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

**[[See GersteinLab convention for editting Google docs at** [**http://goo.gl/kSrS9**](http://goo.gl/kSrS9) **]]**

**[[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]]**

**Introduction Responding to Criticism for NCVARG Grant**

This is a revision of a proposal that we originally submitted about a year and a half ago. The proposal was fairly well reviewed and in the resubmission we are trying to address the main criticisms. The main criticism of the original proposal was that the experimental plan was not completely fleshed out. In the revision we have tried to address this in a number of ways.

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

· In aim 2 we talk about the overall computational plan and how we intend to use medium-throughput validation for parameter tuning and then to use the detailed validation in aim four for assessment.

· 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.

**[[MG(14apr): hy to add stuff on cell lines ]]**

We were also criticized in the original submission for being overly ambitious in planning to call the TCGA variants. We immediately addressed this in the resubmission. We have cut out all of the variant calling and now are planning to rely on the germline variants called by the PCAWG group. (See letter of colloration form J Korbel (head of PCAWG-8).) Also trivially we can get the germline variant from whole genome sequences from a number of recent published papers such as those from the Chris Sander and Erik Larsson groups \cite{25261935,25383969}.

The original proposal was praised for its computational framework and unique focus on rare variants. 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.

**Specific Aims**

**Completed comments**

$**[[MG(6apr)]]**$ & epigenome road map$**[[MG(6apr)]]**$ **[[MRS(12apr) - completed] (e.g. GTEx$[[MG(6apr)]]**$ **[[MRS(12apr) - completed]GEUVADIS RNA-seq data)..**

**$[[MG(6apr)]]**$pcawg$**[[MG(6apr)]]**$**[[MRS(12apr) done]]**

 $**[[MG(6apr)]]**$funseq ==> ELEVAR What this grant is about is updating funseq to create elevar for germline$**[[MG(6apr)]]**$

Prioritizing noncoding variants is a subject ripe for exploration with increasing depth of noncoding functional annotations from the ENCODE and epigenome roadmap projects \cite{22955616,25693563} as well as the many new population-scale functional genomics datasets (e.g. GTEx, GEUVADIS RNA-seq data) \cite{23715323, 24037378}. Most variant prioritization thus far has focused on GWAS SNPs. Here we focus on rare variants, often not strongly correlated with other variants, which may have stronger effects than GWAS SNPs. In particular, we look at rare, germline SNVs (and some deletions and insertions) associated with cancer, trying to prioritize the non-coding variants most associated with disease. This work will be carried out by a team comprising labs specializing in computational biology (Gerstein), experimental cancer genomicis (Rubin), and high throughput experimentation (Yu) that who have worked together for the past decade.

**[[JZ(12apr)toMG: i feel feature section in aim1 is a little bit messy. Network analysis was mentioned at 1 and 3?]][[MG(14apr): no net in aim3]]**

**Aim 1.** Adapt our existing approach for prioritizing somatic variants (FunSeq) to create eleVAR for “**ele**vating” rare germline **VAR**iants that will uniformly score variants for noncoding DNA and RNA. In aim 1, we will develop the features. We will develop 3 categories of features : sites, connectivity and allelic…..

**[[MG(14apr):** rewrite w 3 categories]]

The existing FunSeq pipeline defines the notion of a mutationally "sensitive" region based on population-genetic analysis\cite{FunSeq, FunSeq2}. In developing eleVAR: **(a)** We will expand our analysis of sensitive regions to include RNA regulatory regions, such as regions of miRNA or protein binding. **(b)** **[[MRS(12apr): cut? isn’t this already in funseq2?]]** We will refine our analyses of mutations affecting binding sites, considering gain of function in addition to loss of function. **(c)** We will generate improved enhancer predictions to allow better targeting of noncoding DNA mutations to associated genes. **(d)** We will create a combined network of promoter/enhancer-gene, protein-DNA, protein-RNA, and miRNA-RNA interactions to enable network-based prioritization of variants. **(e)** We further will prioritize the above features based on their allelic activity, how sensitive their activity is to sequence differences, between maternal and paternal alleles.

