**SPECIFIC AIMS**

In this proposal, we will 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 alterations and recurrence in cancer cohorts. We will then assay the actual phenotypes produced by these mutations on three scales: molecular activity, cellular phenotypes, and effects in cultured organoids. Doing these experiments will produce a data resource of prioritized mutations and iteratively refined mathematical models for prioritizing them as a product. It will also allow us to address a number of questions about cancer.

First, 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, what is the relationship between effects of mutations on different scales? Is it the case that a mutation prioritized to give a strong impact in terms of effect on molecular endophenotypes will also have a strong effect on cellular phenotype, and this in turn will have also a strong effect on organismal phenotypes such as contracting cancer? It is unclear that we'll see a similarity between these three levels, and we will be able to ascertain that here.

Finally, we have great interest in what distinguishes mutations associated with aggressive and nonaggressive cancers. Are they associated with particular types of molecular disruptions or cellular phenotypes?

We will focus our analysis on prostate cancer, a disease that our group has studied extensively in the past. In addition to our general mutation prioritization, we will conduct focused investigations related to TP53 and RB, two genes that are particularly important for this cancer. This combination of multiscale mathematical modeling and validation with biologically focused investigation has great potential to yield insights about prostate cancer and will provide a model for similar study of other cancers.

 **AIM 1 Computational prioritization of coding and non-coding somatic mutations.** In aim 1, we will computationally prioritize the mutations on a number of scales. First, we will look for the mutations that score highly in terms of punitive molecular functional impact. This will be ascertained by features, including whether or not they break motifs, create loss of function of coding genes, preserve protein structures, or hit genes or regulators in the center of networks. We will then take the orthogonal perspective of scoring whether mutations recur and are under positive selection and cancer cohorts. We will further integrate these two scores into a combined prioritization model. With our prioritization scheme, we will conduct focused investigations of the potentially impactful mutations around a core sub-network of genes associated with the TP53 and RB proteins that play major roles in prostate cancer. Finally, from looking at the results of the medium-scale assays in the second aim, we will iteratively update our model over the course of the grant to make it more accurate, with the goal of producing a valuable mathematical prioritization model.

**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 *P*r*O*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 and validation of 10 coding and non-coding variants in prostate organoids.** In the third aim, we will select 120 variants that score highly in terms of effects on molecular activity and gene related prioritization scores for investigation of two potential cellular phenotypes related to cancer: growth and cell invasion, which is related to metastasis. The mutations will be introduced into RWPE-1 prostate normal cells through CRISPR/Cas9 knockin mutagenesis. We then will select the top 10 coding and non-coding mutations and evaluate them in a realistic tissue system – organoids derived from normal prostate samples. Our large-scale phenotyping efforts will be lent expertise by two other large NCI sponsored centers that are headed by core investigators of the grant.  One is the U54 Systems Biology Center at Yale (led by Andre Levchenko), and the other is the Prostate SPORE Center at Cornell (led by Mark Rubin).