**SBU01 discussion**

Jing

Nimubs/LARVA

Comparison with germline in 1KG, PCAWG

Compare selection pressure in cancer vs. normal

 Mutations that create cancer in cohorts (or at least are statistically associated)

 Prioritize important regulators/oncogenes

Shaoke

For the grant, we need to make sure that our prioritization scheme matches experiments

Restrict noncoding variants to TF motifs/protein variants to PPI?

**Test coding and noncoding variants for the same gene(s) already known to be associated with cancer**

**Find the most impactful variants based on molecular assays**

**Then compare in cellular assays to known cancer SNPs in these proteins, SNPs that affect PPI, and SNPs that affect activity of linked enhancers and promoters**

**Validate a few variants in organoids**

Sushant

 Think about heterogeneity/subclonal architecture?

 Find mutations associated with tumor heterogeneity?

 Mutation in organoid + single-cell RNA-Seq?

Gene list for cancer drivers

SNP list for known cancer-inducing mutations in protein-coding genes

TERT promoter/other gold standard for noncoding

Update to parameters (possibly gene and mutation-type specific?)

List tools (find others)

Recurrence + Functional impact

FunSeq

AlleleDB

LARVA

Nimbus

NetSNP

Subject: Beginning of aims for SBU-01 grant, where SBU-01 is one word.

 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 molecular endophenotype produced by these mutations.

 The cellular phenotype and also an organism phenotype on three different scales. 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 to determine 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.

 The mutations that we will look at will be, many of the non-coding mutations are 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'll deal with is 'Are these mutations in any sense comparable or are, fundamentally, the coding mutations more deleterious?'

 Specifically, we have four aims. In the first aim, we will prioritize mutations computationally. We will do this in two ways. 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 amounts of data from such things as the TCGA and PCAWG. We will also prioritize mutations computationally through looking at their sequence level molecular impact. This will be done from using a variety of metrics such as one: 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 these two things, 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 iterative machine learning tactics such as a simple Bayesian update.