Budget Justification

Questions to be addressed for personnel

There are three Postdocs that are fairly new and/or recently joined the lab, correct?  Dr. Harmanci was reflected in the -02 progress report for 12.00 cal. months.

Dr. Gursoy is reflected as 3.8 cal. months in the carryover.

**Will there be two post docs working on the grant to compensate for Harmanci’s duties?**

Dr. Kong was listed on the **personnel report of -02** progress report as 1.00 cal. month but the carryover request indicates Dr. Kong recently joined the lab and plans to commit 2.5 cal. months.

**When did he/she join the lab?  What was his/her effort at the time of start on this grant?**

**Can you clarify what Drs. Gursoy, Kong, Kumar will be doing on this grant during the carryover?**

**What part of their duties and associated AIMS could not get done due to the shortened period?**

Personnel:

Dr. Gamze Gursoy, Ph.D., Postdoctoral Associate (3.8 calendar months). The overall coordination of the project and analysis proposed in this grant has been performed by Dr. Arif Harmanci. However, Dr. Harmanci recently left Yale for a faculty position. Dr. Gursoy joined the Gerstein lab mid November 2016 and has since been trained by Dr. Harmanci on the ongoing effort of this project. Dr. Gursoy has extensive experience with functional genomics data and strong background in scientific computation. As part of her Ph.D thesis, she developed advanced methods and tools for construction of 3D organization of genome integrating various functional genomics assays including RNA-Seq. In the Gerstein laboratory, she is currently in process of taking over the overall coordination, supervision and analysis of this project under the direction of Dr. Gerstein. Her responsibilities will include developing the informatics theoretic quantification of sensitive individual characterizing information. Dr. Harmanci could not finish the formalism for quantifying information leakage (Aim 1) due to the shortened period. Dr. Gursoy will finish the mathematical framework during the carryover period. She needs to perform variant calling from RNA-Seq data that requires extensive computing time and storage.

Dr. Xiangmeng Kong, Ph.D., Postdoctoral Associate (2.5 calendar months). Dr. Kong has joined the Gerstein lab March 2017 for her expertise in scientific computing. Before joining the Gerstein lab, she obtained her Ph.D degree from chemistry department of Yale University, focused on computational quantum chemistry. She is currently working on characterizing structrual variations using personal genomes. Understanding the format and information content of personal genomes makes her qualified to assist in developing new file formats that enable efficient and effective distribution of molecular phenotyping datasets in a privacy aware-manner. She was already involved in this project and worked on coordination with GA4GH by attending to GenoPri conference and iDASH. During the carryover period, she will be helping with the development pf practical software based on the mathematical formalism developed by Dr. Gursoy (Aim 1).

Dr. Sushant Kumar, Ph.D., Postdoctoral Associate (2.5 calendar months). Dr. Kumar has extensive experience with the relationship between genotyping and phenotyping data as well as scientific computing. He received his Ph.D degree from Penn State University, focused on protein dynamics using molecular dynamics simulations. In the Gerstein lab, he has developed new algorithms to analyze the relationship between genotypes and the local structures of proteins. He gained extensive experience in detection and analysis of genotyping data using various sequencing data. As variant calling algorithms take long computational time, Dr. Kumar will help Dr. Gursoy with variant calling and comparison of the real and leaked genotype sets (Aim 1).

Dr. Gursoy and Dr. Kumar will be replacing Dr. Harmanci’s duties in the carryover period.

We have previously suggested to purchase **Dell Poweredge R815 computing server. However,** The Yale Center for Research Computing (YCRC) updated their machines and they no longer use Dell Poweredge R815 servers, therefor this purchase has not been completed. We previously requested only computing servers and we did not anticipate that we will need more computer storage at the time. As the increase in the number of projects conducted in the lab, we generated more data and needed more storage. **To be able to successfully give an overall understanding of private information leakage from RNA-Seq data as well to come up with effective privacy-preserving file format system, we realized that we need to calculate the leakage from every step of the data processing pipeline. This includes accessing the raw read files, aligning them to human genome, post alignment processing and genotype calling. Each of these steps require extensive amount of computational storage and time. We also need to store our privacy-aware file formats in our system and do careful comparisons between the original files and the new files to make sure new file formats result in similar biological results as the original files.**

**We already purchased the following equipment for computing servers and storages and we will use the requested funds to pay for these purchases:**

* Red Hat Enterprise Linux OpenStack Platform for Controller Nodes with
Smart Management (w/o guest OS), Standard (2-sockets), 1 Year
Equipment - Computer - Asset Purchase (SC025)
$3,864.40
* Lenovo Purchase $38K servers, $15,000
* Lenovo GS220 GPFS storage system for the Farnam cluster, with 168 10TB drives, for a total of 1.2-1.3PB usable capacity, $11,397

[remove below?]

Questions to be addressed for equipment:

**What is the status of the Dell Poweredge R815?  Was this purchased, if so when?  If not, why?**

**GG: I am guessing this was in the year -02 equipment. I asked MH if we ever bought this.**

**What “computer equipment” are you referring to regarding this research?**

**GG: I asked MH a list of purchases done using this grant with purchase dates.**

**Clarify what this server is and how does it relate to the project?**

**GG: We are calculating the overall information leakage from raw RNA-Seq data and other functional genomics experiments at different coverage for a fair comparison. To do so, we need to store large raw read files and their sub-sampled alignments to the server to do genotyping calls.**

**What is the difference between this storage system and the Poweredge?**

**GG: Figuring it out with MH**

**Why are you asking for $20,000 more than the original amount?**

**GG: At the time, we were using our storage in HPC that is shared among other projects. However, with the increase in the number of big projects Gerstein Lab became part of last couple years, we ran out of torage space in our servers and we did not anticipate this need when we were writing our original budget.**

**This server was not part of the original grant application, How does this relate to the original scope of the grant that was reviewed and approved?**

**How is this an immediate need?  What part of the research couldn’t get done as result of the shortened period?**

**GG: To be able to successfully give an overall understanding of private information leakage from RNA-Seq data as well to come up with effective privacy-preserving file format system, we realized that we need to calculate the leakage from every step of the data processing pipeline. This includes accessing the raw read files, aligning them to human genome, post alignment processing and genotype calling. Each of these steps require extensive amount of computational storage and time. We also need to store our privacy-aware file formats in our system and do careful comparisons between the original files and the new files to make sure new file formats result in similar biological results as the original files.**

**SUBCONTRACT**

University of California, Berkeley – see separate budget and justification.

INDIRECT COST

Indirect costs are calculated at Yale’s federally negotiated rate of 67.5% of modified total direct costs. DHHS agreement dated 02/16/2017.