Introduction

[[HM, SK WM PDM]] introduction, HM with help from SK, WM, and PDM: 1-2 pages on prostate/colon cancer by 5/28

- Update from SKL: prostate cancer very likely to be considered instead of colon cancer

- General intro to cancer
- Learn about prostate ~~colon~~ cancer biology and pathways
- Looking into tumor growth - while getting only a little into pathway related to metastasis
- Is the impact of mutations at different levels consistent (molecular, cell, etc.)?
- “State-of-the-art” of studying prostate/colon cancer
- Come up with a pseudo-hypothesis

- Colon cancer is a mutated (not a pediatric) cancer
- Vogelstein (JHU): 4 key genes related to colon cancer and certain pathways are clearly related. Other genes are "marginally related"
- Current hypotheses: colon cancer is tractable (experiments are easy to do, cells are easy to acquire, etc. with much understood biology)

[Cancer]

The high complexity of genetic variation associated with cancer demands comprehensive approaches that can assess the effects of different types of variants. In light of the fact that cancer cells tend to have more variants than normal cells, a deeper insight is needed to understand the mechanisms and rate according to which variation evolves in tumors. While we know that these variants usually alter genes that control cell growth and division, we still need to prioritize variants with respect to their deleteriousness, especially because certain variants in tumor cells might merely be a result, rather than a cause, of cancer. Such variant prioritization methods align with initiatives to develop the field of precision medicine, as they would help scientists and medical practitioners understand both common and unique combinations of genetic variants each cancer patient has.

[Somatic vs Germline, Coding vs Noncoding, Rare vs Common (?)]

We consider categorizations of genetic variants underlying cancer at two ~~(three?)~~ levels. The first includes heritable variants passed to offsprings through germ cells, called germline variants, and variants acquired during lifetime, called somatic. A second important level depends on the region in which a variant takes place. Namely, coding variants occur in protein coding regions and noncoding

TOO BASIC / NO GERM

variants, which form the majority of single nucleotide variants in a genome, take place out of these regions. Noticeably, both types of variants can have no to high effect on protein formation and function. A third important distinction between variants leads to their classification into common and rare variants depending on their frequencies in the population.

[More on Coding vs Noncoding + Why noncoding and somatic]

Advances in sequencing technologies and efforts by consortia like ENCODE and 1000 Genomes generated a wealth of annotated data, what engenders the need for comprehensive computational, mathematical, and experimental methods and analyses. Genetic variant types aforementioned occur at different rates across cancers [1], and we believe that the study of all of these types is crucial to further decode the genetic components underlying different cancers. There has been a bias towards studying germline variants that take place in coding regions because of their high importance. However, efforts to study noncoding regions has uncovered their significance as they host the majority of disease related variants [2], what indicates that further study of these regions might be critical to develop a better understanding of genetic cancer variants. Consequently, we will leverage our experience and tools to comprehensively prioritize coding and noncoding somatic variants *in silico*, *in vitro*, and *in vivo* (see Aims 1-4).

[Molecular level, cellular phenotype, and phenotype in cultured organoids]

Effects of numerous genetic variants transcend the molecular level and propagate into the phenotype. However, the extent to which variant effects take place at the level of molecular activity, cellular phenotype, and organismal phenotype are still unclear. The assumption that the impact of variants is consistent at all three levels needs to be examined. For that purpose, we plan to leverage our experience and use a variety of pipelines, cell-based assays, and CRISPR-Cas9-based methods, and realistic organoids (colon or prostate?). We will also study the relationship between different mutations and tumor growth and invasiveness (see Aims 2-4).

[Prostate cancer: intro + state of the art]

In our work, we will focus on prostate cancer. Significant efforts have been made to study genetic and non-genetic causes of this cancer type, but quantum leaps forward still need to be taken to develop a more complete etiology of the disease. Along with other major risks, more than 70 genetic susceptibility variants associated with prostate cancer have been identified [3], and suspected loci are continuously being discovered using GWAS studies [4] and genotyping arrays [5]. Known driver genes include *BRCA1*, *BRCA2*, *HOXB13*, and *RNASEL*. Such variants have shown to increase the predictability of the disease [6] and have been associated with altering the expression levels of several genes including *MSMB, NUDT11*, *RBPMS2*, *NEFM*, and *KLHL33*. These variants have also been

preliminarily associated with multiple lipid metabolism pathways, suggesting potential links between them and the disease [5].

