SIGNIFICANCE

\*\*\* Different Scales

The typical cancer has thousands of somatic variants. These manifest their effects on different scales. At the smallest scale is direct effect molecular activity such as the binding of a transcription factor or the transcription of a downstream gene (often dubbed the "molecular endophenotype" [[get ref]]). Cancer manifests itself on a cellular level in terms of the phenotype of the cells -- e.g., growth or invasiveness, the latter related to metastasis. Finally, it also manifests itself on an overall organismic level in terms of, obviously, the cancer phenotype, but more subtly in terms of the severity of the cancer.

The extent to which variant effects take place at the levels of molecular activity propagates to the cellular phenotype, and organismal presentation is unclear.

\*\*\* Systems & Networks

The coupling between the individual variant, its effect on molecular activity, cellular phenotype, and organismal development, is a systems effect. With many variants and genes often connected in both a regulatory and interaction networks. We endeavor to probe these connections here. In particular, many proteins carry out diverse functions through interacting with other proteins [3]. Recent studies have been conducted on genetic coding mutations in the context of the human interactome network [4-7], where on average, a functionally active protein interacts with >5 other protein partners. We will leverage our experience to deploy a novel approach that systematically uses several agnostic functional assays in parallel. This approach serves as a paradigm to prioritize coding variants and provides important insights into mutation mechanisms of interaction from a systems biology perspective.

\*\*\* Evaluation of Coding and Noncoding Variants

Conceptually, Both coding and noncoding variants may vary in their degree of impact on cancer development or protein formation and function. Numerically, the overwhelming bulk of variants in cancer genomes are non-coding (usually by a factor of 50 to 100).[[ref]] Historically, there has been an emphasis towards studying coding variants due to the functional significance of protein coding regions. However, as noncoding alterations constitute the majority of disease-associated variants [1], further study of non-coding regions may also be critical to a better understanding of cancer biology. Accordingly, we will consider a combination of coding and noncoding variants. Moreover, a wealth of non-coding information is available due to advances in sequencing technologies and efforts by consortia like ENCODE and 1000 Genomes.[[ref]]

\*\*\* Weaker effects

Non-coding variants traditionally have been thought to have weaker effects than coding ones -- not disabling a gene or creating a new binding site in one but more subtly affecting regulation. These may come into play in the development of weaker drivers which may have smaller effects on cancer. There has been recent work on these of late. In particular, recent studies \cite{26456849} \cite{23388632} suggest that certain passenger mutations described as "mini-drivers" may have a weak effect on tumor cell fitness and in turn promote or inhibit tumor growth. From a tumor fitness perspective, three categories can thus emerge: positively-selected driver variants, neutrally-selected passenger variants, and negatively-selected mini-driver variants.

\*\*\* Application to prostate cancer

Prostate cancer is a particularly tractable system for us to focus on for a number of reasons. First of all, as we described, we have much preliminary background working on this specific cancer and deep connections with the cancer SPORE grant.[[add ref]] Also, prostate cancer is highly heterogeneous, displaying very different phenotypes, from a highly indolent, almost notice less disease, to a very aggressive condition. These different presentations may be coupled to systems-wide effects.

Significant efforts have been made to study genetic and environmental causes of this cancer type, but major leaps forward are still needed to develop a more complete etiology of this disease that affects XXX million men worldwide. Along with other major factors associated with prostate cancer such as the hormonal action of androgens and estrogens [8], more than 70 genetic susceptibility variants have been identified [9]. Suspected loci are continuously being discovered using GWAS studies [10] and genotyping arrays [11]. Such variants increase the predictability of the disease and have been associated with altering the expression levels of several genes. Some of the most well-known genes associated with prostate cancer are P53 and RB [[add ref]]. These do [[ $$$ add 3 sentences here !!! ]].

\*\*\* Indolent v Aggressive

One of the most interesting questions about prostate cancer is whether if can detect the overall aggressiveness of the disease from its molecular mutation profile as this has direct implications for treatment.[[ref]]

INNOVATION

Our mathematical model, its multi-tiered cutting-edge biological validation in concert and each individually, and the real-time Bayesian update of the former with the latter are fresh, exciting contributions to the field.

\*\*\* Overall Framework

We believe our overall approach is highly innovative in that we have assembled a diverse team of investigators and are probing prostate cancer on many levels, from clinical outcomes to a more cellular, systems-wide experiments, to large-scale molecular experiments to computational prioritization on a variety of scales.

\*\*\* Aim 1 - Mathematical Model

The specific mathematical model that we are developing is innovative for a number of reasons. First of all, it encompasses a wide range of genomic features. Second, of all, it combines information from both the molecular, nucleotide-level scale (biochemical/biophysical, evolutionary, and network) with information about recurrence and whole-organism disease phenotype. Second, we provide an innovative scheme to update our model in a Bayesian framework using large-scale experimental data. The update and the validation will lead to a more accurate and usable model.

\*\*\* Aim 2 - High-throughput Molecular Experiments

eSTARR-seq: this unique barcoding approach allows direct quantification of enhancer activity, with 40-fold increase in sequencing efficiency compared with traditional STARR-seq

InPOINT: this unique technology directly examines the biochemical consequences of coding variants on protein stability and interactions

\*\*\* Aim 3 - Cellular Assays

CRISPR: This genomic editing breakthrough technology can build a cellular variants impact evaluation model to introduce targeted mutation in coding and noncoding regions from normal prostate cell lines, which will grow in prostate organoid to investigate tumor progression effect.

Organoid technology: This technique, successfully deployed differs from traditional cell culture by maintaining cancer cells in three-dimensional (3D) cultures. Benign and cancer cells that are grown in 3D retain cell-cell and cell-matrix interactions that more closely resemble those of the original tumor compared to cells grown in two dimensions on plastic