# **Prioritizing rare variants associated with cancer using non-coding annotation**

[[See GersteinLab convention for editting Google docs at<http://goo.gl/kSrS9> ]]

##

##

## Introduction (Response to Reviewer Criticism of the Previous Submission)

[[MRS (23 Apr) revised -- when pasted into word, 2 lines too long with Arial 12/default spacing.]][[MG(26apr)2ALL: what's font size. Isn't it Helvetica 12? Let's format w/o spc betw para & then have an ident. Single spacing]]

This is a revision of a proposal that we originally submitted about a year and a half ago. The proposal was fairly well reviewed, with the committee expressing “high enthusiasm” for the “likelihood that a useful computational tool for prioritization of noncoding variants will come from the project” and “good potential for being applicable to a variety of traits and diseases.” In the resubmission we attempt to address some critical feedback that our original proposal received, particularly that the “laboratory correlative aspects of [our] proposal are not as well developed as the detailed sophisticated computational aspects.”

In the revision we attempt to address criticism of our approach in a number of ways.

· We have split the original Aim 3, which is the experimental aim into two aims, one focusing on medium throughput experimental validation and the other on detailed characterization.

· In our new Aim 2, we describe how our variant prioritization pipeline can be tuned using medium-throughput validation experiments from new Aim 3 and then assessed using the detailed validation in new Aim 4.

· We have also added a new experimental collaborator, Dr. Haiyuan Yu. Dr. Yu is part of the original FunSeq collaboration between Yale and Cornell and brings significant expertise to the grant in terms of medium-scale enhancer validation. Dr. Yu will implement his newly-established massively-parallel site-directed mutagenesis pipeline, Clone-seq\cite{25502805}, to generate ~1200 specific mutations in ~600 enhancers in three rounds to tune the parameters of the our variant prioritization pipeline and comprehensively evaluate the performance of the pipeline at the end.

The experimental plan was also specifically criticized for not detailing how “cell context can be important” for validation assays and pipeline tuning. Here we address this issue by more explicitly focusing our validations on the prostate cancer system, using a prostate cancer cohort and the LNCaP cell line. This will enable us to show how our variant prioritization scheme can be refined and evaluated for a specific and relevant disease context.

We were also criticized in the original submission regarding the large “computational load for handling all germline variants from TCGA and ICGC whole genome sequences.” We addressed this by collaborating with the PCAWG germline variant calling group (See letter of collaboration form J Korbel, head of PCAWG-8), enabling us to cut in-house variant calling from the proposal. We will also take advantage of recently published studies, in which individual groups called germline variants for large portions of the available TCGA and ICGC whole genome data \cite{25261935,25383969}.

Feedback on our original proposal described our “focus on rare variants [as] a strength as their role has often proven difficult to interpret.” We have preserved this strength of the previous submission; however, as a considerable amount of time has of course passed since the submission and a number of the things that we originally proposed have now been published. In particular, we have published the FunSeq2 paper\cite{25273974}, which is still a paper focused on somatic variants in cancer but which incorporates some of the ideas proposed in the original grant. Consequently, we have moved some of our originally proposed ideas now into preliminary results and elaborated on new ideas focused on allelic variants in our proposal. Overall, our revised proposal provides a strengthened computational framework for rare variant prioritization, with more specifically targeted validation that will enable us to tune and assess our pipeline for prostate cancer as a prototype targeted disease system.

**A. Specific Aims**

[[MG(26apr)2HY+MRS: we need to get the split betw Aim 2 & 3 right]]

[[MG(26apr)2HY+MRS: we need to better integrated aim 4 & elab. functional stuff]]

In this proposal, we will adapt our FunSeq pipeline for prioritizing somatic variants in cancer to create a tool that scores rare germline variants (eleVAR). eleVAR-- **ele**vating germline **VAR**iants--will have the expanded capability to score variants in noncoding DNA and RNA regions uniformly and will contain a flexible weighting scheme, which we will subject to successive rounds of validation and tuning. We will first do this in a generic fashion using publicly available data. Then, we will perform our own validation experiments in prostate cancer cell lines and patient samples to tune and assess a targeted version of eleVAR, as a prototype for how our approach can be targeted to any disease.

**Aim 1.**  Adapt our existing tool for prioritizing somatic variants (FunSeq) to create a generalizable, conceptual approach for prioritizing impactful non-coding variants (eleVAR). Our eleVAR pipeline will build upon the FunSeq approach, which we developed earlier, to prioritize rare germline variants that occur within genomic regions under negative selection within the human population. (a) We will employ this approach and expand the existing set of DNA-level features (ie TF binding sites) to include non-coding RNA features such RNA-binding-protein sites and structured regions. (b) We will further prioritize variants that overlap genomic elements that display strongly allelic activity. (c) We will then use network connectivity from predicted enhancer/promoter-gene linkages, microRNA targeting, and other sources to prioritize variants at hubs and bottlenecks. (d) Finally we will use an entropy-based integrated scoring scheme to combine this diverse set of features into a score for each variant genome-wide.

**Aim 2.**  Implement eleVAR pipeline & develop a workflow for tuning and assessing performance, focusing on prostate cancer as a test case for a specific disease In particular, (a) we will implement eleVAR as a computationally efficient software package, with separate modules for building the data context from annotations, parameter tuning, and scoring variants. (b) We will use eleVAR to generate a generic list of prioritized variants from the PCAWG germline variants, using the framework from Aim 1. (c) We will develop a Bayesian update approach to tune the eleVAR parameters in response to validation data. Then we will carry out tuning (d) first based on publically available validation and then based on our own luciferase reporter experiments in LNCaP cells (see Aim 3).

**Aim 3.**  Medium throughput validation of ~1200 variants using Clone-Seq and luciferase reporter assays. We will perform three rounds of iterative validation and learning of parameter weights to improve eleVAR. **(a)** In each of the first two rounds, we will choose 250 genomic elements, and use our newly-developed massively-parallel Clone-Seq pipeline to generate two variants, one highly prioritized -predicted to be deleterious by eleVAR- and one with a lower score (500 total variants/round). We will assess impact of variants on gene regulation using high-throughput luciferase reporter assays, comparing wild type and predicted high and low impact variants. The luciferase results will be used to fine tune the parameter weights in eleVAR. **(b)** In the final round, purely for assesment, we will choose another 100 genomic elements and two variants with high and low scores on each element and generate clones using Clone-seq. In total, we generate clones for ~600 WT genomic elements and ~1200 variants, which will allow for a comprehensive evaluation of the eleVAR performance.

**Aim 4.**  Detailed experimental validation of top non-coding variants from eleVAR

We will perform a detailed in-depth experimental validation on 6 representative postive variants from eleVAR after medium-scale validation. **(a)** We will use TaqMan assays to genotype our top 6 variants in 4,000 samples from a cohort prostate cancer patients. **(b)** We will further evaluate these 6 variants in cancer samples from the cohort studies for biochemical validation by introducing them into their endogenous loci using the CRISPR-Cas9 system.We will then assay their downstream effects on gene-expression using real-time quantitative PCR analysis, as well as cell viability, migratory potential (for metastasis), and transcription factor binding (ChIP and EMSA).