[Prostate cancer: pseudo- hypotheses]

In addition to prioritizing susceptible cancer variants, we will investigate the following related questions: (1) (in light of preface) are non-coding variants as deleterious as coding ones *w.r.t.* prostate cancer incidence?, (2) do deleterious variants lead to the emergence of more deleterious ones in tumor cells?, (3) is there a fitness benefit for heterozygous mutation *v.s.* homozygous in tumor suppressor genes?, and (4) is there a relationship between mutations that lead to loss of heterozygosity in tumor cells?

References

[1] Lu, C. *et al.* (2015) Patterns and functional implications of rare germline variants across 12 cancer types. Nat. Comm., 6, PMCID: PMC4703835.

[2] Zhou, J. & Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods, 12, 931-943, PMCID: PMC4768299.

[3] Eeles, R.A., Olama, A.A., Benlloch, S., Saunders, E.J., Leongamornlert, D.A., Tymrakiewicz, M. *et al.* (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat. Genet., 45, 385–91, PMCID: PMC3832790.

[4] Olama, A.A. *et al.* (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat. Genet., 46(10), 1103-9, PMCID: PMC4383163.

[5] Penny, K.L. (2015). Association of Prostate Cancer Risk Variants with Gene Expression in Normal and Tumor Tissue. Cancer Epidemiol. Biomarkers Prev., ;24(1), 255-60, PMCID: PMC4294966.

[6] Szulkin, R. (2015). Prediction of individual genetic risk to prostate cancer using a polygenic score. Prostate, 75(13), 1467-74, DOI: 10.1002/pros.23037.

Aim 1: Prioritizing coding and non-coding mutations for functional analysis in prostate cancer

In aim 1, we will prioritize both coding and noncoding colon cancer variants for investigation assays of molecular, cellular, and organoid-level phenotypes, simultaneously validating candidate oncogenic variants and refining tools to predict impactful variants. In doing this we will leverage our extensive experience in both variant prioritization and cancer genome analysis. We have developed numerous tools for both coding and noncoding variants, using a variety of approaches.

A. Prior Experience

Experience prioritizing protein-coding variants

We have developed a number of tools that search for deleterious protein-coding variants. Our variant annotation tool is a utility that helps identify variants that overlap genes or other annotations, including, for example, whether variants induce premature stop codons\cite {22743228}. Since minor disruptions to some protein-coding genes can cause disease, while other genes can experience total loss of function with no observable effect, it is important to identify which genes have important functions that could cause disease when altered. Our netSNP tool integrates protein-protein, transcription factor, and metabolic networks to build a classifier that distinguishes genes that essential from those that are loss of function tolerant (Fig 1) \cite{23505346}. Building upon this, we have developed a pipeline for Analysis of Loss of Function Transcripts (ALoFT) that predicts whether loss of one copy or both copies of a given gene is sufficient to cause disease. ~~We have applied these tools as part of the Center for Mendelian Genomics at Yale, and to diseases including Autism and cancer. [[MRS: Include ALOFT? It's not published yet, but maybe soon?)]]~~ Beyond identification of genes whose mutation can cause disease, we have also developed tools that characterize the effects of specific variants. Our STRESS tool identifies mutations that might affect allosteric hotspots in proteins, which can be key to protein function \cite{27066750}. Along similar lines, our Frustration tool uses calculations of localized structural frustration to identify key functional protein regions \cite{27915290}. Finally, our Intensification tool searches for deleterious mutations particularly within repeat regions of proteins \cite{27939289}.

Experience in noncoding genome analysis

Our interest and expertise in prioritizing noncoding DNA variants rests on our experience analyzing a wide array of genomic assays to characterize noncoding genomic elements. Much of this work has been in connection with the ENCODE and modENCODE consortia \cite{22955616, 25164757, 22955619, 21177976}. We have developed widely used tools to identify ChIP-Seq peaks \cite{19122651, MUSIC}, perform RNA-Seq quantification \cite{21134889, 22238592}, identify and functionally categorize new noncoding transcripts \cite{21177971, 25164757}, and to predict enhancer regions \cite{22950945}, including some that have been functionally validated \cite{#58 from ncvarg grant, find PMID}. [[MRS: do we have anything published on enhancer-gene target linkages?]] In addition to identifying and quantifying noncoding noncoding genomic elements, we have been multiple linear and nonlinear models use epigenetic signals to predict gene expression \cite{22955978, insert others}.

