NIH Advisory Committee to the Director
December 14, 2017

Summary of HeLa Genome Data Access Requests

1. Project #15197, Validation of Our In-house Software for Refined Protein Identification

   University of Geneva,
   Geneva, Switzerland

2. Project #16085, Statistical Methods for Profiling Intra-tumor Heterogeneity from Sequencing Data

   University of Chicago,
   Chicago, Illinois

3. Project #14865, Long Non-coding RNA Transcriptional Landscape in HeLa Cells

   Center for Genomic Regulation,
   Barcelona, Spain

4. Project #14603, Methods Development for the Comparisons of Three-dimensional Genome Structure

   Carnegie-Mellon University,
   Pittsburgh, Pennsylvania

5. Project #4483, Identification of Structural Change of Chromosomes Related to DNA Copy Number Changes in Cancer

   Gwangju Institute of Science/Technology,
   Gwangju, Korea

6. Project #15736, Determining Differences in the HeLa Genome at the CDR1-AS Locus

   Stanford University,
   Stanford, California

7. Project #12668, Mapping Controlled HeLa Transfection MS Data Over Customized Proteome Database Description

   Norwegian University of Science and Technology,
   Trondelag, Norway

National Institutes of Health
Advisory Committee to the Director
HeLa Genome Data Access Working Group
HeLa Genome Data Access Request: Project 15197

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Validation of Our In-house Software for Refined Protein Identification**
Date Received	7/24/2017
Project Summary (Provided by NIH)	• The investigator previously developed and implemented MzVar, a software tool that uses data obtained from mass spectrometry (MS), an analytical method that separates chemical species based on size and charge, to identify proteins. The investigator developed the software mainly using HeLa proteomic data available in public databases. • The investigator proposes to use HeLa cell genome sequence to validate the HeLa proteins identified using the MzVar software and to identify unstable proteins that are present in the HeLa genome sequence but absent in their HeLa proteome database assembled by MzVar.
Institution	University of Geneva, Geneva, Switzerland
Collaborator(s)	None

Working Group Finding	Consistent with the Data Use Agreement
Research Use Statement (Provided by Requestor)	We have developed and implemented MzVar, a software tool to compile customized variant protein and peptide databases in the prospect of refining protein identification from mass spectrometry data. In this project, our aim is to validate the MzVar tool on an extended dataset. We plan to identify the variants in the HeLa cancer cell line that consistently affect protein stability and expression, preventing their identification through proteomics workflows. First, the variants called from the genome sequencing of the HeLa cell line are used to compile a customized protein sequence database using MzVar. In parallel, tandem mass spectrometry (MS/MS) raw datasets from experiments using the HeLa cell line are downloaded from the PRIDE Archive public database. Then, the retrieved mass spectra are searched against our customized database using the X!Tandem open search engine. Finally, the results are filtered based on a stringent FDR threshold. The identifications of variant peptides and their wild-type counterparts are compared across all datasets and a statistical analysis is performed. The HeLa cell has a great value to us by its wide adoption in the proteomics community. This led to the large amount of mass spectrometry data currently available in public databases, which was an essential prerequisite to be able to perform our analysis. MzVar is a free software. We do not anticipate intellectual property (IP) or the development of commercial products and services from our research with HeLa cells. If any of these would change, we agree to notify the National Institutes of Health (NIH) under the terms of the HeLa Genome Data Use Agreement. We plan to publish the results of this study in a peer-review journal and present them in scientific conferences. In any case, we will acknowledge properly the source of the data.
Non-Technical Summary (Provided by Requestor)	Modern medicine greatly relies on understanding the molecular basis of the problem. Proteomics is capable to interrogate a variety of biospecimens for their protein contents, thereby shedding light on the underlying molecular basis. We develop new technologies in instrumentation, data acquisition, and data analysis in the realm of proteomics. The HeLa proteome is one of the most commonly used standards for technology benchmark. We will use the resource obtained from the HeLa Cell Genome Sequencing project to advance our current HeLa data analysis and related benchmarking.