**[[MG(14apr): more parallel to sect ....gen sen net alllele8]]**.

$**[[MG(6apr)]]**What do we do w larva ???$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]][[MRS: deleted puprle highlighted section from aim 2 -- too long!]]** “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}. Similarly, many somatic and germline mutations in the T53 gene and GALNT12 coding exons were implicated in Sonic-Hedgehog medulloblastoma (SHH-MB) tumors \cite{22265402} and colon cancers \cite{19617566}”

**[[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]]**

**Aim 2.** Integrate eleVAR features with a tuneable weighting scheme, implement into a practical pipeline & run on many variarnats. Take results and use to tune parameters. Finally assess.

Implement efficient eleVAR pipeline on TCGA/ICGC germline variants, create framework for tuning pipeline to experimental validations. The Pan Cancer Analysis of Whole Cancer Genomes is producing variant calls for TCGA and ICGC whole genome sequences (estimated >2000 during the grant), and other groups have already produced their own calls on subsets of these data \cite{PCAWG, 25261935,25383969} **(a)** We will implement eleVAR on germline variants from the PCAWG/TCGA/ICGC cancer genomes, **[[MG(14apr): ,,weights n tune is aim 2b]]**.using the entropy-based scoring scheme from FunSeq2 to combine all the features discussed in Aim 1. **(b)** We will include a parameter tuning module in eleVAR, that will perform a Bayesian update of the weights for all the scoring features, using publically available data, variant burden tests, and medium-scale validation experiments (see Aim 3). **(c) [[MRS(12Apr) cut this? we have very little space in spec aims]]** We will develop eleVAR as a practical and efficient computational package to perform large-scale computations, including separate modules for building the complex data context (the annotation from many sources), parameter tuning, and actual scoring of genomic variant sets. **(b)** After each round of validation, we will tune the feature weights of the eleVAR pipeline to and produce refined genome-wide scores for noncoding variant impact.

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

**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 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 noncoding variants from eleVAR. **[[MG(14apr): ,bad]]** We will perform detailed validation on our top 10**[[MG(14apr): wrong]]**variants from eleVAR after medium-scale validation. **(a)** We will use TaqMan assays to genotype our top variants in 4,000 samples from a cohort prostate cancer patients. **(b)** We will select the 6 most recurrent variants in cancer samples from the cohort studies for biochemical validation by introducing the variants into their endogenous loci using the CRISPR-Cas9 system. We will then assay the effects of the variants on phenotypes such as overall gene expression (RNA-Seq**[[MG(14apr): ,wrong]]**), 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]]**

**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[1] **[[MG(14apr): ref]]** 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, repsectively. 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 Loci can be further utilized to investigate underlying disease mechanisms \cite{23482391}.

**B-3 Rare variants are significant to study of cancer & disease in general & have received less attention than common/GWAS variants**

There have been a large number of GWAS studies \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**

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 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 associated with increased disease risk 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. Thus, our approach will be among the first for functional interpretation of these variants. 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.

**[[JZ(12apr)2MG: in our discussion we decided to delete C-2? We just treat the RNA section as new features? Should delete the C-2 paragraph?]]**

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

**C-2 Analyzing variants in ncRNAs [[not sure C-4 Analyzing variants in ncRNAs ]]**

Moreover]]Most previous studies for functional interpretation of noncoding 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 this will form another major innovative component of our approach.

**[[MG(14apr): ,allelic innovative ]]**

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

**C-3 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.

**[[MG(14apr): 2step assess the assessment ]]**

**C-4 Functionally validating rare variants in depth to assess medium scale innovation**

Rare variations in regulatory regions of genome can have a paramount influence on biological processes and might function as primer for recurrent somatic mutations in adjacent genomic regions or might contribute to long range changes in chromatin regulation. Using a comprehensive panel of cell lines and genome editing tools like the CRISPR-CAS system we will introduce the rare variations in cell lines and study the effect on cellular physiology. This innovative approach will allow us to generate a catalogue of biological outcomes that can be attributed to a rare variation in a physiological setting.

