Justifications for BD2K Grant – Gerstein (Yale)

Mark Gerstein, Ph.D. PI (.45 summer months). Dr. Gerstein is the Albert Williams Professor of Biomedical Informatics. His lab (http://gersteinlab.org) was one of the first to perform integrated data mining on functional genomics data and to do genome-wide surveys. His tools for analyzing motions and packing are widely used. Most recently, he has designed and developed a wide array of databases and computational tools to mine genome data in humans, as well as in many other organisms. He has worked extensively in the 1000 genomes project in the SV and FIG groups. He also worked in the ENCODE pilot project and currently works extensively in the ENCODE and modENCODE production projects. He is also a co-PI in DOE KBase and the leader of the Data Analysis Center for the NIH exRNA consortium. In these roles Dr. Gerstein has designed and developed a wide array of databases and computational tools to mine genomic data in humans as well as in many other organisms. He will lead the overall informatics effort in the project.

Dr. Arif Harmanci, Ph.D., Assoc. Research Scientist (12 calendar months). Dr. Harmanci has extensive experience with bioinformatic approaches to genome-wide analysis and a strong background in scientific computation. As part of his PhD thesis, he developed advanced methods for RNA secondary structure prediction. In the Gerstein laboratory, he has developed new algorithms to identify transcription factor binding peaks from ChIP-Seq data. He is currently working on transcriptome, epigenome, and variant analysis of several large scale RNA-seq, DNA-seq, and ChIP-Seq datasets that include the Geuvadis dataset (RNA-Seq on 500 individuals), TCGA dataset, and ENCODE datasets. He will work on the analysis proposed in the grant under the direction of Dr Gerstein. He will work on developing the information theoretic quantification of sensitive individual characterizing information.

Jieming Chen, Postdoctoral Associate (4 calendar months) received his PhD in the Computational Biology and Bioinformatics Program and the Integrated Graduate Program in Physical and Engineering Biology at Yale University in 2015. Prior to Yale, he graduated from the National University of Singapore in 2008 and worked for two years at the Genome Institute of Singapore, where he has co-authored six publications in population genetics, particularly in human population structure and pharmocogenomics. He had also participated in a number of local and international consortia, including the Singapore Genome Variation Project, Pan-Asian SNP Consortium and the International Stem Cell Consortium. His current research focuses on the development of tools and large-scale genome-wide computational analyses of personal genomes. In collaboration with Lynne Regan's Lab at Yale, this involves the integration of a diversity of research perspectives, including protein structures and exomes, allele-specific regulatory elements in the non-coding genome and biological network analyses. Thus far, he has coauthored six scientific publications and is part of the 1000 Genomes Project Consortium. He will work on developing and evaluating the different linking attack scenarios utilizing specific instantiations of genotype prediction and linking approaches.

Sushant Kumar, Postdoctoral Associate (7 calendar months) received his PhD in Bioinformatics from the Pennsylvania State University. As part of his PhD thesis, Sushant applied coarse-grained simulations to systematically investigate the folding and binding process of disordered

FCP1 protein. He will work on development of anonymization strategies for protecting the phenotype datasets against linking attacks.

Timur R. Galeev, Ph. D., Postdoctoral Associate (12 calendar months). Dr. Galeev has a strong expertise in scientific computation. Before joining the Gerstein lab at Yale University, he obtained his Ph.D. degree (2014) from Utah State University. His Ph.D. research was in the field of theoretical and computational physical chemistry, focused both on applications of modern electronic structure methods and development of new theoretical tools and models of molecular structure and bonding. He is currently working on analysis of functional genomics data. Dr. Galeev will work on developing new file formats that enable efficient and effective distribution of molecular phenotyping datasets in a privacy-aware manner.

Fringe benefits are calculated at the rate of 31% for the PI and the Associate Research Scientist and Postdoctoral Associates according to the University guidelines.

EQUIPMENT

In year 1, we need a Dell Poweredge R815 server with 160GB of memory and four AMD Opteron processors. This will be used for processing our data and digital visualization. This is needed to complete the proposed research and will solely benefit this project.

SUPPLIES

We are budgeting an incremental amount of supplies for the individuals named above. This supplies budget will be used to cover computer supplies for them, covering such expenses as: diskettes, tapes, and other miscellaneous computer parts (e.g. replacing worn out surge suppressors), software upgrades, web hosting and "cloud computing fees, and reprint charges. These items are needed to complete the proposed research and will solely benefit this project.

TRAVEL

As this is a collaborative project, we are budgeting considerable funds for travel between sites. Here we are requesting incremental funds for each of the FTEs for airfare, lodging and meal expenses to attend scientific meetings annually that benefit the project. In particular, the travel will include at least 1 trip per year to a scientific meeting of in genomics and bioinformatics such as the ISMB or CSHL Biology of Genomes.

INDIRECT COST

Indirect costs are calculated at Yale’s federally negotiated rate of 66.5% of modified total direct costs. DHHS agreement dated 05/19/2015.