National Institutes of Health
Advisory Committee to the Director
HeLa Genome Data Access Working Group
HeLa Genome Data Access Request: Project 16085

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Statistical Methods for Profiling Intra-tumor Heterogeneity from Sequencing Data**
Date Received	11/3/2017
Project Summary (Provided by NIH)	• Cancer development can result in tumors composed of cells similar to one another or tumors composed of diverse cell type. The cell type variations observed in tumors is known as tumor heterogeneity and has important consequences for cancer diagnosis and treatment. • The investigator requests access to HeLa cell genome sequence data to validate the investigator's computational model, which relies on sequencing to help characterize single cells within tumors to detect tumor heterogeneity.
Institution	University of Chicago, Chicago, Illinois
Collaborator(s)	None
Research Use Statement (Provided by Requestor)	We will develop statistical methods for intra-tumor heterogeneity from single cell sequencing data (SCseq). SCseq measures the DNA content, methylation, DNA accessibility at each single cell resolution. Due to low capture efficiency and low amount of input material, single cell sequencing experiments suffer from a substantial amount of technical noise. Developing a computational framework that can accurately profile clonality from single cell data can facilitate the transition of this novel technology to routine use. HeLa genome sequence data composes several valuable datasets that are measured using single cell sequencing. We request the usage of this dataset to test our computational model and benchmark the performance. We will develop, distribute and support software for the methods proposed in this project. The methodology and results will be reported in a publication. I will not develop a commercial product or service or file Intellectual Property (IP) based on your findings from the proposed research. My findings will not be expected to result in a commercialized product or service. I'm not expecting my plans to change regarding your intention not to seek IP or commercialization. I agree to inform the NIH if your plans for IP or commercialization change.

Working Group Finding	Consistent with the Data Use Agreement
Non-Technical Summary (Provided by Requestor)	The cancer genome is characterized by genetic heterogeneity that is seen across tumor types, among samples of a particular type (inter-tumor heterogeneity) and within an individual tumor (intratumor heterogeneity). Understanding tumor heterogeneity is a prerequisite for personalized tumor diagnosis and treatment. This project aims to develop a comprehensive suite of tools for the analysis of tumor heterogeneity using next generation sequencing data, tailored for the specific characteristics of single cell sequencing and bulk tissue sequencing.

National Institutes of Health
Advisory Committee to the Director
HeLa Genome Data Access Working Group
HeLa Genome Data Access Request: Project 14865

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Long Non-coding RNA Transcriptional Landscape in HeLa Cells**
Date Received	7/5/2017
Project Summary (Provided by NIH)	• Recent evidence suggests that a special class of RNAs, called long non-coding RNAs (lncRNAs), are implicated in various diseases. The investigators seek to understand the expression of lncRNAs and their role in cancer. • The investigators plan to use HeLa genome sequence to find out how genomic alterations affect the regulation of expression of lncRNAs in cancer cells. This should enable a better understanding of how cancer arises, and what role lncRNAs play in this process.
Institution	Center for Genomic Regulation, Barcelona, Spain
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	We plan to study the expression of long noncoding RNAs (lncRNAs) in HeLa cells, and its relationship to cancer. Specifically, we aim to relate publicly available HeLa PacBio lncRNA transcriptome data generated in-house (GEO accession: GSE93848, and doi: https://doi.org/10.1101/105064) to its matching genome. We will investigate how HeLa-specific genome rearrangements, especially HPV insertions, affect the expression of nearby lncRNAs in those cells, leading to a better understanding of the role of lncRNAs in cancer development. This work requires the precise mapping of HeLa long RNA-Seq reads to the HeLa genome, because of its numerous dissimilarities with the publicly available human reference genome. We also plan to use ENCODE HeLa RNA-Seq and epigenomic data in this study. This is a preliminary study; however, results are expected to be disseminated through public presentations and peer-reviewed scientific publications. No phenotypic characteristics will be tested for association with genetic variation. No IP or commercial product is expected to arise from this research. Should our IP, commercial plans or expectations change, we agree to notify the NIH under the terms of the HeLa Genome Data Use Agreement.