**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**

**[[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**

Our proposed work is based on our experience in non-coding annotation. We have made a number of contributions in the analysis of the noncoding 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}, as well as new machine learning techniques \cite{19015141}. In addition, we have also proposed a probabilistic model, referred to as target identification from profiles (TIP), that identifies a given TF’s target genes based on ChIP-seq data \cite{22039215}. 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}

 **[[MG(14apr): put next para into prev… to long]]**

Furthermore, a comparative analysis of transcriptional regulatory features in diverse human, worm, and fly cell types (at different developmental stages and conditions) revealed remarkable conservation of general structural properties of regulatory networks despite extensive divergence of individual network features. \cite{25164757} We reported a large-scale transcriptome analysis \cite{25164755} across three species and discovered co-expression modules shared in animals and enriched in their developmental genes. In addition, a multi-organism comparison of pseudogenes suggested that pseudogenes are much more lineage specific than protein-coding genes, reflecting the different genome remodeling processes in each organism’s evolution \cite{25157146}. We introduced a framework to quantify differences between networks and by comparing matching networks across organisms, found a consistent ordering of rewiring rates of different network types. \cite{21253555} We developed a new comparative genomics tool, OrthoClust, for simultaneously clustering data across multiple species. OrthoClust \cite{25249401} integrates the co-association networks of individual species utilizing the orthology relationships of genes between species and has been used to obtain co-expression modules from worm and fly RNA-Seq expression profiles.

**[[MG(14apr): ,,shorten]]**$**[[MG(6apr)]]**$shrink **D-1-a-ii We have considerable experience processing RNA-seq data and annotating ncRNAs** by 50%$**[[MG(6apr)]]**$

$**[[MG(6apr)]]**$**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 ]]

**D-1-a-ii We have considerable experience processing RNA-seq data and annotating ncRNAs**

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 expression quantification of annotated RNAs and identification of splice sites and gene models \cite{21134889}. In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts \cite{22238592}. Comparisons between RNA-Seq samples, and to other genome-wide data, will be facilitated in part by our Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks \cite{21349863}. We have also developed tools that specifically analyze features of ncRNAs, including incRNA, an ncRNA-finder, and ncVAR, a prototype pipeline that genetic variants across biotypes and subregions of ncRNAs \cite{21177971, 21596777}. Finally, we have developed statistical models relating gene expression levels to chromatin marks and TF binding \cite{22955619,22955978,22060676,21926158}.

**[[JZ(10apr)2MG: a little bit confused here, the whole pipline is called eleVAR, not the RNA section? and this para is what we will do, not experience. suggest to be deleted]]**

**[[MRS2JZ/MG: this is copied from the CMG grant, where we called the RNA-specific tool eleVAR, and included one sentence about the tool before describing preliminary resutls. I added to the above paragraph and think that we can delete this paragraph]]**

**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 noncoding 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 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 noncoding mutation patterns in subtypes of gastric cancer**[[MG(14apr): ,submitted]]** Currently we are playing a leading role in TCGA/ICGA pcawg-2 (analysis of mutations in regulatory regions) group.

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

**[[MG(14apr): sites]]**

**\*\*\* D-1-b-i** identifying features associated with functional and conserved sites

\*\* Workflow. We plan to convert the current FunSeq prototype from its focus on somatic variants to allow the identification of rare variants associated with high functional impact. We will call this pipeline eleVAR . We will do some simple improvements in particular, Overall: we identify functional sites in the genome, some specific. We survey them in human polymorphism data. We find coherent groups of very specific sites that conserved -- ie these are our original sensitive sites. Then we use these as prioritizing features . We will do this consistently for TFBS and ncRNA.

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

**Improvement of sensitive regions definition in prototype FunSeq pipeline**

*Identifying gain-of- and loss-of-function mutations for TF binding sites*

We'll handle TF binding sites as in FunSeq with more emphasis on Loss-of- and gain-of-function variants 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. As the functional impact of the variant reflects the historical event when the polymorphism was first introduced in the human population, we will determine the ancestral allele of the variant to resolve between loss-of-function and gain-of-function.

*Better enhancer and enhancer target prediction*

As part of ENCODE enhancer prediction group, we are working on predicting confident sets of enhancers in human. We are currently developing 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

**More features to be analyzed in the eleVAR pipeline**

*RNA level biochemical interactions*

RNA interactions with proteins and miRNA are thought to be key in the regulation process \cite{22337053}. We will mine these interactions 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}. We will define protein/miRNA footprints on RNA as sensitive to mutation if they are highly enriched for rare genetic variants \cite{24092746, 25273974}. Our preliminary analyses indicate that some binding sites are comparable with or even more sensitive to mutation than coding sequences. We will also use these biochemical interactions to interpret the network context of our variants and prioritize variants that are bound by multiple factors.

