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

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

**Introduction to Reviewer Criticism of the Previous Submission**

[[MG2MRS - intro/ spec aims - 1 page]]

[[MG(6apr)-????: NEW introduction , cut out the relevant quotes from the summary statement and paste them, edit ]] [[MG(14apr): intro: where cut??]] [[MG(14apr): this isn't really rev]][[MRS (23 Apr) revised -- when pasted into word, 2 lines too long with Arial 12/default 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 Aim 2, we describe how our variant prioritization pipeline can be tuned using medium-throughput validation experiments from Aim 3 and then assessed using the detailed validation in 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 medium-scale enhancer validation.

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.

[[MG(14apr)2hy: to add stuff on cell lines ]][[MRS(22apr): I still don’t see HY additions (?)]]

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 immediately 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 the 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 the 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**

[[MG2MRS - intro/ spec aims - 1 page - just 1st 2]][[MRS (23 Apr) edited -- still needs to be shortened some]]

[[?(?)2:?tuning, assessing, validation will be specialized into prostate...]]

[[MRS(16apr): added MG snippet]]

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.

**[[MG(14apr):** rewrite w 3 categories:]][[MRS (23 Apr) done]]

**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 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 to include RNA-level features such binding 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. [[MRS (23 Apr) I thought that we were moving the FunSeq2 scoring scheme into Aim 1, but it looks like we currently still have it in aim 2. Will modify if necessary.]]

[[MRS(16apr): added MG snippet]][[MRS(22 Apr) deleted/rewrote MG snippet]]

[[DC(20apr)2ALL: following sent seems to be incomplete (start of sent seems to have been cut off)]] [[MRS2DC (22 Apr) fragment deleted]]

[[MG(16apr): add in the weighting sys]]

[[DC(20apr)2ALL: following sent seems to be incomplete (ie, the beginning of the sent seems to have been cut off)]][[MRS2DC (22 Apr) fragment deleted]]

[[MRS(16apr): added MG snippet]]

[[MRS(23apr): I worked off of the workflow below from the previous meeting. The current version of aim 2 doesn’t quite match… Will modify aim if current flow is being kept.

**(a) software implementation**

**(b) generate generic prioritized variants**

**(c) tune the parameters using publicly available data**

**(d) tune the parameters in more specific context - our experiments]]**

**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 (a) In Aim 2, we will implement eleVAR as a computationally efficient software package, with separate modules for building 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 then use medium-scale experiments to tune eleVAR parameters using a Bayesian update approach. We will first do this with publicly available enhancer screening experiments, assessing our variant scores using burden tests in cancer genomes. (d) Following this, we will tune prostate cancer-specific parameters for eleVAR using two rounds of luciferase reporter experiments (500 variants per round) in LNCaP cells (see Aim 3). (e) We will assess our final version of eleVAR by conducting reporter assays on 200 additional variants, and through detailed validation of top hits (see Aim 4).

AIM 2$[[MG(6apr)]]What do we do w LARVA ???$cut.?$[[MG(6apr)]]$

[[JZ(12apr)2MG: LARVA cut, only mentioned it in the 2 hits hypothesis part. Put the purple highlighted section to aim2 and removed it from the D sections]]

Get all the variants from PCAWG => JC's numb.

1st round public: enhancer assays, recurrence in prostate in PCAWG+baca

2nd round proj-spec: reweight from internal enhancerseq

[[MG(14apr): rewrite from scratch into new flow]]

[[MG(14apr): combine n tune. Pipeline]][[MG(14apr): , run on vars. workflow,',,]]

[[MG(14apr): Use the lingo Val tuning. Val assess]]

[[MG(14apr)2LS: we need to rework aim3 &4 w/ Dimple & Kellie]]

[[MG(19apr)2ALL+HY: HY will carefully edit this

**Aim 3.**  Medium throughput validation of ~1000 variants using Clone-Seq and luciferase reporter assays for eleVAR parameter tuning

We will perform two rounds of iterative validation and learning of parameter weights to improve eleVAR. **(a)** In each phase, 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.