**B. Significance**

**B-1 Non-coding variants, most of which are regulatory, are significant to the study of diseases but less well studied than coding ones**

Numerous studies have been conducted on the mutations to coding portions of the genome. However, a few initial studies indicate that variants in non-coding regions of genome significantly influence the associated phenotype\cite{17185560} and are often implicated in various diseases\cite{23138309,16728641}. Much of the non-coding variation is contributed by regulatory variants, where cis- and trans-acting variation in the human genome can modulate gene expression\cite{19636342} and this gene expression variation has been implicated in cancer and other diseases\cite{23374354,23348506,23348503,7663520,19165925,18971308}.

**B-2 Rare variants are significant to the study of disease in general & have received less attention than common/GWAS variants**

There have been a large number of GWAS\cite{19474294}, which have primarily focused on associating common genetic variants with diseases. However, growing evidence suggests that rare genetic variants may have strong effects in many human diseases, including cancers\cite{11404818}. Increased disease susceptibility is often attributed to the cumulative effect produced by multiple rare variants\cite{20554195}. For instance, rare germline variants in the CHEK2 gene\cite{16982735} and in the HBOX gene\cite{22236224} were associated with breast cancer and prostate cancer, respectively.

**B-3 Much recent progress in annotating the non-coding genome, making it ripe for variant annotation**

Annotating non-coding regions is essential for investigating genome evolution\cite{16987880}, for understanding important biological functions (including gene regulation and RNA processing)\cite{19148191}, and for elucidating how SNPs and structural variations may influence disease\cite{15549674}. The Encyclopedia of DNA Elements (ENCODE) and the model organism ENCODE (modENCODE) Project provide extensive genomic annotation of human\cite{22955616}, drosophila\cite{21177974} and *C. elegans*\cite{21177976} genomes, respectively. Furthermore, the functional landscape of regulatory variations in the human genome has been investigated by large-scale mRNA and miRNA sequencing\cite{24037378,20220756,20220758,24092820}. Similar efforts have been also directed toward annotating human epigenomic data along with understanding influence of genomic variations on the gene expression profile of the human genome\cite{23715323}. These Expression Quantitative Trait Loci (eQTL) can be further utilized to investigate underlying disease mechanisms\cite{23482391}.

**C. Innovation**

Our method will combine various large-scale genomics data to interpret rare non-coding variants associated with increased cancer risk. Currently no computational pipeline exists with focused analysis for rare germline variants associated with increase cancer risk. Moreover, large-scale consortia, such as The 1000 Genomes Project and ENCODE, have produced data that have been used to interpret other genomic studies. However, these resources have not been fully exploited to understand the functional implications of variants associated with cancer risk. The integration of these data would be an important innovative component of our approach. The specific innovative components of our approach are listed below.

**C-1 Identifying and interpreting rare non-coding variants, consistently for TF binding and ncRNAs, using population-scale polymorphism data**

The GWAS catalog contains many common variants associated with diseases. However, as discussed above, many rare variants may increase susceptibility to various diseases including cancer. Currently, no standard methods exist to functionally interpret such variants, especially in non-coding regions. Futhermore, prior studies for functional interpretation of non-coding variants have primarily focused on regulatory regions associated with transcription factor binding sites or regions of open chromatin. Our approach will also analyze impact of variants in ncRNAs and thus will be one of the first comprehensive approach to decipher the functional interpretation of these variants.

**C-2 Prioritizing elements based on allelic activity**

[[MG(26apr)2JC: we need to define why this innovative]]

Previous studies have prioritized allele-specific variants by taking into account of their regulatory role. However, in the proposed work we will be prioritizing variants based on their presence in different allelic elements of the genome. Allelic elements will be defined by aggregating allelic variants from hundreds of individuals in distinct part of the genome. Each allelic element will be assigned an 'allelicity’ score, which will be dependent on the enrichment as well as recurrence of allelic variants in that element across multiple individuals.

**C-3 Developing a feature weighting system for prioritization & a plan for tuning its parameters by multiple rounds of medium-throughput validation**

We will develop an integrative framework, which will employ an iterative approach to predict ‘high-impact’ rare variants. In the first iteration, we will implement a weighted scoring scheme by assigning weights to various features based on publicly available polymorphism data. Each variant will be assigned a weighted score based on weight of individual features associated with that particular variant. In the second iteration of this workflow, we will apply a Bayesian learning strategy to tune weights based on experimental observations. Subsequently, these updated weights will be assigned to prioritize rare variants.

**C-4 Clone-Seq: a massively-parallel site-directed mutagenesis pipeline leveraging next-generation sequencing**

Current protocols for site-directed mutagenesis require the selection of individual colonies and subsequent sequencing of each colony using Sanger sequencing, which makes them labor intensive, expensive and unscalable for genome-wide surveys. Using Clone-Seq, we can generate clones for ~3,000 mutations in one lane of an Illumina HiSeq run and decrease the cost by more than 10-fold\cite{25502805} (see D-3-a-iv). Please note that Clone-seq is entirely different from previously described random mutagenesis approaches \cite{20711194,\cite{23509263,\cite{23035249,\cite{20947767}. In Clone-Seq, each mutant clone has a separate stock and different clones can therefore be used separately for completely different downstream assays.

**D. Approach**

**D-1 Approach Aim 1 - Convert the prototype FunSeq non-coding somatic variant pipeline to prioritize germline variants and elaborate it with additional features**

**D-1-a Preliminary results for Aim 1**

**D-1-a-i We have experience annotating non-coding regions of the genome, both TF-binding sites and non-coding RNAs**

[[MG(26apr)2LL+ANS: maybe D-1-a-i is too RNA focussed]]

Our proposed work is based on our experience in non-coding annotation. We have made a number of contributions in the analysis of the non-coding genome, as part of our 10-year history with the ENCODE and modENCODE projects. Our TF work includes the development of a method called PeakSeq to define the binding peaks of TFs\cite{19122651}, target identification from profiles (TIP) to identify a TF’s target genes\cite{22039215}, as well as new machine learning techniques\cite{19015141}. Furthermore, we have developed machine-learning methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers\cite{20126643}, which we have partially validated\cite{22950945}. We have also constructed regulatory networks for human and model organisms\cite{22955619}\cite{21430782}, and completed many analyses on them\cite{22125477,21177976,20439753,15145574,14724320,17447836,15372033,19164758,16455753,22955619,22950945,18077332,24092746,23505346,21811232,2160691,21253555}. Furthermore, we have conducted large-scale multi-organism regulatory network comparison along with transcriptome and pseudogene lineage analyses for the modENCODE project\cite{25164757,25164755,25157146,21253555}. Recently, we developed a comparative genomics tool, OrthoClust\cite{25249401}, for simultaneously clustering co-expression networks across multiple species that utilizes the orthology relationships of genes between species.

We also have extensive experience conducting integrated analyses of large sets of RNA-Seq data, such as through the ENCODE, modENCODE, BrainSpan and exRNA consortia\cite{22955616,22955620,21177976,0000001,0000002}. In particular, for general RNA-Seq analysis, we have developed RSEQtools, a computational package that enables splice site identification and gene model creation\cite{21134889}. In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts\cite{22238592}. We have also developed tools that specifically analyze features of ncRNAs, including incRNA, an ncRNA finder, and ncVAR, a prototype pipeline that integrates genetic variants across biotypes and subregions of ncRNAs\cite{21177971, 21596777}.

