NIH Advisory Committee to the Director
September 5, 2014

Summary of HeLa Genome Data Access Requests

1. Analysis of Gene Expression, Splicing Regulation, and mRNA Degradation in Human Cells
   University of California Berkeley

2. Development of Methods to Infer the 3D Structure of the Genome
   University of Washington

3. Searching for Infectious Cancer Agents
   University of Pittsburgh

4. Polymorphisms in the Shifted Self Peptidome Following Viral Infection
   Vanderbilt University; University of Texas, El Paso

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

HeLa Genome Data Access Request

Working Group Finding	Consistent with Data Use Agreement

Project Title	**Analysis of Gene Expression, Splicing Regulation, and mRNA Degradation in Human Cells**
Date Received	May 2, 2014
Requestor's Organization	University of California Berkeley
Project Overview	• The requestor plans to evaluate the effects of variations in the HeLa genome with that seen in other human cell lines, in order to learn more about gene regulation, structure and function.
Research Use Statement (Supplied by Requestor)	We would like to use the HeLa genome as a reference genome for various high-throughput experiments performed in HeLa cells, particularly short read transcriptome sequencing (RNA-seq). Given the differences in this genome compared to the reference human genome, it is ideal to use the HeLa genome in the analysis of the RNA-seq data, especially since we are particularly interested in novel splice variants and proper read mapping is important. We also want to investigate the HeLa-specific characteristics of genes related to our systems of interest. The projects we will use the genome for include investigating the global prevalence of alternative splicing coupled with nonsense-mediated mRNA decay in human cells, building a splice factor regulatory network, and pharmacogenomics and genome variation projects which would benefit from having the HeLa genome for mapping high-throughput sequencing reads and for investigating the differences between it and the reference genome. 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 are likely to be disseminated via publications and presentations.
Non-Technical Summary (Supplied by Requestor)	We are investigating the regulation of gene expression, particularly the mechanisms by which genes can produce multiple forms of mRNAs, which then can be either degraded or translated into different proteins that have different functions. We primarily use high-throughput sequencing of the mRNAs in order to investigate their structures and how these structures change when the regulatory system is perturbed. Since our main model system is the HeLa cell line, the HeLa genome will help us differentiate results specific to the cell line from those more general to all human cells.

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

HeLa Genome Data Access Request

Working Group Finding	Consistent with Data Use Agreement

Project Title	**Development of Methods to Infer the 3D Structure of the Genome**
Date Received	May 20, 2014
Requestor's Organization	University of Washington
Project Overview	• The overall goal of the project is to use the HeLa genome dataset to test hypotheses about three-dimensional folding of chromosomes, which is important for cellular function. • The knowledge and complex nature of the HeLa genome are valuable for evaluating algorithms to infer three-dimensional genomic architecture. • The requestor will collaborate with a Mines ParisTech investigator, who has submitted an independent Data Access Request in accord with the HeLa Genome Data Use Agreement.
Research Use Statement (Supplied by Requestor)	We will use this Hi-C data set as a means to test our algorithmic approaches to inferring 3D genome architecture. The complex karyotype and detailed haplotype information provide an excellent framework for evaluating structural inference algorithms. • The HeLa cell sequence data are valuable for the proposed research because (1) Hi-C data is available, which is true for only a handful of cell types, (2) this cell type's haplotype structure has been fully characterized, and (3) this cell type's karyotype is very complex, making it a particular challenge for structure inference algorithms. 1. We do not anticipate IP or the development of commercial products or services. 2. We not foresee that IP or commercial products or services arising from our research with HeLa cells. 3. If our IP or commercial plans or expectations change, we agree to notify the NIH. • We plan to disseminate the results of our proposed research in scientific journals and, potentially, at scientific conferences.
Non-Technical Summary (Supplied by Requestor)	Several lines of evidence suggest that the 3D structure of DNA in the nucleus of the cell is important for cellular function. A recently described assay called "Hi-C" provides pairwise measurements of contacts in 3D between distant locations along the genome. We are studying how to infer the complete 3D structure of the genome on the basis of this type of pairwise contact data.