Experience in noncoding variant prioritization

We have extensively analyzed patterns of variation in noncoding regions, along with their coding targets[90,95,114]. In recent studies[26,27], we have integrated and extended these methods to develop a prioritization pipeline called FunSeq (Fig 2). It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). FunSeq links each noncoding mutation to target genes, and prioritizes such variants based on scaled network connectivity. It identifies deleterious variants in many noncoding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. Integrating large-scale data

from various resources (including ENCODE and The 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations. Using FunSeq, we identified ~100 noncoding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples[28]. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

Experience in background mutation rate estimation and recurrence analysis

In cancer research, background models for mutation rates have been extensively calibrated in coding regions, leading to the identification of many driver genes, recurrently mutated more than expected. Noncoding regions are also associated with disease; however, background models for them have not been investigated in as much detail. This is partially due to limited noncoding functional annotation. Also, great mutation heterogeneity and potential correlations between neighboring sites give rise to substantial overdispersion in mutation count, resulting in problematic background rate estimation. We developed a new computational framework called LARVA, which integrates variants with a comprehensive set of noncoding functional elements, modeling the mutation counts of the elements with a beta-binomial distribution to handle overdispersion. LARVA, moreover, uses regional genomic features such as replication timing to better estimate local mutation rates and mutational hotspots. We demonstrate LARVA's effectiveness on 760 whole-genome tumor sequences, showing that it identifies well-known noncoding drivers, such as mutations in the TERT promoter. Furthermore, LARVA highlights several novel highly mutated regulatory sites that could potentially be noncoding drivers.

Experience in cancer genome analysis consortia

We have participated extensively in consortium analysis of cancer genomes as part of The Cancer Genome Atlas (TCGA) and the Pan Cancer Analysis of Whole Genomes (PCAWG) groups. Using whole genomes of kidney papillary cancer patients from TCGA, we uncovered noncoding mutations in the MET driver gene potentially associated with this cancer's etiology and also investigated the mutational processes that underlie this cancer's development \cite{28358873, 26536169}. We have also extensively used TCGA RNA-Seq data in the development and application of tools \cite{Loregic, 25884877}. We are currently leading the PCAWG group investigating the impact of so-called passenger mutations on cancer development, progression and prognosis. We are also conducting a study integrating cancer genomes from the PCAWG consortium with ENCODE data to provide a resource for studying noncoding variants in cancer.