**D-1-a-iii We have experience in allelic analyses**

A specific class of regulatory variants is one that is related to allele-specific events. These are variants that are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and allele-specific expression (ASE)\cite{20567245,20846943}. We have previously developed a tool, AlleleSeq,\cite{21811232} for the detection of candidate variants associated with ASB and ASE. Using AlleleSeq, we have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and The 1000 Genomes Project\cite{22955620,22955619,24092746}. Overall, we found that these allelic variants are under differential selection from non-allelic ones\cite{22955619,24092746}. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and expression\cite{22955619}. Furthermore, we have implemented the AlleleSeq tool, generated lists of detected allelic variants, and constructed a personal diploid genome and transcriptome of NA12878 on\cite{0000003}.

**D-1-a-iv We have experience in relating annotation to variation & based on this have developed the FunSeq pipeline for somatic variants**

We have extensively analyzed patterns of variation in non-coding regions along with their coding targets\cite{21596777,22950945,22955619}. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations\cite{21596777}. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region\cite{22955616}. Further studies by our group showed relations between selection and protein network structure, e.g. hubs vs periphery\cite{18077332,23505346}.

In recent studies\cite{24092746,25273974}, we have integrated and extended these methods to develop a prioritization pipeline called FunSeq. FunSeq identifies sensitive and ultra-sensitive regions, i.e. those annotations under strong selection pressure as determined by human population variation. It links each non-coding mutation to target genes and prioritizes them based on scaled network connectivity (compute the percentile after ordering centralities of all genes in a particular network). It identifies deleterious variants in many non-coding functional elements, including transcription-factor (TF) binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitivity sites and detects their disruptiveness of TF binding sites (both loss-of and gain-of function events). By contrasting patterns of inherited polymorphisms from 1092 humans with somatic variants from cancer patients, FunSeq allows identification of candidate non-coding driver mutations\cite{24092746}. Our method is able to prioritize the known TERT promoter driver mutations and scores somatic recurrent mutations higher than non-recurrent ones. In this study, we integrated large-scale data from various resources, including ENCODE and The 1000 Genomes Project, with cancer genomics data. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples. We have also submitted a paper applying our method to investigate non-coding mutation patterns in subtypes of gastric cancer \cite{submitted}. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

**D-1-b Research plan for Aim 1**

We plan to convert and extend the current FunSeq prototype from its focus on somatic variants to allow the identification of rare germline variants associated with high functional impact. Our new pipeline is called eleVar. In particular several improvements would be done to eleVar to tailor it for germline analysis, including 1) identifying functional sites among the conserved regions of the human genome and ncRNA regulatory elements; 2) investigating the allelic elements; 3) considering network connectivity. by As a result, we would propose an unified scoring system to consistently prioritize the variants based on these features.

**D-1-b-i Prioritizing non-coding functional elements through human polymorphism data**

In the first aim, we are going to extend the FunSeq approach into eleVAR by defining specific sites on the genome, looking at their polymorphism characteristics (from The 1000 Genomes Project) as well as cross-species evolutionary conservation using classical measures such as GERP score to define rare variants with highly impactful events.

First we are going to update the TF binding non coding elements from the original FunSeq approach. Here we are going to try to use the better enhancer definition coming from the Epigenome roadmap and now from the ENCODE project. Specifically we are going to develop a new machine learning framework that utilizes pattern recognition within the signal of various epigenomic features and transcription of enhancer RNA (eRNA) to predict active enhancers across different tissues. In addition to the enhancers identified in the ENCODE project, we will also include enhancers from the Roadmap Epigenomics project.

Second, RNA regulatory elements will be added as prioritization features in a consistent fashion to the approach for TF-binding sites. Specifically, we will mine RNA interactions with proteins/miRNAs from publically available data, such as CLIP-Seq, CLASH and computationally predicted (TargetScan) to create a compendium of biochemical interactions with RNA\cite{25416797, 24297251, 20371350, 23622248, 21909094}. Our initial analyses indicate that some binding sites are comparable with or even more sensitive to mutation than coding sequences. In addition, we will also incorporate RNA structural elements. our initial survey indicates that more rigid RNA structures, such as stems, are under higher selection pressure than other RNA regions, and that those variants that incur a larger free energy change of the structures tend to be rarer in human populations. We will define sensitive regions based on folding free energy and folding z-score cutoffs that are enriched for rare genetic variants.

**D-1-b-ii Identifying impactful motif breaking & creating mutations**

For impactful events from the TF binding sites we are going to use motif breakers and motif formers to define loss of function and gain of function events respectively as these events are more likely to cause deleterious impact \cite{23512712,24092746,21596777,23348503,23348506,23530248,23887589}. Variants altering the position-weight matrix (PWM) scores for TF binding sites could potentially either decrease (loss of function) or increase (gain of function) the binding strength of TFs. One of the key innovations that we plan to utilize this time is to employ ancestral alleles to get a more accurate determination of these events.

For miRNA/protein bindings sites likewise we are going to use the specific binding site of the microRNA and whether the mutation moves one closer or further from the canonical pattern. We also plan to use specific RNA binding motifs and to look for motif breakers in a consistent way to TF binding sites.

Studies of RNA processing and function have identified key motifs associated with events ranging from RNA splicing to chemical RNA base modifications\cite{18369186}. We have found that intron-exon junctions, polyadenylation sites, and intron lariat structures are much more sensitive to mutation than other genomic regions, particularly for motif-breaking variants. Variants that occur in regulatory motif regions will be scored based on the degree to which they break the motif.

**D-1-b-iii Identifying likely target genes of distal regulatory elements & assessing impact of variants on network connectivity**

To interpret likely functional consequences of non-coding variants, we will define associations comprehensively between many non-coding regulatory elements and target protein-coding genes. The correlation between enhancer and promoter activity across the ENCODE cell-lines and different tissues will be used to identify significant associations between regulatory elements and candidate target genes as done in \cite{20126643}. A single regulatory variant may affect the expression of multiple genes either because it directly regulates multiple genes or because the target gene is itself a regulatory factor.

We will use the regulatory element - target gene pairs to connect the non-coding variants into a variety of networks -- e.g. regulatory network, metabolic pathways, etc. We will examine their network centralities, such as hubs, bottlenecks and hierarchies, as we know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious\cite{23505346,18077332}.For RNA regulatory elements, we will also use protein/miRNA biochemical interactions to interpret the network context of our variants, using RNA molecules as nodes and RNA-protein and miRNA-RNA interactions as edges. We will prioritize variants that are bound by multiple factors, and those within whole RNAs that are bound by many RNA-binding proteins.

**[[MG(26apr)2JC: more detail on geuvadies, na12878, &c]]**

**D-1-b-iv Variant prioritization based on Allelic activity**

The evident regulatory roles of the allele-specific variants assert that they will be useful for identifying functional variants. For example, if we can associate the differential binding effect of a particular transcription factor with different alleles of an SNV, we can identify loci that have potential functional impact in regulation. However, because allelic variants are enriched for rare variants\cite{24037378} that occur only in a few individuals, it will be difficult to match the specific rare variants to prioritze in an individual against those earlier determined to be allelic in a functional genomics experiment on a cellline. Hence, instead of prioritizing by allelic variants, we need to prioritize by 'allelic elements', or allelic regions in the genome.