]]

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

[[MG(14apr): ,bad]] We will perform a detailed in-depth experimental validation on our top 6[[MG(14apr): wrong]]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 using real-time quantitative PCR analysis (RNA-Seq[[MG(14apr): ,wrong]]), as well as cell viability, migratory potential (for metastasis), and transcription factor binding (ChIP and EMSA).

$[[MG(6apr)]]$ update for GTEx in significance $[[MG(6apr)]]$[[MG(14apr): ,no... No road either]]

[[MG2SK - signfic/innovation]]

**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 variants**

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}. However, experimental evidence of inherited variants, allele-specific effect on enhancer/promoter activities and transcriptional influence (short and long range) are lacking.

**B-2 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}.

**B-3 Rare variants are significant to the study of cancer & 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 and can act as a primary driver of 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.

**C. Innovation**

[[MG2SK: do this section C , doing 3 of the 4 innovations]]

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 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.

**[[MG(14apr):** **integ. C2 into C1?]]**

**Analyzing variants in ncRNAs [[?(?)2?: not sure C-4 Analyzing variants in ncRNAs ]] [[SK2MG:done]]**

**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. 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.

 [[SK2MG: How does PCAWG germline variants fits in here ?]]

The 1000 Genomes Consortium has created a deep catalog of genetic variation across many populations. Our approach will use the allele frequencies of variants in ~2,500 individuals from 1000 Genomes data to understand which genomic regions are tolerant to common mutations without conferring disease risk. We will then use this knowledge to identify rare variants that may be associated with increased disease risk.

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

**[[MG(14apr): allelic innovative ]] [[SK2MG:done]]**

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 weights (‘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 & experimental that can be tuned by multiple rounds of medium-throughput validation**

**[[MG(14apr): multi round innovative w tuning ]] [[SK2MG:done]]**

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.

**[[MG2SK : leave alone for HY]][[MG(19apr)2HY: can you shorten by 20% & add in a little about multi-round tuning]]**

**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. In Clone-Seq, we put single colonies of each mutagenesis attempt into one pool (Fig. xxx) and combine multiple pools through multiplexing for one Illumina sequencing run. As described in D-3-a-iv, we can identify correct clones with high accuracy for ~3,000 mutations in one lane of an Illumina HiSeq run and decrease the cost by more than 10-fold\cite{25502805}. Every step of Clone-Seq has been significantly optimized for high-throughput operations. For example, we have to implement customized variant calling software because existing pipelines (e.g., GATK\cite{20644199}) cannot be applied due to our pooling strategy.

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. In random mutagenesis, a pool of sequences containing different mutations for one gene is generated. Therefore, it is not possible to separate one mutant sequence from another and the whole pool can only be used for the same assay(s) together. Furthermore, it is not possible to control which or how many mutations are generated on each DNA sequence, making it difficult to distinguish the functional impact of each individual mutation

**D. Approach**

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

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

[[LL/ANS confer and edit below, condensed to 1 para, de-empaszinng but not removing the modencode ]] [[ANS2MG(23apr): Done!]][[LL2MG(23apr): Portions of the text culled from the 2nd paragraph are highlighted in dark orange]]

**[[MG(10dec)-to-SK add in cmptxn cmpreg cmppgene]]$[[MG(6apr)]]$$[[MG(6apr)]]$**, shrink para to 2 sent$**[[MG(6apr)]]$**

**[[d1-1-a-i cut 2nd to 1 sent. we also add cmptxn cmppgenes]]**

**D-1-a-i We have considerable experience annotating non-coding regulatory regions of the genome & networks**

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.

**[[MG(14apr): put next para into prev… too long]][[LL2MG(23apr): See highlighted text above]]**

**D-1-a-ii** [[DC(20apr)2ALL: a sub-section title is needed here]] [[ANS: how about following?]] **We have extensive experience in RNA-Seq analysis**

**[[LL MG(14apr): ,,shorten by 50%]][[LL2MG(23apr): Done]]$[[MG(6apr)]]$shrink D-1-a-ii We have considerable experience processing RNA-Seq data and annotating ncRNAs by 50%$[[MG(6apr)]]$[[LL: This title looks fine]]**

**$[[MG(6apr)]]$D-1-a-ii We have considerable experience processing RNA-Seq data and annotating ncRNAs [[should shorten by 67% prelim should be integrated into earlier ]]**

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}.