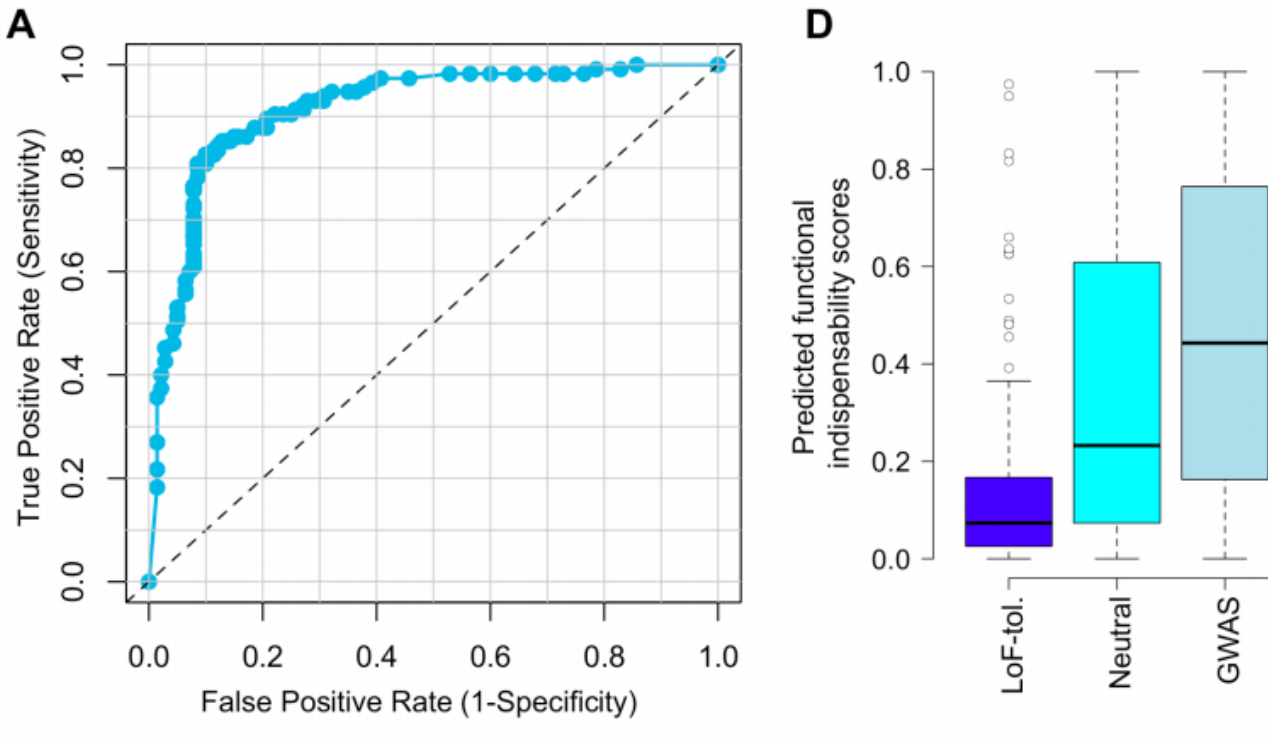


Fig 1. netSNP classifier. A. ROC curve shows distinction of LOF tolerant from Essential genes by netSNP score. B. netSNP score for genes with different levels of functional importance.