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

HeLa Genome Data Access Request

Working Group Finding	Consistent with Data Use Agreement

Project Title	**Searching for Infectious Agents**
Date Received	June 18, 2014
Requestor's Organization	University of Pittsburgh
Project Overview	• The requestor would like to search the genomic sequence of various tumors to potentially discover new types of viral infections that cause cancer. • Tumor sequences will be compared to normal human and HeLa cell genome sequences to identify new genomic variations associated with infectious agents.
Research Use Statement (Supplied by Requestor)	Viruses, such as HPV, are known to contribute to cancer. In addition, unidentified agents are present in tumors and discerning their contribution to cancer has not been pursued. During mapping to the human genome, many sequences are discarded because of genomic rearrangements, chimeric reads, or contaminating sequences. Also, sequences of infectious agents present in human samples will be discarded. We built tools to search the discarded sequences for viruses and evidence of their integration. We identified 400 samples in TCGA that contain sequences and integrations of viruses. With continued access to TCGA (phs000178.v8.p7.c1), we will explore the role of virus integration in determining tumor behavior, and search environmental metagenomes for agents present in cancer. Our goals are to determine: 1) how virus integration influences viral and cellular gene expression. We will determine the coding potential of chimeric human-virus transcripts and study how orientation of integration influences patterns of cellular gene expression 2) if sequences found in metagenomic and TCGA data represent novel infectious agents. We have assembled sequences, likely representing novel viruses, from metagenomic and TCGA databases. We will assess their phylogeny and genome structure to determine whether these sequences are present in tumors as a contaminant or infectious agent. We will correlate all our observations with TCGA clinical data. We do not forsee that our analysis of the requested datasets will create any additional risks to participants. We will use the HeLa genome sequence (phs000640.v2.p1.c1) to determine the exact nucleotide sequence of the integrated portion of HPV18 and to obtain a list of SNPs compared to the human reference genome. Neither of these goals is possible without access to the HeLa genome sequence. We do not anticipate nor forsee IP or the development of commercial products or services from our research with the HeLa genome. We agree to notify the NIH under the terms of the HeLa Genome Data Use Agreement if our IP or commercial plans change. We will dissemintate our research

HeLa Genome Data Access Request

	findings through publications and presentations as appropriate
Non-Technical Summary (Supplied by Requestor)	Most microorganisms that exist on planet earth have not yet been discovered. These agents can be detected by genetic signatures present in metagenomic sequencing experiments. We will search these metagenomic sequences for agents present in human tumors.

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

HeLa Genome Data Access Request

Working Group Finding	Consistent with Data Use Agreement

Project Title	**Polymorphisms in the Shifted Self Peptidome Following Viral Infection**
Date Received	July 1, 2014
Requestor's Organization	Vanderbilt University; University of Texas, El Paso
Project Overview	• The requestor plans to study how the set of proteins found on the surface of cells change after a viral infection and how these vary among the population. • HeLa cells, which will be used as a control, are an example of a cell that has been infected by a virus and displays abnormal surface proteins.
Research Use Statement (Supplied by Requestor)	The objective of this research proposal is to determine the degree of polymorphisms present in self peptides differentially displayed before and after viral infection. Identification of the self peptides presented by MHC class I molecules before and after infection of a derivative of HeLa cells with vaccinia virus revealed a profound shift in presentation. This could have a profound effect on allopeptide specific immune responses if polymorphisms are significantly present within the human population. Therefore, our data set will be searched against the human genotype and haplotype databases. Searching against the HeLa genome will serve as a control and allow us to rule out any sequences unique to HeLa cells. This control is important to eliminate any profound variation caused by the use of cancerous HeLa cells. This study will be observational only and not result in IP; however, if that changes the NIH will be notified pursuant to the terms of the HeLa Genome Data Use Agreement. This work will be disseminated through publication and presentation with appropriate acknowledgements to the HeLa Genome.
Non-Technical Summary (Supplied by Requestor)	Short proteins (peptides) are displayed by cells to report to the immune system the internal workings of the cell. If something changes then these peptides change too. When all is well, these peptides are pieces of the cells own proteins. In an infection, they are parts of the microbe's proteins. We have found that after a virus infection the peptides from the cell do not return to normal. This can be harmful for transplants and other medical conditions if not everyone has the same peptides. We, therefore, want to know if these cell peptides vary among the population. Using the HeLa genome as a control (since the virus infected a derivative of HeLa cells), we will look for differences among the population at these cell peptide locations.