We derive allelic elements by first identifying allelic variants from hundreds of individuals. These individuals will be amassed from multiple disparate studies with RNA-Seq and ChIP-seq experiments. We will harmonize the heterogeneous data and detect allelic variants in a uniform fashion. The detection pipeline (AlleleSeq) will be also extended to account for the overdispersion of read distributions empirically observed in ChIP-Seq and RNA-Seq datasets\cite{25223782,20671027,22499706} and to remove problematic "outlier" datasets.

Subsequently, allelic variants (rare and common) identified across hundreds of genomes can be aggregated into ‘allelic genomic elements’. Each element will be assigned an ‘allelicity’ score based on not only its enrichment of allelic variants within the element 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) enables incorporation of ASE and ASB into the main prioritization scheme: input variants, even if rare, but lie in highly-ranked allelic genomic elements, will be upweighted according to their scores.

[[MG(apr26)2JC: we need better integ. eQTL, in the paper too.]]

We also plan to integrate another category of regulatory variants, eQTL. These are typically common variants and multiple databases have already been set up\cite{20220756,18462017,17873874,23715323}. Input variants will be upweighted if they overlap with known QTL in these databases.

**D-1-b-v We will use a unified weighted scoring scheme for combining all eleVAR features to prioritize variants**

[[MG(26apr)2YF: how asssess dependency? We should introduce F better]]

To integrate the various features mentioned in Aim 1 to predict ‘high-impact’ variants, we plan to elaborate the weighting system in FunSeq, taking into account the relative importance of each feature \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. “in a particular functional annotation or not”) 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 **d**, we calculate the probability **pd** that it overlaps a common polymorphism. Then we calculated the information content to denote its weighted value **wd**.

$$w\_{d}=1+p\_{d}\*log\_{2}p\_{d}+(1-p\_{d})\*log\_{2}(1-p\_{d})$$

 (1)

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 **c**, which is associated with a score **vc** , we will calculate feature weights for each **vc** . In particular, we discretize at each value and compute **wcvc** . Now, when we come to evaluate the continuous feature **c** for a particular variant, we calculate its weighted value using the corresponding **vc**.

$$w\_{c}^{v\_{c}}=1+p\_{c}^{⪰v\_{c}}\*log\_{2}p\_{c}^{⪰v\_{c}}+(1-p\_{c}^{⪰v\_{c}})\*log\_{2}(1-p\_{c}^{⪰v\_{c}})$$

 (2)

Feature **F** is binary with value 1 if observed. For each variant, we score it by summing up the weighted values of all its features (3). We will also consider the dependency structure of features when calculating the scores.

 $$ S= \sum\_{d}^{}w\_{d}f\_{b}+\sum\_{c}^{}w\_{c}f\_{c}$$

 (3)

**D-2 Approach Aim 2: Implement an efficient eleVAR pipeline & develop a workflow for tuning model parameters and assessing performance**

[[MG(26apr)2SK: rewrote this bit]][[MG(26apr): computation gives millions (genome-scale computational prioritization), medium-throughput on 1000s (high-throughput experimental characterization) , detailed, focused validation on 10s. Think about the characterization on different scales.]]

**D-2-b Workflow for Aim 2[[MG(26apr): need to figure out numbering here]]**

In the previous aim we developed a list of features that were associated with highly impactful variants and a way of weighting these features relative to each other to construct an overall score. Here in this aim we will take our feature weighting scheme and construct a practical piece of software that we can high throughput on many genomic variants. We will then develop a specific large body of genomic variants from the existing cancer genomics data. We will run the pipeline on these variants and prioritize many of them. Then we will compare the prioritization of the variants to publicly accessible validated variants and elements to readjust the parameters in our prioritization scheme. Finally we will compare the newly prioritized variants after this first round with the results of our median throughput experimental validations. We will do this successfully in two rounds updating the parameters at each time.

After comparing our prioritized variants to the publicly available data we will re-update and re-tune our parameters. Finally we will compare our newly prioritized list against the output from the medium-scale validation. This will be done integrally over the course of the grant. At each time we will assess the performance and then update the weights using a Bayesian approach. At the end we will perform an unbiased testing and pick a number of variants for in depth evaluation.

**D-2-b Research plan for Aim 2**

**D-2-b-i Statistical framework for feature integration and parameter tuning**

[[MG(26apr)2YF: need to discuss tuning]]

**D-2-b-i-(2) Parameter tuning using Bayesian update**

The initial feature weights **W** assigned in D-1-b-v will be further optimized with newly available “gold standard” datasets. We plan to tune these parameters using incremental bayesian learning strategy.

The probability that a variant **v**is functional, given the eleVAR score **S** (equation 3 in D-1-b-v), follows logistic function $$P(y=1|W,F)=\frac{1}{1 + exp(-k \* (S-a))}$$

 , with **F** denotes the feature vector of **v**and **W** is the vector of feature weights. To update **W**, we implement Bayes’ rule -$$ P(W|Y,F) ∝ P(Y|W,F)P(W)$$

, the probability of observing **W** given training data **Y** is proportional to the probability of observing **Y** given **W** and **F**, times the prior probability of **W**. Assuming independency between data points in **Y**, $$P(Y|W,F)P(W)=\prod\_{i=1}^{n}P(y\_{i}|w\_{1},w\_{2},...,w\_{m}, f\_{1},f\_{2},...,f\_{m})P(w\_{1},w\_{2},...,w\_{m}) $$

, with $$y\_{i}=1$$

means positive result, whereas $$y\_{i}=0$$

 denotes negative result, $$w\_{m}$$

is the weight of feature $$f\_{m}$$

, given **m** number of total features.

We will maximum this function to find the most probable weights **W**, based on training data, as our updated weights. The updated **W** will then be used as tuned parameters in eleVAR to prioritize variants. The procedure will be iterated in several rounds. In the first round of tuning, feature weights obtained in D-1-b-v will be used to construct priors $$P(W)$$

. In the following rounds, the updated weights will be set as new priors.

**D-2-b-ii Software implementation, taking into account of data context and depenedency graph**

[[MG(26apr)2STL: what do you think of "In particular, they will be able slightly perturb the data context with the addition of a single targeted functional genomics experiment."]]

This software suite will be efficient, robust and yet flexible for users to parameterize and customize for their own research projects. We will host it on a user friendly web server for researchers to query interactively. Researchers are also able to download this software and install on their own computers or deploy it on the cloud. We will provide a downloading version that has been configured in a Docker container to minimize portability issue. We will publish the source code on Github (https://github.com), aiming to distribute the software to the entire research community and ensure the reproducibility of our analysis results.