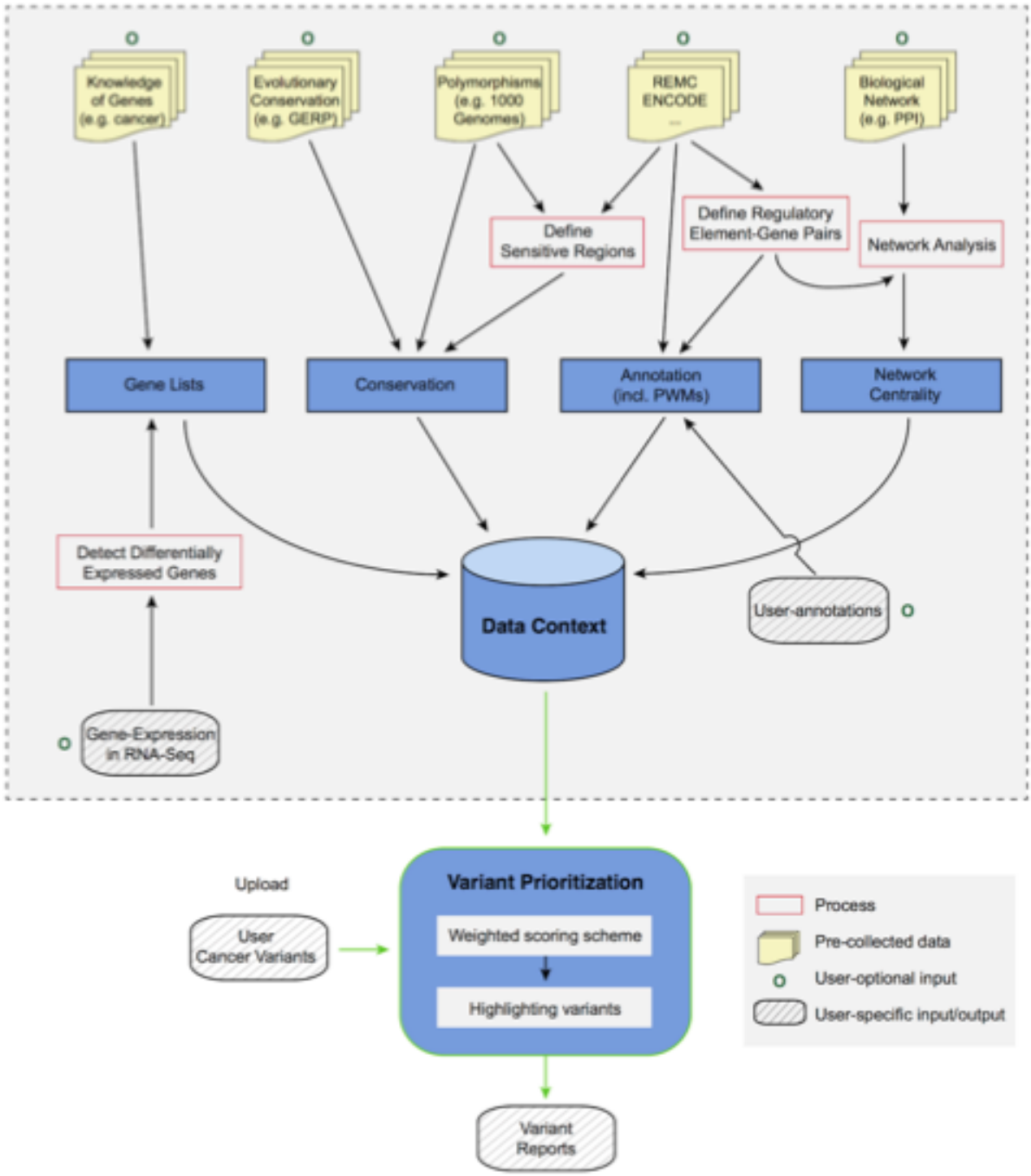


Fig 2. Description of FunSeq workflow and data context

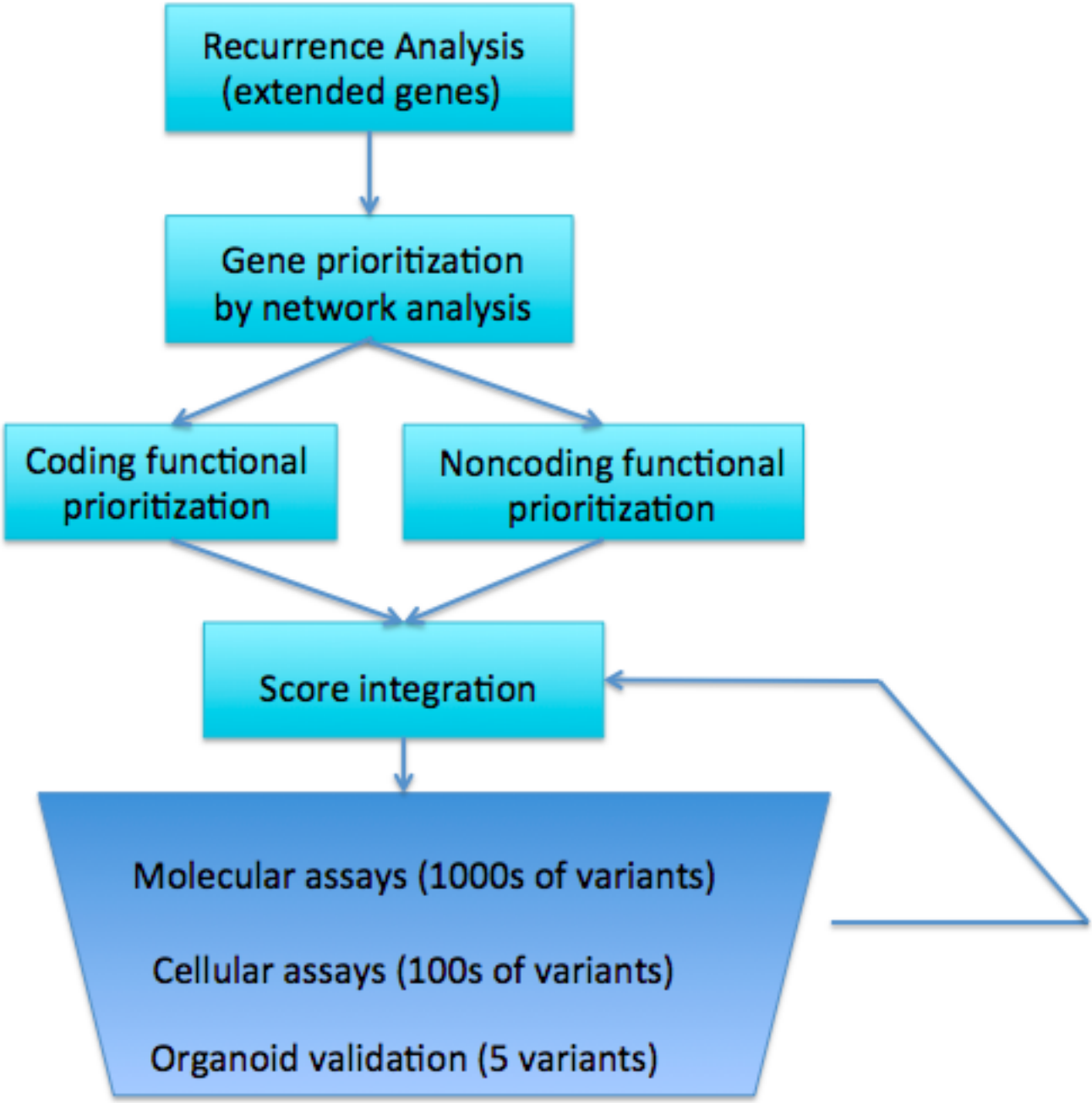


Fig 3. Workflow of prostate cancer variant prioritization

B. Research plan

We will prioritize both coding and noncoding mutations for a set of genes of interest. We will first identify putatively important genes in prostate cancer through a combination of recurrence analysis and biological network analysis. We will then functionally prioritize both coding and noncoding mutations for this set of genes.

Identification of recurrently mutated genes

[[MRS text to replace SKL below?]] To identify genes whose mutation is important to the development of prostate cancer, we will search for genes that are recurrently mutated in prostate cancer patients. Since both coding and noncoding can be cancer drivers, it is desirable to assess recurrence on the basis of both the gene itself and the important noncoding elements, such as enhancers and promoters and protein binding sites in untranslated regions, that are linked to the gene. Specifically for enhancers, we defined a compact list through an ensemble method: enhancer candidate identification by integration of pattern recognition based algorithm on ChIP-seq and DNase-seq signals and STARR-seq pipeline, enhancer-target linkage prediction using JEME method, and then filtered through high resolution Hi-C experiments. The “extended gene neighborhoods” that we construct will then be tested for mutation recurrence using models for background mutation rate built from epigenetic data from prostate cancer specifically. This approach both increases power to identify recurrently mutated genes by considering noncoding mutations, in addition to coding mutations, and also gives an unbiased starting point from which to functionally prioritize both coding and noncoding mutations from genes identified as recurrently mutated.

