*** Descriptive title of proposed activity**
Evaluating the functional impact of coding and non-coding somatic mutations over multiple scales

*** Name(s), address(es), and telephone number(s) of the PD(s)/PI(s)**
Haiyuan Yu, PhD
Department of Biological Statistics and Computational Biology
Weill Institute for Cell and Molecular Biology
Cornell University
335 Weill Hall
Ithaca, NY 14853
Phone: 607-255-0259
Email: haiyuan.yu@cornell.edu

Mark B. Gerstein, PhD
Computational Biology & Bioinformatics Program
Yale University
MBB, PO Box 208114
New Haven, CT 06520-8114 USA
Phone 203-432-6105
Email: mark.gerstein@yale.edu

*** Names of other key personnel**
Andre Levchenko, PhD
John C. Malone Professor of Biomedical Engineering
Director of Yale Systems Biology Institute
Director of Cancer Systems Biology@Yale (CaSB@Yale)
Yale University

Sidi Chen, PhD
Yale Systems Biology Insitute
Department of Genetics
Yale University

Mark Rubin, MD
Director, Caryl and Israel Englander Institute for Precision Medicine
Homer T. Hirst Professor of Oncology in Pathology
Cornell University

*** Participating institution(s)**
Cornell University
Yale University

*** Number and title of this funding opportunity:**
PAR-16-131
Emerging Questions in Cancer Systems Biology (U01)

In this proposal, briefly, we plan to develop mathematical models to prioritize and rank non-coding and coding mutations in similar terms. These models will rank the impact of mutations causing cancer in terms of their underlying genomic alteration. We will then assay the actual phenotypes produced by these mutations on three scales: molecular activity, cellular phenotypes, and phenotypes in cultured organoids. Doing these experiments will produce a data resource of prioritized mutations and iterated mathematical models for prioritizing them as a product. It will also allow us to address a number of questions about cancer.

First of all, cancer genomics has revealed that there are often thousands of mutations per tumor genome but only a small fraction of them are in coding regions. Yet, almost all of the known driver mutations in cancer are in coding regions. Is this because, fundamentally, non-coding mutations have less impact than coding ones, or just simply because of an ascertainment bias on our part?

Second of all, is it the case that a mutation prioritized to give a strong impact in terms of effect on molecular networks binding will also have a strong effect on cellular phenotype and this will have also a strong effect on organismal phenotypes such as contracting cancer. It's not clear that we'll see a similarity between these three levels and we will be able to ascertain that here.

To focus our analysis, we will prioritize both coding and noncoding variants in linked enhancers and promoters on a matched set of genes, including both validated and putative cancer drivers, as well as some control genes with no known cancer association. Non-coding mutations are potentially directly involved in our regulatory networks sitting in regulatory regions of the genome and they can be matched, in a system sense, to many of the coding mutations which directly effect protein-protein interfaces involved in protein networks. One question we will investigate is 'Are these mutations in any sense comparable or are, fundamentally, the coding mutations more deleterious?'

AIM 1 Computational prioritization of coding and non-coding somatic mutations

First, we will do this in a classical sense by looking for mutations under positive selection in cohorts that are recurrent in particular regions of the genome i.e. in particular domains of a protein or in particular non-coding elements and to do this we will use the recently constructed large datasets, e.g. from TCGA and PCAWG consortia. We will also prioritize mutations computationally by looking at their sequence level molecular impact. This will be done from using a variety of metrics such as: the degree to which the mutation directly breaks the functional site i.e. breaks the TF motif or protein-protein bind interface; the degree to which it effects central positions in the overall network; the degree to which it's associated with a site that has an obvious allelic effect and sensitivity to sequence; the degree to which it sits in a functional element; and the degree to which it shows obvious conservation across organisms or within the human population, for instance as measured from GERP score.

From the combination of positive selection and functional impact, we will develop mathematical models to prioritize mutations and lists to prioritize mutations that we will then hand off to the validation components of the proposal. We will take the results each year from the validation components and use it to refine our models by a variety of simple iterate machine running tactics such as a Bayesian or online conjugate gradient updates.

AIM 2 High-throughput *in vitro* quantification of molecular phenotypes of ~2500 non-coding and ~1500 coding mutations

We will select ~500 coding and ~1000 non-coding mutations and subject them to a number of high-throughput in vitro assays to look at their molecular readout. We will take advantage of our novel Clone-seq pipeline to generate these mutant clones in large-scale. As an integral part of the Clone-seq pipeline, each mutant clone will be fully sequence verified by next-generation sequencing to ensure quality. Furthermore, we will assay the non-coding mutations using eSTARR-Seq and Promoter-seq the coding mutations to quantify their effect on enhancer and promoter activities. We will also assay the coding mutations using our high-throughput protein-protein interactome screening methodology described in our previous publications[8-11], *IN*tegrated *PrO*tein *IN*teractome per*T*urbation screening (InPOINT). This pipeline combines six different functional assays to examine experimentally the impact of hundreds of coding variants on protein stability and specific protein-protein interactions. From this we will be able to rank this pool of ~1500 variants in terms of their strongest molecular readouts.

AIM 3 Medium-throughput *in vivo* quantification of cellular phenotypes of ~300 mutations using cell growth and migration assays with CRISPR/Cas9 mutagenesis

In this aim we will look at cellular phenotypes associated with the mutations. We will evaluate ~150 coding and ~150 non-coding mutations in terms of their phenotypes for cell growth and also invasiveness, which is related to metastasis, using a variety of cell-based assays. The mutations will be introduced into CCD-18Co cells through CRISPR/Cas9 mutagenesis.

AIM 4 *In vivo* validation of 10 coding and non-coding mutations using CRISPR/Cas9 knock-in colon organoids

In aim 4, we will select the top 10 coding and non-coding mutations and evaluate them in a realistic tissue system – organoids derived from normal colon samples. We will see if these mutations are actually associated with promoting cancer in this model system and then show the degree to which we can find non-coding mutations with as much functional impact as coding ones. We will further investigate the mechanisms through which mutations lead to cancer. For non-coding mutations, we will test alterations in transcript levels, H3K27Ac/H3K4me3 marks and transcription factor binding, comparing gene-edited and isogenic control colon organoids. For coding mutations, we will perform co-IP, protein stability and selected functional assays in gene edited and isogenic control organoids. Throughout the process, we will feedback the results of each of the assays into our overall computational model and prioritization scheme developing a more accurate scheme. So with each year of the grant we will develop a more accurate model, eventually culminating near the end of the grant with a highly accurate model and a refined prioritization list.