NIH Advisory Committee to the Director
June 14, 2018

Summary of HeLa Genome Data Access Requests

1. Project #17040, Analysis of Chromatin Organization in HeLa Cells
   University of North Carolina-Chapel Hill, North Carolina

2. Project #17520, Genome-Edition of HeLa Cells for the Study of Genes Involved in Endocytosis
   Catholic University of Louvain, Belgium

3. Project #18446, Finessing predictors of cognitive development (part 2)
   University of Auckland, Auckland, New Zealand

4. Project #17477, The Role of Architectural Proteins in Shaping the Promoter Interactome
   Babraham Institute, United Kingdom

5. Project #16442, RNA-editing Analysis in Hela S3 Cells by Single-cell RNA-seq and Machine Learning
   Shenzhen BGI Technology Company, Shenzhen, Guangdong, China

6. Project #17438, Role of Nuclear Structure in Transcription Regulation
   Duke University, North Carolina

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 17040

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Analysis of Chromatin Organization in HeLa Cells**
Date Received	1/30/2018
Project Summary (Provided by NIH)	• In the cell nucleus, DNA is tightly folded around proteins called histones and together form a structure called chromatin. The physical properties of chromatin (e.g. the tightness of compaction and proximity of different sections of chromatin to one another) has been shown to exert control over gene expression implicating chromatin organization in normal biological processes and disease. • The investigator proposes to use the HeLa cell genome sequence to improve upon existing software to detect the physical properties of chromatin (e.g. mark the frequency by which discrete sections of chromatin interact in proximity with one another) to provide a resource tool to the scientific community that will allow investigators to query the impact certain regions of chromatin have on one another in normal biological processes and disease.
Institution	University of North Carolina-Chapel Hill, North Carolina
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	We plan to use the chromatin data from the HeLa cell line to explore chromatin organization alongside of other cell lines and primary tissues as part of testing our own software to visualize this type of chromatin data (current version at http://yunliweb.its.unc.edu/HUGIn/, improvements and addition of more chromatin interaction datasets in process). Since HeLa cells are widely used in genetics research, it is important for researchers to know how chromatin is physically organized (and thus what genetic variants might reasonably be able to interact with which genes/segments of DNA) in this cell line. No sequence level data will be displayed in our browser, only statistics indicating how frequently particular 40 kb segments of DNA are expected to interact. We have published a paper already on the current iteration of our web-based tool (PMID 28582503) and plan to publish another paper describing improvements to the browser, appropriately acknowledging the inclusion of HeLa cell data. Our plans do NOT include development of commercial products or intellectual property applications; we are trying to create a free, web-based tool for the scientific community. We do not expect these plans to change, but we agree to inform the NIH if plans for intellectual property or commercialization change.
Non-Technical Summary	The DNA in each individual cell would be very long (~6 feet!) if not organized and compacted in some way; this organization can make pieces of

HeLa Genome Data Access Request: Project 17040

(Provided by Requestor)	DNA that seem like they would be far apart actually quite close together in the context of a cell's nucleus. This organization can differ in different types of cells or in different individuals. It is important for researcher's to know what pieces of DNA are in physical proximity, as this may change what genes can by influenced by a particular change in DNA. We are developing software to display the chromatin organization structure of HeLa cells and other cell lines and tissues in a user-friendly format (see current version at http://yunliweb.its.unc.edu/HUGIn/). Sequence level detail will not be available to users; only whether or not (and how often) 40,000 base pair long segments of DNA interact. We hope to create a valuable tool for researchers studying genetics and chromatin organization in many cells and tissues, including the very widely used HeLa cell line.