Identification of key regulators using TF network analysis

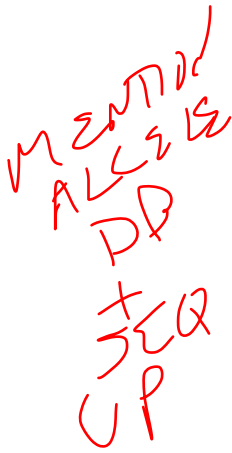
[[SKL: shrink]] We will then investigate the global topology of the transcriptional regulation network by comparing the inbound and outbound edges of each transcription factor (TF). TFs in different levels of the hierarchy reflect the extent to which they directly regulate the expression of other TFs \{cite 25880651}. When comparing the common regulators in approximately matched tumor and normal regulatory networks, rewiring (i.e., target changing) analysis may help to identify cancer-associated deregulation. Our rewiring analysis not only considers direct connections associated with a given TF, but also the whole neighborhood of connections with which a TF associates through membership and topic models, which used a mixed-membership model to look more abstractly at local gene neighborhoods to re-rank the TFs.

[[SKL: shrink]] We then investigated the global topology of transcriptional regulation network by comparing the inbound and outbound edges of each factor. TFs in different levels of the hierarchy reflect the extent to which they directly regulate the expression of other TFs \{cite 25880651}. TFs in different levels of the hierarchy reflect the extent to which they directly regulate the expression of other TFs \{cite 25880651}. When comparing the common regulators in approximately matched tumor and normal regulatory networks, rewiring (i.e., target changing) analysis may help to identify cancer-associated deregulation. Our rewiring analysis not only considers direct connections associated with a given TF, but also the whole neighborhood of connections with which a TF associates through membership and topic models, which used a mixed-membership model to look more abstractly at local gene neighborhoods to re-rank the TFs.