Working Group Finding	Consistent with the Data Use Agreement
Non-Technical Summary (Provided by Requestor)	We are trying to understand how a special class of human genes, called Long Noncoding RNAs (lncRNAs), are involved in cancer. LnRNAs are genes that do not code for proteins, and there is accumulating evidence that many of them are implicated in various diseases. Genome alterations, such as those present in HeLa cells, sometimes lead to cancer development. We will use the HeLa genome sequence to find out how these mutations affect the regulation of expression of lncRNAs in these cells, and compare it to healthy cells. This should enable us to better understand how cancer arises, and what role lncRNAs play in this process.

National Institutes of Health
Advisory Committee to the Director
HeLa Genome Data Access Working Group
HeLa Genome Data Access Request: Project 14603

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Methods Development for the Comparisons of Three-dimensional Genome Structure**
Date Received	6/8/2017
Project Summary (Provided by NIH)	• The investigator seeks to develop a method to quantify the similarity and differences in 3D chromosome structure across cell types and different disease states. • The investigator proposes to use the HeLa cell genome sequence to develop and validate this method to reveal the role of chromosome structure in disease, as well as to improve the understanding of how chromosomes are organized in HeLa and other cell types.
Institution	Carnegie-Mellon University, Pittsburgh, Pennsylvania
Collaborator(s)	None
Research Use Statement (Provided by Requestor)	The goal of this project is to characterize the preservation and divergence of three-dimensional (3D) chromosome structure across various cell types and disease states. We are developing a method to quantify the similarity between topological architectures of different cell lines, and to identify which regions of the chromosome are most similar and different from each other. This would allow us to characterize the structural heterogeneity of chromosomes, and analyze which factors are critical to determining chromosomal architecture. The high-resolution HeLa cell Hi-C data would help us to develop and validate this method, as well as revealing the role of 3D structure in a complex disease state. We aim to improve understanding of the similarities and differences between the HeLa chromosomal structure and that of other cell types. To that end, we will combine the HeLa data with Hi-C data of other cell types in order to compare their structures. We currently have no plans to commercialize or patent any intellectual property resulting from the use of this data, but agree to inform the NIH in the event of any IP produced from this work. The methods developed and results of the research will likely be available through publications and/or conference presentations.

Working Group Finding	Consistent with the Data Use Agreement
Non-Technical Summary (Provided by Requestor)	Recent technology has allowed us to study the three-dimensional shape of the chromosomes within the nucleus, but how that shape changes with different cell types and diseases is still unknown. We are developing a computational method to compare chromosome structures in different cell types in order to understand how much or how little this structure varies, and what factors may determine its variation. This dataset will help us to test our method and give insight into the chromosomal structure of a complex disease and its relationship to other cancer and healthy cell lines.

National Institutes of Health
Advisory Committee to the Director
HeLa Genome Data Access Working Group
HeLa Genome Data Access Request: Project 4483

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Identification of Structural Change of Chromosomes Related to DNA Copy Number Changes in Cancer**
Date Received	10/17/2017
Project Summary (Provided by NIH)	• Aberrant genomic alterations such as chromosome addition, deletion, or duplication occur frequently in cancer. Such aberrant genomic alterations are known as copy number aberrations (CNAs) and may contribute to cancer development. • To investigate the contribution of CNAs to cancer, the investigator requests to use the HeLa cell whole genome sequence to develop a computational method to search the cancer genome for CNAs. HeLa cell genomic structural images will be used to validate the CNAs identified by their computational approach.
Institution	Gwangju Institute of Science/Technology, Gwangju, Korea
Collaborator(s)	None