As our software uses features coming from large scale genomic datasets, calculating score is very time-consuming, space-inefficient and probably computationally intractable for some researchers. To address this problem, first, we will provide precalculated scores for all possible variants in the genome. Also, we will analyse and optimize data flow in our model, aiming to eliminate data dependencies and modularize the calculating process. We will recognize critical interprocedural interfaces (e.g. intersections where multiple flows merges) that are likely to get updated and save intermediate data files to facilitate used in fast rebuilding and recovery. After updating some data sources or partial corruption of runtime data files, our software ,will use a data flow map to identify the flow paths that require rebuilding. All other unperturbed paths will use the nearest intermediate data files and do minimal recalculation. By carefully removing data dependencies and mapping data flow paths, we give users the ability to customize and constantly update our model and software at minimal cost. In particular, they will be able slightly perturb the data context with the addition of a single targeted functional genomics experiment.

**D-2-b-iii Generating an initial list of prioritized variants**

**D-2-b-iii-(1) We will run eleVAR on all the variants & prioritize them**

The PCAWG-8 group will be generating high-quality germline call sets (comprising SNPs, Indels, and SVs) for relatively high-coverage whole-genome datasets. The germline SNP call sets will be generated by the four of the most state-of-the-art variant callers, including GATK HaplotypeCaller\cite{21478889} which will be run by the Broad Institute, and Caveman \cite{21995386} run by Wellcome Trust Sanger Institute. These call sets will then be integrated as priors into the tool FreeBayes\cite{arXiv:1207.3907} [[JC(apr22)2ALL: note that this is NOT a pubmed ID]], which will generate the final call set for further downstream analyses.

[[JC(apr22)2MG: need to know number of prostate genomes & give this a name prostate200?... prostate compendium ]][[MG(26apr)2JZ: need to fill in]]

As we will be focussing on prostate cancer, we will add to this list a number of other whole-genome sequences of prostate cancers (tumor and normal).[[MG(26apr)2ALL: add baca.]] We will call the variants in these genomes consistently to what is done in PCAWG. We estimate that we will have ~200 prostate cancer genomes in total from this group plus what is in PCAWG.

We will run eleVAR on the rare variants resulting from our variant calling on both all of PCAWG and on our set ~200 prostate cancer whole-genome sequences.

**D-2-b-iv First round of parameter tuning based on publically available datasets**

To perform initial round of performance assessment and parameter tuning, we plan to use publicly available datasets from various resources and highly mutable regions in cancer detected by burden test.

[[MG(26apr)2ALL: can we add in ClinVAR?]]

The Human Gene Mutation Database (HGMD)\cite{12754702} collects substantial number of regulatory disease-causing mutations. Rare mutations close to GWAS tag SNPs are probably more deleterious than variants elsewhere. Several high-throughput technologies have been developed to test functional impact of non-coding genomic variants. For example, Kwasnieski et al. used CRE-seq\cite{23129659} to assay > 1,000 single and double nucleotide mutations in promoter regions. Kheradpour et al.\cite{23512712} used MPRA to test variants affecting regulatory motifs in >2,000 human enhancers. We will use these datasets to perform comparison with other variant prioritization method, such as CADD\cite{24487276}, to obtain initial idea of method performance and then tune our parameters.

We will further investigate the germline mutation burden of the cancerous individuals with normal ones. Specifically, we will use the more than 2500 normal samples from 1000 genomes project as the control data, and run the mutation burden test using available software such as SKAT\cite{23684009}. Different from the generally used binning process, which is relative *ad-hoc*, we will aggregate rare mutations in each regulatory element in our updated sensitive feature list to evaluate the cummulative effects of rare variants in cancerous patients. As a results, a list of heavily mutated regulatory elements in cancerous individual but not in controls will reported as candidate regions and would be upweighted during the scoring process.

In addition, since the validation work is done in prostate cancer cell lines, we would further focus on our compendium of prostate cancer WGS (see above) to investigate the germline mutation burden on the noncoding regulatory elements. Both genome wide and regional tests based on known loci will be analyzed to check for suspicious sites.

[[MG(26apr)2JZ: we shud discuss somatic interrelationship as sim to skate burd here for concis. Rework this]]

The interplay between germline and somatic variants may increase the cancer risk but are less investigated in cancer studies. For example, germline and somatic mutations in the promoter regions of some genes have been associated with particular cancers, e.g. telomerase reverse transcriptase (TERT) promoter mutations in cutaneous melanoma\cite{23348503, 22265402,19617566}. In our study, we will also analyze the somatic mutation burden in our feature list and look for overlapping with the germline analysis.Regions that are heavily mutated by both germline and somatic variants should be weighted in eleVAR.

**D-2-b-v Second round of tuning and performance assessment based on medium-throughput experimental results in this project**

We expect an average of ~40K rare germline variants per genome\cite{23128226}. Since they recur rarely at the exact same position, we anticipate a prioritized list of ~8M variants (=40K \* 200 genomes).

In addition to just non-coding variants, we will prioritize non-coding elements of functional importance. Any variants occurring in high-priority functionally important regions (from Aim 1) are more likely to have an impact. [[YF (apr 23): not sure this is how we pick elements ?? I thought we pick elements based on highly ranked variants, instead of directly testing functionally important elements…]][[MG(26apr)2YF: to discuss ]] Thus, overall, we expect this analysis to yield a list of 500 functional regions of appreciable size, which also contain the recurrence of at least two rare variants over the entire sample set of 200 genomes. Thus, assuming ~8M variants are distributed evenly across the human genome, taking also an average element size of 3kb and a recurrence of at least two variants in each of the 500 elements, the number of variants per element will be 4. For each element, we will prioritize at least one of these variants to be of high impact and the other variants to be of a differential impact. Specifically, we will have 1000 variants in total - 500 variants high impact and 500 low impact.

Subsequent tuning and refinement of the eleVAR parameters will be based on further experimental characterization of these 1000 variants (500 highly prioritized and 500 lowly, respectively). We will validate these variants through medium throughput functional genomic screens using the Clone-Seq technology coupled with luciferase reporter assay. Overall, this refinement will be accomplished in two rounds, each round per year, as explained in details at Aim 3. [[HY]] I am under the impression that we will pick 2 variants on each element, one highly and one lowly prioritized. However, here we divide the variants into 3 groups: high, medium, low. I like this three tiered approach, but we need to be consistant[[HY]][[MG(26apr): we need to be consistent]] [[MG(26apr)2ALL: we should just do high & low]]

Finally, during the last year of the grant, we will perform a careful assessment of our model. We will again prioritize our full list of variants and select a final set of 200 top ranked variants for an unbiased validation. This will allow us to construct a precise ROC curve in order to test the accuracy of our predictions.

**D-3 Approach Aim 3 - Medium scale validation of the prioritized variants**

[[MG(26apr)2HY: we an intro sentence]]

**D-3-a Preliminary results related to validation**

**D-3-a-iv Performance, throughput, and cost of our Clone-Seq pipeline**

To set up our Clone-Seq pipeline, we attempted to generate clones for 1034 mutations on 223 genes, including 40 mutations for *MLH1*. We picked 4 colonies for each mutation (4106 total). After sequencing these colonies using one lane of a 1×100 bp Illumina HiSeq run, we were able to identify at least 1 colony containing the intended mutation with no unwanted ones for each allele (100% success rate), including all 40 *MLH1* mutations. Normally 100× sequencing coverage is sufficient for even a conservative variant calling pipeline to identify mutations with high confidence\cite{23201682,\cite{23128226}. The average coverage of these 1034 alleles is > 300×. Therefore, our Clone-Seq pipeline has the capacity to generate > 3,000 mutations in one full lane of a HiSeq run with 1×100 bp reads, drastically improving the throughput and decreasing overall sequencing costs by at least 10-fold\cite{25502805}.