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

HeLa Genome Data Access Request: Project 17520

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Genome-Edition of HeLa Cells for the Study of Genes Involved in Endocytosis**
Date Received	3/2/2018
Project Summary (Provided by NIH)	• Endocytosis is an essential cellular process required for the uptake of nutrients and proteins from the cell environment into the cell. HeLa cells have been used to investigate the biological properties of the endocytic pathway. • The investigator proposes to use the HeLa cell genome to design molecular editing tools, such as the CRISPR-Cas9 editing protein complex, to investigate how endocytic pathways dynamically operate in HeLa cells.
Institution	Catholic University of Louvain, Belgium
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	Our lab is studying endocytosis in mammalian cells. Endocytosis is an essential cellular process required for uptake of nutrients from cell environment and turnover of plasma membrane components. Many endocytic pathways have been discovered over the last 30 years, depending or not on the clathrin coat protein. A big part of discoveries in the field have been made with HeLa cell line. However, most studies have been performed on cells overexpressing fluorescently-tagged versions of proteins involved in endocytosis, which may disturb those dynamic processes and induce artifacts. Nowadays, we have the possibility to label those proteins in the genome using the CRISPR-Cas9 editing approach, in order to observe them at their endogenous expression level. Unfortunately, HeLa genome is substantially different from NCBI reference human genome, which makes very difficult the design of sgRNAs or cassettes for homology-directed repair (HDR). Hence, our aim is to use the HeLa genome sequence for the design of CRISPR-Cas9 editing tools that will allow the endogenous labeling of proteins involved in endocytic processes. Overall, our studies should bring new crucial information on how endocytic mechanisms dynamically operate, as fluorescently labeled proteins produced by genome-edited HeLa cells will be observed by various microscopy techniques. We do not plan to combine this requested HeLa genome dataset with other datasets outside dbGaP. We do not anticipate intellectual property (IP) or the development of commercial products or services from our experiments. The genome information will be used exclusively for the design of effective sgRNAs, the generation of recombination cassettes for HDR, and the design of sequencing primers to verify the resulting CRISPR/Cas9-modified Hela clones. If IP should be generated from this data, NIH will be notified immediately. We plan to publish the results of our research in peer-reviewed research journals,

HeLa Genome Data Access Request: Project 17520

	and we will acknowledge use of the HeLa genome sequence in all publications and presentations according to NIH guidelines.
Non-Technical Summary (Provided by Requestor)	Endocytosis is an essential cellular process required for uptake of nutrients from cell environment and turnover of plasma membrane components. Many endocytic mechanisms have been discovered over the past 30 years, and are still being discovered and characterized today. The recent advances in genome-editing techniques open the door to the endogenous labeling of proteins involved in these processes. In order to design genome editing tools, the correct genetic sequence of the specific cells being edited must be known.

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

HeLa Genome Data Access Request: Project 18446

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**Finessing Predictors of Cognitive Development (part 2)**
Date Received	5/24/2018
Project Summary (Provided by NIH)	• In In the cell, DNA Is compacted into organized structures called chromatin, which form 3-D architectural structures within the cell. Previous studies have shown that manipulation of the 3-D genomic architecture can influence normal development as well as disease. • How genomic architecture influences cognitive development is poorly understood. The investigator proposes to characterize the regulatory networks associated with cognitive development by comparing the HeLa cell genome sequence that dictates the 3-D architecture of the HeLa genome ("Hi-C sequence") to brain cell genomic architectural data to identify regions of the genome that play a role in cognitive development.
Institution	University of Auckland, Auckland, New Zealand
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	We are requesting access to phs000640: The chromatin contact map in Hela cells. The HeLA Hi-C data were published in Rao et al (2014; doi:10.1016/j.cell.2014.11.021). The data will be used for Health Medical research as outlined below. Cloud computing will not be used. As for most polygenic disorders the genetic variants that are associated with mental conditions often fall outside of genes. We propose to investigate the regulatory network for SNPs associated with mental disorders. Our objectives are to incorporate Hi-C data on genome organization with functional read-outs (eQTL data) to identify regulatory networks that are impacted by cognition associated SNPs. The HeLa cell Hi-C data represents a critical dataset for the identification of conserved and unique functional spatial regulatory connections. The numbers of High-resolution Hi-C maps that are available to data mine are slowly increasing in number. The HeLa data we are requesting represents one of the highest resolution Hi-C maps that is currently available. The use of this data will enable the identification of possible spatial regulatory connections between enhancers and genes that form within the human genome. By comparison with those connections that form in other cell types and tissues, we will be able to identify tissue/cell-type specific and common connections. Extension with the GTeX database, will identify those connections that are functional in terms of direct gene regulation. Our findings will be disseminated through conference presentations and peer reviewed publications. The data that will be shared as part of this will only refer to regions of the genome that form contacts (or do not form contacts)