Working Group Finding	Consistent with the Data Use Agreement
Research Use Statement (Provided by Requestor)	With the high-resolution data sets such as single nucleotide polymorphism (SNP) microarrays, several methods were developed to identify cancer-driving genes. For instance, our previous work, the wavelet-based identification of focal genomic aberrations (WIFA), was successfully applied to SNP microarrays from glioblastoma (GBM) and lung cancer patients. It integrated DNA aberration regions from multiple samples and detected focal aberrations consistent across multiple samples with high accuracy. Since next generation sequencing data from several cancer data sets are currently available, there is a growing chance that more accurate focal aberration regions might be detected. NGS data can be used to detect structural variations in genomic regions so that we can identify copy number changes by combining with structural variations such as gene fusion, chromothripsis, and break and fusion bridges. Update on 16 October 2017 regarding HeLa cell; We are currently developing a computational method to reconstruct chromosomes with structural variations using WGS data. By combining copy number alterations and structural variations, we may improve accuracies of genome reconstruction in cancer cell. Traditionally, structural changes have been investigated using Fluorescence In Situ Hybridization (FISH). To validate that our method can successfully reconstruct cancer genomes, we attempt to compare reconstructed genomes with FISH images. Because both high resolution WGS and FISH images are available for the HeLa cell line, we would like to validate our computational method using the HeLa cell line. Dissemination: We will publish our methods and research findings in peer reviewed journal and present them in seminars. We do not have any plan to develop a commercial product or service or file Intellectual Property (IP) based on findings from the HeLa cell, and we agree to inform the NIH if our plans for IP or commercialization change based on findings from the HeLa cell.

Working Group Finding	Consistent with the Data Use Agreement
Non-Technical Summary (Provided by Requestor)	Cancer development is closely related to copy number aberrations (CNAs) in DNA. Among these copy number changes, some CNAs play more important roles than others, called driving CNAs. Identifying these driving CNAs is an important research topic since it gives a chance to find cancer related biomarkers. This research can be advanced by analyzing next generation sequencing data from multiple patients. Update on 16 October 2017 regarding HeLa cell; Cancer genomes have many structural changes including insertion, deletion, translocation, and transversion of genomes compared to the genomes in normal cells. Traditionally, these structural changes have been investigated using Fluorescence In Situ Hybridization (FISH). With availability of the sequencing technology, we can reconstruct cancer genomes using sequencing data. However, because the accuracy of reconstructed genomes in a whole chromosomal scale is still low, we need to develop a new computational method to reconstruct cancer genomes, and the accuracy of these methods can be validated using FISH data.

National Institutes of Health
Advisory Committee to the Director
HeLa Genome Data Access Working Group
HeLa Genome Data Access Request: Project 15736

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Determining Differences in the HeLa Genome at the CDR1-AS Locus**
Date Received	9/29/2017
Project Summary (Provided by NIH)	• Previous work by the investigator identified that a gene important for normal brain function is expressed at very low levels in HeLa cells, but not many other cell lines. The cause of the low expression level in HeLa cells is unclear. • The investigator proposes to compare the HeLa cell genome sequence to the human reference genome to see if genomic differences in HeLa cells can explain the low expression level of the gene.
Institution	Stanford University, Stanford, California
Collaborator(s)	None

Working Group Finding	**Consistent with the Data Use Agreement**
Research Use Statement (Provided by Requestor)	CDR1-AS is a gene now known to play a role in proper neuronal functioning. We are now studying important features of this gene, such as the repertoire of transcripts arising from this locus, the splicing patterns generated, the promoters, and regulation required for its expression. In our study, we and others have noted that this gene is not expressed highly in HeLa cells, and while we believe the epigenetic state of the promoters is the likely reason for its lack of expression, we need to rule out the possibility that there are genetic differences in HeLa cells that are responsible. It is for this reason we are requesting the HeLa genome. Specifically, we will align the region of the X chromosome (approximate coordinates: chrX:140,672,000-140,895,000) from the Human reference to the HeLa Genome to determine whether genomic differences arise that can explain the differences in CDR1-AS expression we observe (for example, whether putative promoters or enhancers are absent from the HeLa genome, or potentially important single nucleotide variants exist). This research will be disseminated in a future publication. In fact, this request is in response to reviewer comments on an already submitted paper to the journal PLOS Genetics. We may also include this work in a presentation, but we have not formalized any plans for this. In either case, presentation or publication, we will formally acknowledge the source. We have no intention to develop a commercial product or service or file Intellectual Property (IP) based on our findings and do not expect these intentions to change. We will inform you otherwise if our plans do change regarding IP or commercialization. Given that this work is basic science, we do NOT find it reasonably likely that it would be adapted for a commercialized product or service.
Non-Technical Summary (Provided by Requestor)	The human genome contains many genes whose functions are not completely understood. In addition, the exact ways in which these genes are expressed (or made into a functional product) is still unclear. We know that some genes are expressed in some cells, but not in others. The reason for this is not always obvious. Sometimes the gene's DNA is marked with a signal that tells the cell not to express it. Sometimes the cellular machinery needed to express the gene is missing. And sometimes the gene or an important part of it is absent from the genome. This last example can happen especially in cancer cells that have very fragile and unstable genomes that are prone to breakage and rearrangement. We have found that HeLa cells do not express a gene that has been found to be important for the normal functioning of the brain. We believe the reason for this is because the DNA of HeLa cells is marked in a way that restricts expression, but we cannot rule out the possibility that the absence of expression is actually because part of the genome needed to make this gene is missing in HeLa cells.