**[[MG2JC: look at]]**

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

[[dec12 JC]]

A specific class of regulatory variants is one that is related to allele-specific events. These are cis-regulatory 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 provided the AlleleSeq tool, lists of detected allelic variants, and the constructed personal diploid genome and transcriptome of NA12878 on\cite{0000003}.

**D-1-a-iv We have extensive experience in relating annotation to variation & based on this experience have developed the prototype 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 also applied our method to investigate non-coding mutation patterns in subtypes of gastric cancer**[[MG(14apr): ,submitted]]** Currently we are playing a leading role in TCGA/ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

########### comments secrion ###########

**[[MG2JZ/ANS ]]**

**[[MG(14apr): sites -- JZ/ANS SITES - CONSISTENTLY, G0F/LOF & POLY, use another layer of heading, motif discussion might be different from the sites...**

**MAYBE**

**D-1-b-i conserved**

**D-1-b-ii motifs ]][[ANS&JZ:Done!]]**

[[DC(20apr)2MG,JZ,ANS: the above sub-heading scheme (i.e., “D-1-b-i conserved” and “D-1-b-ii motifs”) has not been implemented because these subsection titles seem to be only tentative]]

**[[MG: in D1b preamble we should say that elevar will treat ncRNAs and tfbs on the same level]]**xxxxxxx

########### comments section ###########

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

**[[JZ(22apr)2MG&ANS:moved this paragraph here for an introduction of aim1]] [[ANS2MG&JZ (23 apr): I edited a bit to emphasize ncRNA sites]]**

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, which 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 Identifying functional elements through human polymorphism data**

**[[JZ(22apr)2MG&ANS:combined all the sites info into this section]]**

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) [[JZ(22apr)2MG: try to avoid conserved, its so confusing]] as well as cross-species evolutionary conservation using classical measures such as GERP score to define rare variants with highly impactful events.

**[[JZ(22apr)2MG&ANS:should we separate the TF and enhancer? What do you think ANS?]][[ANS(23Apr)2MG&JZ: I think this is okay - much better after enhancer moved up and not a separate section anymore]]**

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 into the features. 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 inital survey indicate 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 features associated with functional motifs**

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 the 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 similar 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.

[[ANS2JZ: This needs to be put in the right place in the RNA section or here along with motif discussion - please decide!]][[JZ(22apr)2ANS: lets put it here]][[ANS2JZ: I agree. Looks better with the enhancer discussion moved]]

[[ANS Apr 23 - can be removed now as merged with motif section]][[JZ2ANS: removed]]

[[JZ/ANS - fix up the network, maybe D-1-b-iii]]

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

[[JZ(12apr)2MG: why enhancer targets appear twice? in appear also in the D-1-b-i section. Should we just put it at just one place?]]

\*\* 1 we better connect enhancers to targets

We use metatracks approach and elaborate w work encode

\*\* 2 we connect rna sites into networks

we look at clipseq & miRNA making networks and then prioritizing hubs

[[ANS2MG (12apr) - I moved/edited the target prediction to this paragraph.]]

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}**[[MG(14apr): ,as,done in((ref))]] [[ANS 23 Apr:done]]** 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.

**[[MG2JC elaborate the alleleic stuff to 1 more para]] [[MG(14apr): ,more how… elaborate]] [[MG(14apr): ,element]] [[MG D-1-b-iv Variant prioritization based on Allelic activity & eQTL association (AlleleDB module) - JC more logic + shorten eqtl]]**

**D-1-b-iv Variant prioritization based on Allelic activity & eQTL association (AlleleDB module)**

[[dec12 JC]]

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 justify the impact of these rare allelic variants in a prioritization scheme which requires substantial supporting evidence. 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 DNA sequencing, 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}.

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.