*Other RNA level regulatory elements*

RNA structure is fundamental to the function of most well-studied ncRNAs \cite{24895857}. We will predict RNA secondary structures using RNAshapes and compare properties of structured and unstructed regions \cite{16357029}. Our preliminary results**[[MG(14apr): ,Init tuning]]**, 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.

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-ii 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**[[MG(14apr): ,as,done in]]** 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}.  **[[MRS(12Apr)]]For ncRNAs**, 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(14apr): ,more how… elaborate]] [[MG(14apr): ,element]] [[MG D-1-b-iii Variant prioritization based on Allelic activity & eQTL association (AlleleDB module) - JC more logic + shorten eqtl]]**

**D-1-b-iii 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, variants that are associated with allele-specific binding help us identify loci that might have potential functional impact in the differential binding of a particular transcription factor on each strand of the DNA. However, **[[MG(14apr): no oh the curr]]**variant prioritization scheme incorporating allele-specificity. One of the main challenges appears to be that allelic variants are enriched for rare variants\cite{24037378}. This is because since rare variants occur only in a few individuals, it will be difficult to justify the impact of these rare allelic variants in a prioritization scheme based on very limited supporting evidence. Hence, to enable prioritization based on allelic activity, we need to define allelic genomic elements. We derive allelic elements by first identifying allelic variants from hundreds of individuals. To better identify allelic variants, we will extend the AlleleSeq pipeline to account for the overdispersion of empirical read distributions observed in ChIP-Seq and RNA-Seq datasets \cite{25223782,20671027,22499706}. Subsequently, allelic variants (rare and common) identified across hundreds of genomes will be aggregated into ‘allelic genomic elements’. Each element will be assigned an ‘allelicity’ score based on its enrichment of allelic variants. The scoring system allows ranking of genomic elements based on its allelic impact, thereby enabling 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, expression QTLs (eQTLs). 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 QTLs in these databases.

**D-2 Approach Aim 2 - Implement an efficient & easy-to-use FunSeq pipeline & run on all the germline variants in TCGA/ICGC**

In this aim, we will provide an efficient implementation of FunSeq, including the development of a weighting system to bring together all its features, $**[[MG(6apr)]]**$,elaborate weighting system in FunSeq!$**[[MG(6apr)]]**$, the calling all the rare germline variants in sequenced tumor genomes, and then running FunSeq on them to develop a prioritized variant and element list. Overall, using FunSeq functional prioritization plus screening out the common variants will allow us to identify the rare variants **[[JC2MG]]** on a haplotype block with the greatest impact. We cautiously note that unlike GWA studies, 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-b Research Plan for Aim 2**

**D-2-b-i Math for gluing together features**

**[[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-1 We will develop a unified weighted scoring scheme for combining all eleVAR modules to consistently prioritize variants**

To integrate the various modules to predict ‘high-impact’ variants, we will develop a weighted scoring scheme, based on the mutation patterns observed in the 1000 Genomes polymorphisms \cite{25273974}. Constrained by selective pressure, natural variations tend to arise in functionally unimportant regions. Thus, features that are frequently observed 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’). For each discrete feature $d$, we will calculate the probability $p\_{d}\_{}\_{}$that it overlaps a natural polymorphism. Then we will compute 1-Shannon entropyas its weighted value $w\_{d}$.

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

The situation is more complex for continuous features, as different feature values have different probabilities of being observed in polymorphisms. Thus one weight cannot suffice. For a continuous feature $c$, which is associated with a score $v\_{c}$(e.g. PWM change), we will calculate feature weights for each $v\_{c}$. In particular, we will discretize at each value and compute $w\_{c}^{v\_{c}}$using (2). Now, when we come to evaluate the continuous feature $c$ for a particular variant, we calculate its weighted value using the corresponding feature score.

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

Finally, for each variant, we will score it by summing up the weighted values of all its features. We will also consider the dependency structure of features when calculating the scores.

