. The information will include the Approved User's name and institution, project name, Research Use Statement, and a Non-technical Summary of the Research Use Statement. In addition, citations resulting from the use of NIH genomic datasets will be posted on NIH data repository websites. o We agree not to attempt to contact family members of Henrietta Lacks. o We agree to retain control over the data and further agree not to distribute data obtained through this DAR to any entity or individual not covered in the submitted DAR. NIH genomic datasets obtained through this DAR, in whole or in part, will not be sold to any individual at any point in time for any purpose. o We agree that if we change institutions during the access period, we will submit a new DAR in which the new institution agrees to the NIH data use policy before data access resumes. Any versions of data stored at our prior institution use will be destroyed and documented through a final Data Use Report. o We agree to keep the data secure and confidential at all times and to adhere to information technology practices in all aspects of data management to assure that only authorized individuals can gain access to the HeLa datasets. This agreement includes the maintenance of appropriate controls over any copies or derivatives of the data obtained through this DAR. o We (including the institutional Information Technology Director) agree to handle the requested dataset(s) according to the current dbGaP Security Best Practices, including its detailed description of requirements for security and encryption. o We agree to notify any unauthorized data sharing, breaches of data security, or inadvertent data releases that may compromise data confidentiality within 24 hours of when the incident is identified. All notifications and written reports of data security incidents will be sent to: helagenome@nih.gov. o We do not anticipate any intellectual property (IP) or the development of commercial products or services currently from our research, but we agree to notify the NIH if our IP or commercial plans change. o We agree that the HeLa genome sequence data will be used in a pilot study to generate preliminary findings for subsequent research that involves data solely from other sources. The publication will acknowledge Henrietta Lacks and her descendants if the HeLa sequence data informed the subsequent research. oResearch findings from the proposed research will be presented in scientific meetings and publications.

HeLa Genome Data Access Request: Project 27115

Non-Technical Summary (Provided by Requestor)	DNA is double stranded, also called Watson and Crick strands. The replication of DNA is similar to bidirectional freeway, with one way (lagging strand) blocked by many small pieces of DNA nucleotides (called okazaki fragments), while the other way (leading strand) continuously synthesized. Scientists have developed an exquisite approach to characterize replication directionality by calculating okazaki fragment coverage aligned to Watson and Crick strands in HeLa cells. The goal of this study is to explore a new mechanism of DNA replication by re-analyzing whole genome sequencing data of HeLa cells. We hope to identify more evidence from the sequencing data to support our previous observation showing unexpected U-turn in some newly synthesized DNA.