National Institutes of Health
Advisory Committee to the Director
HeLa Genome Data Access Working Group
HeLa Genome Data Access Request: Project 12668

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Mapping Controlled HeLa Transfection MS Data Over Customized Proteome Database Description**
Date Received	7/17/2017
Project Summary (Provided by NIH)	• The investigators previously identified and characterized the HeLa proteome (all of the proteins within HeLa cells) using a technique that tagged individual proteins in HeLa cells with a stable isotope of carbon and identified the tagged proteins using mass spectrometry (MS), an analytical technique that can identify chemicals in a mixture based on their physical characteristics. • The investigators propose to use HeLa cell genome sequence to validate their initial characterization of the HeLa proteome and compare the HeLa genome to the proteome to check for proteins not previously identified.
Institution	Norwegian University of Science and Technology, Trondelag, Norway
Collaborator(s)	Internal

Working Group Finding	Consistent with the Data Use Agreement
Research Use Statement (Provided by Requestor)	We have recently sequenced the proteome of HeLa cells using a SILAC approach in order to compare differences related to transfection using different plasmid expression vector systems (Hagen, Sharma et al. 2015). In order to quantify we have used Human proteome provided by Uniprot (Apweiler, Bateman et al. 2014) to match the mass spectra. However, given the divergence of genome and transcriptome data for HeLa cells (Landry, Pyl et al. 2013) we would like to re-analyze our data using the HeLa Cell Genome Sequence Data in dbGaP. We plan to use the reads from genome/transcriptome and create a custom database in order to check if there are novel peptides missed during our search based on the database provided by Uniprot (Calviello, Mukherjee et al. 2016). Any results from the above study will be disseminated in peer-reviewed international research journals. We hereby declare that we have no plans to develop a commercial service or file Intellectual Property (IP) based on our findings. It is furthermore not reasonable to expect that the findings will result in a commercial product or service. We do not expect that our plans regarding our intentions not to seek IP or commercialization, will change. We agree, however, to inform the NIH immediately if our plans for IP or commercialization change. References: Apweiler, R., et al. (2014). "Activities at the Universal Protein Resource (UniProt)." Nucleic Acids Research 42(D1): D191-D198. Calviello, L., et al. (2016). "Detecting actively translated open reading frames in ribosome profiling data." Nature Methods 13(2): 165-+. Hagen, L., et al. (2015). "Off-target responses in the HeLa proteome subsequent to transient plasmidmediated transfection." Biochimica Et Biophysica Acta-Proteins and Proteomics 1854(1): 84-90. Landry, J. J. M., et al. (2013). "The Genomic and Transcriptomic Landscape of a HeLa Cell Line." G3-Genes Genomes Genetics 3(8): 1213-1224.
Non-Technical Summary (Provided by Requestor)	It has been known that cell lines derived from cancer cells widely differ from normal cells. There can be various changes at the level of chromosomes, such as aneuploidy, translocations, deletions etc. as well as at the level of genes, including point mutations etc. We believe this might lead to changes regarding the type and quantity of proteins expressed, and that we can measure using mass spectrometry. This has potential to lead to novel findings in terms of protein expression, which can be of value to future researchers studying HeLa cells.