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

[[DC(21apr)2ALL: following sent seems to be incomplete (start of sent seems to have been cut off)]]

In this aim, we will provide an efficient implementation of eleVAR, including the development of a weighting and tuning system to bring together all its features, **$[[MG(6apr)]]$,elaborate weighting system in FunSeq!$[[MG(6apr)]]$** and prioritizing all the rare germline variants in sequenced tumor genomes. [[DC(20apr)2?: Finish sentence here]]. Overall, using eleVAR functional prioritization plus screening out the common variants will allow us to identify the rare variants on a haplotype block with the greatest impact. We cautiously note that unlike GWAS, which look for association signal, our method prioritizes variants based on functional information. Thus, the variants identified by our pipeline are highly likely the causal variants.

**D-2-a Preliminary results (none)**

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

**[[MG2YF: fix WEIGHTING system ... will talk about it in the future ]][[MG(14apr):** **p of w ? Or use w as prior ? Inf sign in bayes ? .expl logos.mention elevar score earlier]]**

**[[MG: YF rewrite , fix D-2-b-i Math for gluing together features]]**

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

**D-2-b-i-(1) We will use a unified weighted scoring scheme for combining all eleVAR modules to consistently prioritize variants**

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

The initial feature weights **W** assigned in D-2-b-i-(1) 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-2-b-i-(1)), 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-2-b-i-(1) 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 ??** [[DC(20apr)2ALL: a subsection title is needed here]]

**[[MG2STL]]**

**[[MG2STL: try your best, dependency is interesting... ]]**

**[[MG(14apr): ,separating the rebuild of data context from the lookup of a variant and putting as much as possible into this core score file -- earlier version of the grant + ]]**

In the above, we have developed features and a mathematical system for putting them together. Here we will describe how we will build a high throughput software implementation that integrates all the functionalities we have described above. 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.

**[[STL: Data flow analysis, ideas from Software Engineering and Compiler Optimization]]**

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 uses 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.

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

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

**[[MG JC to edit and check and add refs]][[MG(14apr): ,checked?????]]**

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

**D-2-b-iii-1 Data source/provenance - use PCAWG + TCGA (sander) [[mg(6apr): flesh out]]**

**D-2-b-iii-(2) Get many variants from calling on PCAWG & then run elevar**

[[MG We want to cut this out & say that we’re using the germline calls from PCAWG8]]

The PCAWG-8 group will be generating high-quality germline call sets (comprising SNPs, Indels, and SVs) for relatively high-coverage whole-genome datasets. [[JC(apr22)2MG: need to verify this]] 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.**[[MG(14apr): more how merge callers -- where the PCAWG variants come from]]**

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

**[[MG for JZ/YF D-2-b-iv First round of tuning based on publically available datasets]]**

[[YF/JZ]]

**CRE-Seq [[dc(22apr)2all: insert a new sub-section number here, or leave this as part of the previous section?]]**

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.

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.

[[JZ]][[JZ(22apr)2MG: done! but a little bit confused. do u want to use these burden test as a validation or use them in the weighting scheme]]

1) rare variant burden in cancer v normal - SKAT of PCAWG v 1000G

2 Were going to be focussing prostate context later on so we'll also try for prostate

**3) introduce 2hit: deleterious germ var might generate the somatic varaints; upweight the 2-hit burden**

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 more than 100 prostate cancer samples collected from collaborator [[JZ(22apr)2MG: obviously should not use collaborator since MarkR is also in the grant, how should we call here?]] 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.

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 overlappings with the germline analysis. Regions that are heavily mutated by both germline and somatic variants should be weighted in eleVAR.

**[[JC/YF try to edit this]]**

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

**[[MG(19apr)2HY: please look over this and edit + send comments to YF(Yao) & MG ]]**

2) Element selection [[MG2:JC for prostate cancer with 100 variants ]]//

**We'll take just the variants in the PCAWG that are in prostate genomes + we'll add a few more porstate genoem such Baca et all... we est that we'll have ~100**

[[JC(apr22)2MG: need to know number of prostate genomes ...]]

We will run eleVAR on the rare variants resulting from our variant calling on ~200 TCGA/ICGC prostate cancer whole-genome sequences. 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).

We note that unlike GWAS, which look for association signal, our method prioritizes variants based on functional information. Thus, the variants identified by our pipeline are likely to be the causal ones.

