Cancer genomics has revealed that there are often millions of mutations per 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. Many tools have been developed to identify drivers and prioritize non-coding mutations. But currently, the relative impact of coding and non-coding variants, especially how to prioritize them in a uniform context, is largely unknown.

In this proposal, we will develop systems-level, 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 (at the nucleotide-level) and recurrence in cancer cohorts. We will then experimentally assay the actual phenotypes produced by these mutations on different scales: molecular activity and cellular phenotypes. Our proposed research will produce as products: a data resource of prioritized cancer mutations and iteratively refined models for prioritizing them.

In addition to the relative impact between coding and non-coding, our proposal will allow us to ascertain the relative impact of mutations on different scales? Is it the case that a mutation prioritized to give a strong impact in terms of its molecular "endo-phenotype" will also have a strong effect on cellular phenotype, and this, in turn, will be associated with the organismal phenotypes of contracting cancer?

We will focus our analysis on prostate cancer, a disease that our group has studied extensively in the past. In addition to our genome-wide mutation prioritization, we will conduct focused investigations related to the sub-networks involving TP53 and RB, two genes that are particularly important for this cancer.

**AIM 1 Computational prioritization of coding and non-coding somatic mutations.** Here, we will computationally prioritize the mutations on a number of scales. First, we will look for the mutations that score highly in terms of putative molecular functional impact. This will be ascertained by features of the genomic sequence, including whether or not they break motifs, create loss of function events, preserve interfaces in protein structures, are associated with allelic activity, or hit genes or regulators that are highly conserved or are in the center of networks. We will then take the orthogonal perspective of finding mutations recur in large cancer cohorts and thus under positive selection to cause cancer in humans. We do this by developing a whole-genome burdening formalism integrating many known genomic covariates of mutational rate (e.g., replication timing). We will further integrate these scores on two-scales into a combined prioritization model. Finally, from looking at the results of the large-scale assays in the second aim, we will iteratively and systematically update our model over the course of the grant to make it more accurate, with the goal of producing a practical prioritization model.

**AIM 2 High-throughput *in vitro* quantification of molecular phenotypes of ~1500 non-coding and ~1000 coding mutations.**Based on the prioritization above, we will select ~200 coding and ~300 non-coding mutations per year and subject them to a number of high-throughput *in vitro* assays in the RWPE-1 cell line (a match to normal prostate tissue) to look at their molecular activity. In total, we will examine ~1000 coding and ~1500 non-coding mutations. We will take advantage of our Clone-seq pipeline to generate these mutant clones on a large-scale. Furthermore, we will assay the non-coding mutations using eSTARR-Seq and Promoter-seq 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, *IN*tegrated *P*r*O*tein *IN*teractome per*T*urbation screening (InPOINT). This pipeline combines five 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 the pool of ~3000?? variants in terms of their strongest molecular activity and then pass this to the next aim.

**AIM 3 Medium-throughput *in vivo* quantification of cellular phenotypes and validation of 10 coding and non-coding variants in prostate organoids.** We will further select 120 high impact variants, from those top-ranked in aim 2, 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 cells through CRISPR/Cas9 knock-in mutagenesis. We then will select the top 10 coding and non-coding mutations and evaluate them in a more realistic tissue context – organoids derived from normal prostate samples. 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 prostate organoids. For coding mutations, we will perform co-IP, protein stability and selected functional assays in gene edited and isogenic control organoids. Our large-scale phenotyping efforts will involve substantial interactions with two large NCI sponsored centers that are headed by co-investigators of this grant, which will allow us to probe mutational impact in progressively more complex contexts: 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).