One major advantage of our Clone-Seq pipeline is that it allows us to carefully examine whether other unwanted mutations have been inadvertently introduced during PCR-mutagenesis in comparison with the corresponding wild-type alleles, since we obtain reads spanning the entire gene. This is highly important because there is a ~0.013% error rate in our mutagenesis PCRs, in agreement with previous studies\cite{21906038}. The detection of unwanted mutations, especially those distant from the mutation of interest, is achieved in traditional site-directed mutagenesis pipelines by Sanger sequencing through the gene of interest. This is costly and labor-intensive, especially because multiple sequencing runs and internal primers are needed for one long gene.

Clone-Seq is suitable both for generating a few mutations across many genes as well as a large number of mutations on a few genes. The former situation is applicable when one wants to generate many mutations/variants from large-scale studies (e.g., whole-genome or whole-exome sequencing) since they typically identify mutations/variants on a large number of genes\cite{21798893,\cite{22810696}. The latter situation usually arises in a study focused on a single pathway with a few genes of interest (e.g., an alanine-scanning mutagenesis to determine functional sites on a gene of interest\cite{2471267}). In fact, our Clone-Seq pipeline can generate many more than 40 mutations for a single gene through a two-round barcoding approach: generate groups of 40 mutations and barcode them differently for one HiSeq run. Ten such groups will enable us to generate ~400 mutations for a single gene\cite{25502805}.

In total, we have used the Clone-Seq pipeline to successfully generate 1034 clones with the desired mutant alleles. The results confirm the scalability, accuracy, and throughput of our Clone-Seq pipeline. Through careful considerations, we are confident that this approach can successfully generate the ~1200 non-coding SNVs as proposed.

**D-3-a-v Experience with Reporter luciferase assays confirming validity of in silico TF binding sites**

[[MG(26apr)2HY: we a trans. sentence]]

We have a great amount of experience with developing reporter assays for TF binding. In particular, we have done an earlier study where we validated many binding sites for ER related to prostate cancer. Reporter assays using promoter luciferase constructs encompassing upstream regulatory regions of NEAT1 and corresponding to two ER binding sites are described in Fig. XXX. Interestingly, we discovered that NEAT1 is associated with chromatin and regulates transcription of key prostate cancer genes. Recruitment of NEAT1 was evaluated by ChIP assay and influence on key target genes like PSMA was validated using ChIP and reporter assays (Fig. XXX). Functional validation of NEAT1 functions revealed a predominant tumorigenic role as overexpression of NEAT1 was sufficient to augment proliferation, invasion and migratory behavior of prostate cancer cells (Fig. XXX).

**D-3-b Research plan related to validation**

**D-3-b-i Overview of validation strategy**

Because of the throughput of our Clone-Seq and luciferase reporter assays, we will perform iterative learning and validation in three rounds. In each of the first two rounds, we will select and clone 250 enhancer elements and two variants on each elements that have high and low eleVAR scores, respectively (500 variants total per round). Based on the reporter assay results, we will fine tune the parameters of the learning algorithm as described in **Aim 2**, and then perform the predictions again. In the third round, We will select and clone another 100 enhancer elements and one high scoring and one low scoring variant on each element to confirm the performance of our algorithm. Top candidate ncSNVs that are shown to significantly alter gene expression will be selected for further *in vivo* validations as described in **Aim 4**.

**D-3-b-i-(1) *High-throughput cloning of ~600 WT enhancer elements.*** Forward and reverse sequence-specific primers are combined with attB1 and attB2 sequences, respectively\cite{17207965}. Using human genomic DNA as template, 50 µL PCR reactions are set up on ice in 96-well PCR plates with Phusion polymerase (NEB M0530) according to manufacturer’s manual. We will perform large-scale Gateway BP reactions to clone each PCR product into pDONR223 vector. *E. coli* competent cells are prepared in 96-well plates with 20 µL cells per well. 5 µL of BP reaction products are added to the competent cells using the Tecan robot. After heat shock, 800 µL of SOC recovery medium is added to each well using the Tecan robot and the plate is incubated at 37 °C for 1 hr with vibration. A 20 µL aliquot of the cells is then spotted onto LB + Spectinomycin plates in a fully automated fashion using the Tecan robot. The cells are then spread out in the plates through vigorous shaking with glass beads, as is routinely done in the lab. The plates are incubated overnight at 37 °C. The next day, four colonies per allele are picked for Illumina sequencing. We have already carefully titrated the amount of cells plated so that almost all plates have well-separated single colonies.

**D-3-b-i-(2) *Illumina library preparation and HiSeq sequencing.*** *E. coli* cells for all four colonies of all WT alleles are individually cultured in 96-well deepwell plates over night to the same OD600. 200 µL cells for one colony of each allele are mixed and maxiprepped for DNA plasmids. Four libraries representing one colony of each allele are generated according to Illumina protocols and labeled with distinct barcodes. These four libraries are then mixed into one pool for one 1×100 bp HiSeq run. Correct clones without any unwanted mutations are identified using our customized variant calling software.

**D-3-b-i-(3) *High-throughput cloning of ~1200 mutant enhancers using Clone-Seq.*** Primers for site-directed mutagenesis are designed by our automated web tool:<http://www.yulab.org/Supp/MutPrimer>. 50 µL mutagenesis PCR reactions are set up on ice in 96-well PCR plates using Phusion polymerase. PCR products are digested by *DpnI* (NEB R0176L) overnight at 37 °C (30.5 µL PCR product, 3.5 µL 10× NEBuffer 4, 1 µL *DpnI*). 10 µL of *DpnI*-digested PCR products are added to the competent cells for *E. coli* transformation as described above. The next day, four colonies per allele are picked for Illumina sequencing.

**D-3-b-i-(4) *Functional consequences: Luciferase Reporter assays*.** Reporter assays that employ either luciferase or next generation reporter vectors can provide direct insight to functional relevance of SNPs on target gene. We use a Gateway compatible version of the firefly luciferase reporter vector, pGL4.23-GW (Addgene 60323). All WT and mutant enhancer constructs will be clones into pGL4.23-GW through large-scale Gateway LR reactions. After *E. coli* transformation, individual DNA plasmids for all WT and mutant clones are mini prepped using our fully-automated 96-well miniprep pipeline.

We will use prostate cancer as a model for the validation but we expect that the results will be generalizable to a number of cancers. LNCaP cells will be seeded in 96-well plates and transfected with WT and mutant enhancer constructs. 48 hrs after transfection, enhancer activity will be measured following manufacturer’s instructions (Promega E2940). Assay values will be normalized using internal renilla luciferase as control. Our expectation is that *in vitro* luciferase assays will inform us if a particular mutation had any effect on transcription.