HeLa Genome Data Access Request: Project 18446

	with specific genes and the functional outcomes of these physical interactions. We do not have any plans to develop a commercial product or service or file Intellectual Property based on our findings from the proposed research. We do not foresee this work resulting in a commercialized product or service. We do not expect that our plans will change regarding our intention to seek IP or commercialization but we will inform the NIH if our plans for IP or commercialization change
Non-Technical Summary (Provided by Requestor)	The causes of mental disorders remain poorly described. However, it is likely to be a combination of genetic and environmental factors that contribute to their development. As for most complex disorders the mutations that are associated with mental conditions often fall outside of genes. We think that these mutations work by changing the three-dimensional wiring of cells. We will join together complex datasets to untangle the wiring networks of brain cells. We will then perform a series of experiments to prove that these networks are correct. This study will tell us how mutations affect the gene networks that direct development.

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

HeLa Genome Data Access Request: Project 17477

Working Group Finding	**Consistent with the Data Use Agreement**
Project Title	**The Role of Architectural Proteins in Shaping the Promoter Interactome**
Date Received	2/28/2018
Project Summary (Provided by NIH)	• Previous work has shown that specific DNA conformations influence transcriptional activity, an early step in gene expression, such as topologically associated domains (TADs) and specific DNA looping interactions between DNA fragments and promoters, the start site of transcription. • The investigator proposes to use the HeLa cell genome sequence, in particular the sequence that describes the 3D chromatin architecture of the HeLa genome ("Hi-C sequence"), to use as a reference to validate their approach that seeks to identify critical regions that control DNA conformation (i.e. DNA looping and TADs), transcriptional activity, and gene expression in HeLa cells.
Institution	Babraham Institute, United Kingdom
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	The way DNA is folded in the nucleus is of major importance to genome integrity and transcriptional activity. Multiple levels of DNA spatial organisation have been identified, from chromosomal territories and topologically associating domains (TADs) to specific looping interactions between individual DNA fragments. In a recent publication (Javierre et al. Cell 2016) we showed that TAD boundaries show pronounced, albeit incomplete insulation of the DNA looping interactions involving gene promoters, including enhancer-promoter interactions. We now want to study the effects of perturbations in the binding of architectural proteins on specific promoter-anchored interactions. The HeLa cell line provides an excellent model for our experiments, since HeLa cells, in which CTCF and Cohesin can be readily depleted, are available (Wutz et al., EMBO J 2017). In the proposed project, we will study the promoter interactome using promoter capture Hi-C in these cells. We would therefore like to request access to the HeLa genome sequencing data published by Landry (Landry et al. Genes, Genomes, Genetics 2013) to map Capture Hi-C data to the appropriate reference. The proposed analyses will provide novel insights at an unprecedented resolution into the dynamics of genome organization and promoter interactions response to perturbations of architectural proteins. This, in turn, will further our understanding of the mechanisms underlying transcriptional regulation (with implications for carcinogenesis). This research is of a fundamental nature and we aim to gain and disseminate scientific knowledge rather than patent the findings. We will however

HeLa Genome Data Access Request: Project 17477

	inform the National Institute of Health if this study results in intellectual property that can be exploited commercially, in accordance with the HeLa Genome Data Use Agreement. For our publication history please see: http://www.babraham.ac.uk/our-research/nuclear-dynamics/mikhail-spivakov/publications
Non-Technical Summary (Provided by Requestor)	Each cell in the human body contains approximately two meters of DNA, which is folded into the very small space of the cell nucleus. The exact pattern of DNA folding plays a very important part in activating and switching off genes. Problems with DNA folding can cause cells to function improperly and may even cause cancer. We are investigating the mechanisms by which genes are regulated by the three-dimensional structure of DNA. Specifically, we are looking at the changes in chromosomal organization in response to depletion of two proteins that facilitate DNA-DNA interactions: Cohesin and CTCF. We have data on the chromosomal architecture in HeLa cells under normal conditions and when Cohesin and CTCF are removed. We want to map the location of these elements onto the HeLa genome to better understand how these proteins control DNA folding.