Variant prioritization based on allelic activity

Allele-specific variants(ASVs) potentially provide a most direct readout of the functional impact of a variant. We derive allelic elements by first identifying allelic variants from

individuals will be amassed from The 1000 Genomes Project\cite{23128226}. We will match them with their corresponding RNA-Seq and ChIP-seq experiments from multiple disparate studies, such as gEUVADIS\cite{24037378} and ENCODE\cite{22955616}. After reprocessing and harmonizing the heterogeneous data, we use the beta-binomial test to remove the effect of overdispersion distribution of dataset and detect the ASV in a uniform way. However, because ASVs are enriched for rare variants, we will prioritize by the 'allelic genomic element' with the presence of ASVs. Each element will be assigned an 'allelicity' score based on not only its enrichment of allelic variants within the element (in comparison to accessible variants within the elements and having sufficient coverage to make an allelic activity call), but also across the number of individuals having allelic variants in a consistent allelic direction. The scoring system by element is useful in two ways: (1) it allows continuous ranking of genomic elements based on its allelic impact across multiple individuals (as opposed to defining a threshold to make a binary decision of whether an element is 'allelic') and (2) it enables incorporation of ASE and ASB into the main prioritization scheme; input variants (even those which are rare, but lie in highly-ranked allelic genomic elements) will be up-weighted according to their scores.



Functional prioritization of coding mutations

Once we have identified putative driver genes through a combination of recurrence and biological network analysis, we will score the functional importance of mutations that overlap the coding regions of these genes. We will use our VAT and ALOFT tools to identify mutations that may completely inactivate copies of genes. For potentially impactful variants that do not fully eliminate gene function, we will combine GERP score, a measure of evolutionary conservation, FunSeq2 score, and an ensemble method that combines scores from many tools that score the functional impact of coding variants \cite{PredictSNP, 24453961} [[MRS: do we want to mention structure-based tools?]]

Functional prioritization of noncoding mutations

To integrate the various features, we will expand the weighting system in FunSeq\cite{24092746} and Funseq2\cite{25273974}. Constrained by selective pressure, common variations tend to arise in functionally unimportant regions. Thus, features that are enriched with common polymorphisms are less likely to contribute to the deleteriousness of variants and are weighted less. In general, features can be classified into two classes: discrete (e.g., within or outside of a given functional annotation) and continuous (e.g., the PWM change in 'motif-breaking'). We will weigh these two sets of features with different strategies.

For each discrete feature, we calculate the probability that it overlaps with common polymorphisms. We then calculate the information content to denote the value of discrete features .

The situation is more complex for continuous features, as different feature values have different probabilities of being observed in natural polymorphisms. Thus, one weight cannot suffice for varied feature values. For a continuous feature , which is associated with a value , the probability is first estimated using common variants: . The score of continuous feature is defined as .

The score () is calculated as . We will also incorporate the feature dependency structure when calculating the scores by removing redundant features using feature selection or by performing dimensionality reduction

Parameter tuning after experimental validation

Let () represent the initial feature parameters chosen at random, where is the number of features. will be optimized using an iterative learning scheme by incorporating new experimental information produced in Aims 2. Because of the high throughput of MegaMut and xxx-seq, our strategy is to implement for the first time a iterative learning scheme : the first stage initial learning, the second stage real-time experimental parameter optimization, and the third stage final assessment.

In the first stage, we will randomly select ~500 driver gene as defined by recurrence analysis, PCAWG and TCGA. We will first generate the WT clones of these genes and promoters using xxx-seq. Then, we will select 2 coding variants in coding region and 2 non-coding variants from the promoter region on each gene and generate all ~2,000 variant clones through MegaMut. Their effects on coding and non-coding variants will be quantified by xxx. Starting from the initial tuned , we tune according to the results of ~2000 variants in the first stage. For a specific variant , we define as Bernoulli distributed random variable with indicates that is functional. The expectation of can be predicted through a logistic regression:xxxx (are scaling parameters). To update with experimental validation results , we implement Bayes' rule:. We will use MCMC (Monte Carlo Markov Chain) sampling to search over the parameter space and find the most probable xxx . We will predict the functional impact of all noncoding variants genome-wide, .

In the third stage of final assessment, we will select xxx variants (400 with predicted high impact, 200 with medium impact, and 400 with low impact) on previously cloned driver genes. We will measure their impact on xxxx activities quantitatively through xx-seq.