In addition to just non-coding variants, we will prioritize non-coding elements of functional importance. ...**[[MG(14apr): ,.how var to elt]]**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…]]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.**[[MG(14apr): ,',recurr elsewhere ]]** 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 at high impact, 250 of medium impact and 250 at low impact.

Subsequent tuning and refinement of the eleVAR parameters will be based on further experimental characterization of these 1000 variants (500, 250 and 250, highly medium and lowly prioritized 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.

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. As further described in Aim 4, for a vigorous in-depth experimental evaluation, we will select 6 highly prioritized variants and subject them to Taqman in large cohort, as well as genetically engineer them using the very promising CRISPR-Cas9 technology in order to measure each variant’s effect. This will enable us not only to characterize the accuracy of our computational algorithm by the medium scale pipeline, but also provide an in-depth analysis of a highly prioritized variant’s biological effect.

**[[MG(12apr): pasted in text from Dimple on 8 Apr, starting below. It has a greatly modified Aim 4 & a slightly modified Aim 3. LOCK starts only HY edits aim3]]**

**,,,,blue text.....**

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

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

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

,,,,topic sent....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 Reporter luciferase assays confirm validity of in silico TF binding sites**

,,,,trans sent....Using an *in silico* approach we determined genome-wide distribution of ER in prostate cancer. Intriguingly, we observed a robust recruitment to non-coding genome and identified several intergenic sites that correlated with high ER occupancy. Analysis of recruitment vs transcript profiles confirmed that ER recruitment was associated with productive transcription of long non-coding RNA. Recruitment of ER upstream of NEAT1 lncRNA was addressed in greater detail. Reporter assays using promoter luciferase constructs encompassing upstream regulatory regions of NEAT1 and corresponding to two ER binding sites are described in Fig. 9. 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. 10). 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. 11).

**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. That is, we will first clone and test ~150 candidate ncSNVs predicted by our computational learning algorithm. 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. We will then clone and test another ~150 ncSNVs 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-a Preliminary results related to validation:**

**D-4-a-i Low-frequency functionally active intronic & intergenic inherited variants predisposing to cancer**

Emerging insights into the genetics of constitutional disease etiology demonstrate that germline polymorphisms are associated with a variety of diseases including Alzheimer’s, Parkinson’s, mental retardation, autism, schizophrenia\cite{19715442}and cancer\cite{19536264,18685109}. 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)* demonstratedtranscript 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.

**D-4-a-ii In vitro characterization of SNPs within enhancer elements bound by AR and/or ER**

The Tyrol Early Prostate Cancer Detection Program cohort 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. As part of our Trento-Innsbruck-Cornell collaboration, we further studied the genetics of prostate cancer individuals coupling serum levels and genomics data. Specifically, we studied the impact of genetic variants relevant to the metabolism of Dihydrotestosterone\cite{20056642}(DHT), the most potent form of androgen, and investigated the incidence of common genomic rearrangements with respect to PSA levels and age at diagnosis\cite{23381693}.

It has been shown that a significant fraction (26%-35%) of inter-individual differences in transcription factor binding regions coincides with genetic variation loci and that about 5% of transcripts levels are associated with inherited variant states\cite{20299548}. Genotype-transcript associations have been reported at large for multiple types of inherited variants\cite{21479260,20220756,20220758,21862627,17289997}, however experimental evidence of inherited variants allele-specific effect on enhancer activity are lacking. In order to study the potential role of inherited genetic variants within regulatory elements in the context of hormone dependent human, we have performed an unbiased computational search for AR/ER bound enhancers elements containing SNPs followed by *in vitro* characterization of selected variants. **Table 1** shows counts of SNPs from the dbsnp137 set within AR\cite{20478527} and/or ER (Chakravarty D, *submitted*) binding sites that intersect peak ENCODE data\cite{22955616} generated from 20 cell-lines and ChIP-seq experiments for H3K4m1, H3K4me1+H3K4me3, H3K9ac, H3K27ac, Dnase-seq and FAIRE-seq. For each marker the consensus was generated as the merge of all the regions that are present in at least 2 cell lines and comply with a set of filters. **Fig. 8** shows examples of AR-responsiveness and SNPs impact on putative enhancer elements in MCF7 cells (Garritano S, Demichelis F, *unpublished*).