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

HeLa Genome Data Access Request: Project 16442

Working Group Finding	Consistent with the Data Use Agreement
Project Title	**RNA-editing Analysis in Hela S3 Cells by Single-cell RNA-seq and Machine Learning**
Date Received	11/28/2017
Project Summary (Provided by NIH)	• RNA editing is a molecular process that can make discrete changes to specific nucleotide sequences within a RNA molecule after it has been generated from DNA. • The investigator requests access to HeLa cell genome sequence to develop a computational tool that discovers locations within the genome that have undergone RNA editing. Additionally, the investigator proposes to use single-cell RNA sequencing technologies to identify unique RNA editing sites in HeLa cell populations and compare them to HeLa cell genome sequence.
Institution	Shenzhen BGI Technology Company, China
Collaborator(s)	Internal
Research Use Statement (Provided by Requestor)	Hela S3 is infected by HPV, but whether the RNA editing level is the same with normal cell or not is still unknown. Since the Hela S3 genome and RNA-seq data is available and there are many research found that there is much heterogenicity among single cell, we then suspect that RNA editing may contributed to this phenomenon. So we plan to develop a bioinformatic method aimed to predict the RNA-editing sites in Hela S3 cells more efficiently and accurately. We wish to use this data set to evaluate the performance of our tool, which will then be published and shared with the scientific community. Besides Hela S3 whole genome data, we also need Hela S3 RNA sequence data which we already have by RNA-seq using Hi-seq. We plan to combine the genome and RNA data of Hela S3 cell, and to explore the potential RNA editing site by compare those two data sets. It creates no additional risks to participants since it is a mature pipeline many people done nowadays. Furthermore, if we fortunately find some interest RNA editing site, we will use those information to explain the basic phenomenon in our body which we think would contribute greatly to the development of basic science. So that is meaningful for our group and biological research. We do not have any plan to develop a commercial product or service or file intellectual property, I do not think our findings would result in a commercialized product or service. It is possible that our plan will change when there is no obvious difference between Hela S3 genome and RNA when it comes to RNA editing, but it is acceptable since no one have study this point, if so, we will search other tissue or cells like neuron to verify whether there are different for RNA editing by ADAR1.I totally agree that NIH is informed when we plans for IP or commercialization change.
Non-Technical Summary	Viral infection causes multiple forms of human cancer, and HPV infection is the primary factor in cervical carcinomas. Recent single-cell RNA-seq studies highlight the tumor heterogeneity present in most cancers, but virally induced

HeLa Genome Data Access Request: Project 16442

(Provided by Requestor)	tumors have not been studied. HeLa is a well characterized HPV+ cervical cancer cell line. We want to find the heterogeneity about RNA-editing in Hela cells.

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

HeLa Genome Data Access Request: Project 17438

Working Group Finding	Consistent with the Data Use Agreement
Project Title	**Role of Nuclear Structure in Transcription Regulation**
Date Received	3/14/2018
Project Summary (Provided by NIH)	• In the cell, DNA is tightly compacted around proteins called histones. Together, the compacted DNA and histones form a structure called chromatin. Special nuclear proteins associate with the chromatin to modify its overall architecture, which influences gene expression by controlling how regions of chromatin interact with each other and the genes in those regions. • The investigator proposes to use the HeLa cell genome sequence, in particular the sequence that dictates the 3D architecture of the HeLa genome ("Hi-C sequence"), to evaluate the role of nuclear architectural proteins that associate with chromatin and change gene expression.
Institution	Duke University, North Carolina
Collaborator(s)	None
Research Use Statement (Provided by Requestor)	We would like to use the HeLa genome as a reference genome for various high-throughput experiments we perform in HeLa cells, particularly gene expression profiling (RNA-seq) and chromatin immunoprecipitation (ChIP)-Seq assays. We are interested in defining novel nuclear architectural proteins that participate in gene regulation and understanding how they functionally impact transcription, and chromatin structure. Our studies thus would benefit highly from having access to the HeLa genome and HeLa Hi-C topology data (phs000640) that will collectively provide a comprehensive understanding on mechanisms that causally link nuclear architecture to regulation of chromatin topology and chromatin structure during gene expression. We do not anticipate any foreseeable IP or commercial products/services arising from our research, but we agree to notify the NIH if this changes. Any research findings from these studies will be disseminated via publications and presentations.
Non-Technical Summary (Provided by Requestor)	Proteins that form the nuclear architecture associate with chromatin (DNA wrapped around proteins called histones) and regulation of these interactions is critical for expression of genes. We are interested in using the HeLa genome and HeLa chromatin organization (HiC) data to evaluate role of novel nuclear architecture proteins in gene expression.