**D-4 Detailed Validation of Specific Variants**

[[MG(26apr)2HY: think about this & ~10]]

In this aim we strive to characterize in detail functionally ~10 of the variants that we find as positive through the median throughput validation. The luciferase assays in Aim 3 are in vitro characterizations. Here, the goal is to really understand what positive means, both in vivo and in terms of disease. We first describe our preliminary results in applying the CRISPR/Cas technology and in screening against a large cohort and then we describe how we will carry this out for the six variants culled. We will choose the six variants to be representative positives from the 1200 tested. These will be variants with both very high eleVAR scores, which also scored positive in screen. We will then use this to obtain an assessment of the median throughput screens.

**D-4-a Preliminary results related to detailed validation**

**[[MG(26apr)2ALL: see how I reworded heads]]**

**[[MG(26apr)2dimple: we need to emphasize more here your expertise in prostate cancer]]**

**[[MG(26apr): we need to shrink cohorts prelim results by half]]**

**D-4-a-i We have experience with prostate cancer cohorts**

Relevant to this proposal our group recently performed a large scale profiling study for 2,000 individuals from the Tyrol Early Prostate Cancer Detection Program \cite{18321314,16829552}cohort. This cohort is part of a population-based prostate cancer-screening program started in 1993 and intended to evaluate the utility of intensive PSA screening in reducing prostate cancer specific death. By genotyping DNA extracted from peripheral blood samples, we annotated the cohort on more than 5,000 CNVs and 900,000 SNPs and then queried inherited low frequency deletions variants \cite{20059347} for their impact in driving prostate cancer \cite{20479773} and the more aggressive form of the disease \cite{10351184}. We reported on coding and non-coding functionally active risk variants. Among the top hits of the case-control study, an intronic variant in the Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C (MGAT4C) demonstrated transcript abundance association with genotype states both in prostate and in lymphoblastoid cells, significant increase in cell and migration upon overexpression in benign and cancer prostate cell lines, and significant decrease in proliferation upon knock down of MGAT4C expression with siRNA. In addition, we suggested that intergenic PCA risk variants affect gene regulation through modified transcription factor binding activity of the Activator Protein 1 (AP-1)\cite{20299548,21862627}. Altogether, we demonstrated that inherited variants may directly or indirectly modulate the transcriptome machinery of known oncogenic pathways in prostate cancer facilitating carcinogenesis.

**[[MG(26apr): we need to shrink "n vitro characterization of SNVs within enhancer elements" prelim results by half. We have to move the cohorts stuff up]]**

**[[MG(26apr): this whole aim goes to aim 3]]**

#### **D-4-a-ii We have experience in the In vitro characterization of SNVs within enhancer elements**

We and others have largely studied the genetics of prostate cancer individuals coupling serum levels and genomics data. Specifically, we characterized the impact of genetic variants relevant to the metabolism of Dihydrotestosterone [131](DHT), the most potent form of androgen, and investigated the incidence of common genomic rearrangements with respect to PSA levels and age at diagnosis [132]. The Tyrol Early Prostate Cancer Detection Program cohort we profiled for those studies is a well-characterized cohort with centralized data collection that ensures proper patients’ follow-up annotations and availability of well-preserved tissues and blood samples. The cohort currently includes more than 3,000 men. Genotype-transcript associations have been reported at large for multiple types of inherited variants [8, 9, 25, 26, 133]; however, experimental evidence of inherited variants allele-specific effect on enhancer activity are lacking. The challenge is to demonstrate how single variants or combinations can increment disease susceptibility or progression by perturbing the expression of a transcript, disrupting the function of a protein or affecting regulatory sequences. In order to study the potential role of inherited genetic variants within regulatory intergenic elements in the context of hormone dependent human tumors, we recently performed an unbiased computational search for AR/ER bound enhancers elements containing SNVs followed by *in vitro* characterization of two selected variants (Garritano et al, Oncotarget REF PMID: 25693204). ChIP-Seq ENCODE data analyses across 17 cell-lines with H3K4m1 and H3K4m3 data availability was performed including 136 transcription factors. Briefly, we applied the following filters to select regions of interest: i) chromatin signature of enhancer activity (H3K4m1, H3K4me1+H3K4me3), ii) binding by ER and AR and iii) presence of a SNP. Using dbsnp137 and AR [134] and/or ER (D Chakravarty, *et al REF to include*) binding data a total of 41 loci was identified (see Table 1 of Garritano et al, Oncotarget 2015). Based upon the minor allele frequencies of the SNPs of interest, two were selected for *in vitro* characterization (on 1q21.3, rs2242193 MAF=0.038 and 13q34, rs9521825 MAF=0.235), here referred to as Locus 1 and Locus 2. Selected loci were cloned in pGL4.26 plasmid (plasmid with alternative allele was also generated) and then validated and characterized *in vitro* by luciferase assay with and without DHT treatment in MCF7 cells. Both constructs reached high responsiveness to DHT treatment hinting at their strong enhancer role. Moreover, the SNP variant on 1q21.3, rs2242193, demonstrated a role in the transcription regulation (p=0.028, Student’s t-test) (**Figure XX).** As both loci were confirmed as transcriptionally responsive to DHT by luciferase assay, we next opted for ChIP assays with AR antibody (or with normal IgG as a control) using MCF7 cells that are heterozygous at rs2242193 in Locus 1, but homozygous for the reference allele at Locus 2. Using quantitative PCR (qPCR), we were able to detect AR binding to both selected loci in MCF7 cells transiently over-expressing AR (**Figure YYB**). Occupancy levels at KLK3, KLK2 and TMPRSS2 were measured as positive controls (**Figure YYA**). Moreover, to assess whether AR showed allele-specific DNA binding at rs2242193, we amplified AR-enriched Locus 1 region by standard PCR followed by double-strand direct DNA sequencing analysis. Quantification of the electropherograms showed that the A allele was significantly enriched (p<10-22, Fisher test) in chromatin fragments immunoprecipitated with antibody against AR compared to input genomic DNA (**Figure YYC**). Altogether, our results show that unbiased genome-wide search for polymorphic regulatory regions (PRRs) is an efficient methodology to discover new functional cis-elements relevant to hormone driven diseases and beyond by providing experimental evidence for selected variants mapping to regulatory regions.