**D-4-a-iii Modeling mutations in cell lines using CRISPR-Cas9 system**:

Mutation in the MAP3K7 gene is seen in castrate resistant prostate cancer patients. Inorder to determine the functionality of the mutation we have used the CRISPR-Cas9 system to generate the mutation in cell lines. We have successfully introduced cancer-specific MAP3K7 mutation in VCaP cells using the CRISPR-Cas9 system. Sequencing of cell lines confirmed mutation. Next 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 metastatic prostate cancer patients. We have used the CRISPR-Cas9 system to generate FANCA deletion in prostate cancer cell lines..

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.

**D-4-a-iv Validation and functional evaluation of physiologic role of somatic mutation predicted by FUN-seq bioinformatics pipeline**. Mutation in RET promoter was determined using insilico FUN-seq pipeline. Bioinformatic analysis using FUN-seq pipeline predicted gain of AP1 motif in promoter of RET promoter. Using a luciferase based reporter assay we studied the promoter activity of WT and mutant RET promoter in LnCaP and DU145 cell lines. Luciferase activity confirmed that mutant promoter was X fold active than the WT promoter. Further addition of AP1 inhibitor compromised the activity, indicating that the observed increase in promoter activity was indeed due to AP1 binding at promoter elements.

**Approach:**

**D-4-b-i Targeted genotyping:** We will determine if any or all 10 variants selected based on successful validation in Aim 3are associated with cancer or cancer causing characteristics. We will achieve this by studying the specific variant in test cohort. 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 include men enrolled at three sites as part of the Prostate Cancer Clinical Validation Center that prospectively enroll 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.

We will first take the highest prioritized variants then subject them to validation. Overall we plan to start the validation pipeline with the top ~10 elements identified from the reporter assays(as described above). TaqMan assays for these 10 variants will be performed on 4,000 cases to see if the precise variants recur in a larger cohort. From this group, we will select top third of the variants (~6), based on recurrence, that we will follow up for detailed functional screening, to be discussed below. This functional screening will be through various reporter assays (e.g. luciferase) looking for the effect on the target gene and also from using the CRISPR-Cas9 system. 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 are 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 functional consequence of variants**

Based on the Taqman results, we will pick the top ~6 variants for functional follow up.

**D-4-b-ii-(1) Functional consequences: RNA-Seq**

We have RNA sequencing data for 85% of the individuals enrolled in the cohort. To fill out the dataset, RNA sequencing will be completed on the remaining where we see recurrent variants (on up to ~160 individuals). The RNA-Seq will be done according to the protocols in\cite{21036922}. This analysis will inform us if a SNP (in promoter or enhancer regions) has any effect on transcription of target gene. This analysis will provide a comprehensive list of SNPs that might correlate with loss or gain of expression. 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 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-Cas9 system**

We will utilize the newly discovered CRISPR-Cas9 system\cite{0000009} to generate endogenous mutations in target genes 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-Cas9 mediated genome-engineering method\cite{23643243} we will directly generate mutations within promoter/enhancers of target genes. Theoretically we 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. 12). In addition to loss or gain of cognate coding transcript, it is quite conceivable that the SNPs might alter expression of non-coding transcript. To capture the complete influence of rare nominated SNPs at genomic and transcriptomic level we will perform RNA seq. The schematic (Fig. 12) shows representative iterations of plausible genomic changes that will be captured in this validation.

For modeling mutations in non coding RNA, prostate cell lines will be screened for the expression of the non coding RNA, and in cells having a high endogenous expression of the ncRNA, CRISPR-Cas9 system will be used to generate the mutation.

**D-4-b-ii-(3) Functional consequences:**

The mutant and WT cell lines generated using CRISPR-Cas9 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.

**D-4-b-ii-(4) Functional validation of mutation in non coding RNA:**

Total RNA will be extracted from cell lines expressing the wild type and the mutant ncRNA and RNA sequencing will be performed to determine the mutation specific gene signature.

**D-4-b-ii-(5) 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 to a 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. 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.

**[[MG(12apr): end of lock & Dimple paste]]**

**Timeline$[[MG(6apr)]]$,my 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** |