**D-4-a-iii We have experience Modeling mutations in cell lines using CRISPR CAS system**

We have successfully used the CRISPR CAS system to generate mutations and deletions in genes. We have evidenced the presence of somatic mutation in the MAP3K7 gene in castrate resistant prostate cancer patients. In order to determine the functionality of the mutation we have used the CRISPR CAS system to generate the mutation in cell lines. We have successfully introduced cancer-specific MAP3K7 mutation in VCaP cells using the CRISPR-CAS system. Sequencing of cell lines confirmed mutation (Figure XXA). We studied the genomic influence of MAP3K7 mutation in evolution of castrate resistant prostate cancer. Another example is the deletion of the FANCA gene evidenced in 16% of localized prostate adenocarcinomas (11 of 69 cases) and 14% of advanced prostate cancers (4 of 29 cases). In some patients deletion of FANCA was associated with increased cisplatin sensitivity. We have used the CRISPR CAS system to generate FANCA deletion in prostate cancer cell line 22RV1. .Briefly, the CRISPR/Cas9 plasmid (Px459) was obtained from Addgene (Cambridge,MA). Using Ran et al(15) protocol we identified a FANCA CRISPR DNA target sequence using algorithms based on analysis in Hsu et al(16). The corresponding oligonucleotides were ordered (IDT Coralville, IA) and were cloned into Px459 vector. Sanger sequencing confirmed integration of the FANCA target site into the vector. CRISPR deletion of FANCA in 22 RV1 cells lead to increased cisplatin sensitivity (Figure XX2)

**[[MG(26apr): this whole aim goes to aim 3 & emphasize that this stems from a collaboration & HY should add in some thing on the prelim results saying he already made some mutations for the original FunSeq publication in Science ]]**

**D-4-a-iv We experience in functional evaluation of physiologic role of**

**mutation prioritized by the FUNseq pipeline.**

We have done validation for FunSeq prototype pipeline. This is similar to what will be done here but was for a somatic rather germline variant. In particular, a mutation in the RET

promoter predicting a gain of an AP1 motif was determined using the in silico FUN-seq pipeline. Using a luciferase based reporter assay we studied the promoter activity of the WT and mutant RET promoter in the DU145 cell line. Luciferase activity confirmed that the mutant promoter was 1.2-1.3 fold more active than the WT promoter (Fig XX3).

**D-4-b Approach to detailed validation**

**[[MG(26apr)2Dimple: We mention EDRN here. Why not intro in prelim results? We should completely intro Tyrol in prelim results]]**

**D-4-b-i Approach to perform genetic validation in cohorts**

We will determine if any or all 6 variants selected based on successful validation in Aim 3 are associated with cancer or cancer causing characteristics. We will achieve this by studying the specific variant in test cohorts. We will use both the Tyrol cohort (described above) and the Early Detection Research Network (EDRN) \cite{0000005} prostate cancer cohort with thousands of prostate cancer individuals as well as normal controls. The prostate cancer cohort includes men enrolled at three sites as part of the Prostate Cancer Clinical Validation Center that prospectively enrolls individuals at risk for prostate cancer at Beth Israel Deaconess Medical Center (Harvard), at the University of Michigan (Michigan) and at Weill Cornell Medical College (Cornell). Cases are defined as men diagnosed with prostate cancer and controls are men who have undergone prostate needle biopsy without any detectable prostate cancer and no prior history of prostate cancer.

[[MG(26apr)2JC: we should have consistent controls above]]

TaqMan assays for these 6 variants will be performed on 4,000 cases to see if the precise variants recur in a larger cohort. Then, we will follow up for detailed functional screening, to be discussed below. For controls, we will utilize deeply sequenced control cohorts (individuals with no cancer) that are already available, including deeply sequenced trios from the 1000 Genomes Project \cite{0000006}, 500 individuals with Complete Genomics sequencing also from 1000 Genomes \cite{0000007} and non-cancerous individual from the UK10K project \cite{0000008}.Superior allelic discrimination is achieved in these assays as they utilize TaqMan minor groove-binding (MGB) probes. This technique generates a low signal to noise ratio and affords a greater flexibility. The Taqman probes are functionally tested to first ensure assay amplification and optimization for amplification conditions.

Methods: Genomic DNA will be extracted from the blood cellular-EDTA samples in a high-throughput fashion using the QIAamp 96 DNA Blood Kit (Qiagen). All DNAs will be evaluated by NanoDrop spectrophotometer (NanoDrop, Thermo Scientific) and gel electrophoresis (2% agarose). For TaqMan Real-Time Quantitative PCR, each DNA sample will be diluted to 10 ng/ml with nuclease-free water.

**D-4-b-ii Evaluation of molecular consequence of variants**

**D-4-b-ii-(1) Impact on gene expression**: Real-Time Quantitative PCR

Real-time quantitative PCR analysis of the genes downstream of the 6 selected variants will be performed on individuals that have been identified as recurrent for the variants and a similar sized group of non-recurrent individuals. We will look for perturbed gene expression in the target genes. This analysis will inform us if a SNP (in promoter or enhancer regions) has any effect on transcription of the target gene. Recurrent rare SNPs will be further validated by PCR assays using primers that can amplify the genomic region encompassing the SNP. PCR will be followed by direct sequencing of the amplicon using an ABI 3730 DNA Sequence Analyzer on a subset of tumor-normal pairs to verify the individual promoter/enhancer mutations for further confirmation.

**D-4-b-ii-(2) Functional consequences: CRISPR/CAS system**

We will utilize the newly discovered CRISPR/CAS system \cite{0000009} to generate

endogenous mutations in TF binding site in a panel of prostate cancer cell lines (VCaP,

LnCaP, DU145 and PC3). This unique system will provide us an opportunity to

directly modulate endogenous genes and minimize artifacts due to the transfection

based reporter assays. Using CRISPR/CAS mediated genome-engineering method

\cite{23643243} we will directly generate mutations within promoter/enhancers of

target genes. Theoretically we will generate 6 individual SNPs in each cell line and will

study functional relevance of these changes compared to WT. In case of rare

mutations, which occur within both promoter and enhancer regions of the same

gene, we will develop cell lines having these combinatorial mutations.

Mutations within regulatory regions like promoter and enhancer regions might

contribute to one or more biological effects as described in the schematic (Fig. XXX).

The mutant and WT cell lines generated using CRISPR/CAS system will be

monitored for a) phenotypic changes by confocal microscopy and actin staining to

determine effects of mutation on cytoskeletal reorganization b) Influence on

proliferation by MTT and CellTiter-Glo® Luminescent Cell Viability Assay (Promega)

c) Influence on invasive and migratory potential using, matrigel coated invasion and

boyden chambers in 24 well format d) senescence by Bgal staining e) apoptosis by

tunnel assay.

**[[MG(26apr)2Dimple: think we should emphasize this more relative to above - maybe we need to grow this by a para]]**

**D-4-b-ii-(3) Effect of the mutation on TF binding**

In vitro EMSAs will confirm specific binding to WT or mutant sequence by a

particular transcription factor. EMSA (electrophoretic mobility shift assay) is a common technique employed to study protein-DNA interaction. We will use the WT and the MT sequences to determine binding of the transcription factor predicted to be present at the site of mutation.

Chromatin immuno-precipitation (ChIP) assays for TFs overlapping the variant will be conducted to determine if the variant can distort TF binding in vivo. This would help validate the variants that are predicted to be motif breakers. Alternatively for the SNVs predicted to create a new motif, ChIP experiments will help validate binding.

**Timeline$[[MG(6apr)]]$,move to figs$[[MG(6apr)]]$,**

|  |  |
| --- | --- |
| **Year I** | **Aim 1: Development of extended FunSeq pipeline for annotating non-coding variants****Aim 2: Optimization & beginning of variant calling****Aim 3: development of validation assays** |
| **Year II** | **Aim 2: Germline variants called from TCGA/ICGC data****Aim 2: Prioritization of most variants for validation experiments****Aim 3: Begin functional validation experiments** |
| **Year III** | **Aim 2: Finishing prioritization of variants****Aim 3: Functional annotation of prioritized variants****Aim 2: Interpreting validation results in light of prioritization** |