**D-2-b-i-1 Tuning parameters with bayesian update**

We plan to tune eleVAR parameters using incremental bayesian learning strategy. The probability that a variant $v$is functional given the eleVAR score $\sum\_{i=1}^{m}w\_{i}f\_{i}$ follows logistic function $P(y=1|W,F)=\frac{1}{1 + exp(-k \* (\sum\_{i=1}^{m}w\_{i,}f\_{i}-a))}$ , with $F$ denotes different features of $v$and $W$ is the vector of feature weights. To update $W$ with newly available experimental data, we implement Bayes’ rule. $ P(W|Y,F) \infty 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 $Y$, $P(Y|W,F)P(W)=\prod\_{i=1}^{n}P(y\_{i}|w\_{1},w\_{2},...,w\_{m}, f\_{1},f\_{2},...,f\_{m})P(w\_{1},w\_{2},...,w\_{m})$, $y\_{i}=1$means positive result, whereas $y\_{i}=0$ denotes negative result, $w\_{m}$is the weight of feature $f\_{m}$. We will maximum this function to obtain updated weights $W$based on training data, using the initial weights $W$ as priors in $P(W)\_{}$. The updated weights will then be used as tuned parameters in eleVAR to prioritize variants.

**[[MG(14apr): ,data contxt issue ]]**

**[[MG-to-STL: update D-2-b-ii We will implement Elevar
In the above, we have developed features & an mathematical sys for putting them together... here we will desc how we will build a high throughput software implementation.... (2 para)]]**

**]]**

In the above, we have developed features and an 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 flexible for users to parameterize and customize it 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 do analyses on their own computers or deploy it on the cloud. 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 features that get integrated into our software come from large scale genomic datasets, a naive query approach will be very time-consuming and space-inefficient. To address this problem, we will pre-process the data, build relationships between features and compress data into smaller context. To facilitate efficient processing, we will build our software on top of Apache Spark and Parquet to better scale with cluster computing and leverage fast accessing of memory. As security is a big concern in genomics research, we will be using https link and crypto algorithms to protect users information. Also, we allow users to clean up their computational footprints. Last, to address the portability issue, we will provide a download version that is contained in a Docker container.

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

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

**[[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. We will be utilizing these germline variant calls for further downstream analyses.**[[MG(14apr): more how merge callers]]**

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

**D-2-b-iii-2 We will run elevar on all the variants & prioritize them**

We will run elevar on the rare variants resulting from our variant calling on ~2500 TCGA/ICGC whole-genome sequences. We expect an average of ~40K rare variants per genome \cite{23128226}. Since they recur rarely at the exact same position, we anticipate a prioritized list ~100M variants (=40K \* 2500 genomes). We note that unlike GWA studies, 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 are more likely to have an impact. 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 (2500 genomes).**[[MG(14apr): ,',recurr elsewhere ]]** Thus, with ~100M variants distributed evenly across the human genome, assuming 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 50. 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.

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

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

assess v publically available functional datasets

We will iterate a number of times with the experimental validation in aims 3 n 4 to tune the parameters and finally to create an unbiased assessment of the results. We will conduct our first round of tuning with publicly available validation results. Several high-throughput technologies have been developed to test functional impact of noncoding 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.

**[[MG(14apr):** mention recurr & burden validation]] **$[[MG(6apr)]]$,move B4 two elsewhere into aim2.... Shorten,,, [[MG(14apr): ,,2hit & prostate focus]]**

We would further carefully evaluate **[[MG(14apr): ,validate]]**the germline mutation burden of cancerous individuals against the more than 2000 controls samples from the 1000 genomes project. Specifically, we will aggregate rare mutations in each regulatory element in our updated sensitive feature list, and use association methods (such as SKAT \cite{23684009}) to evaluate the cummulative effects of rare variants in cancerous patients. **[[MG(14apr): ,PROSTATE[[MG(14apr): ]]**.As a results, a list of heavily mutated regulatory elements in cancer individual but not in controls will be reported as candidate regions. To gather further information on these regions, we would also run our !**[[MG(14apr): no mention larva]]**pipeline on the somatics variants in cancer samples to evaluate their somatic mutation burden. Regions with both heavy germline and somatic mutation burden would be more likely to have deleterious phenotypic effect. We would further investigate the disease GWAS datasets and HGMD database, and discovery of disease causing mutations in our discovered regions would provide further evidence of our results.

**D-2-b-v 2nd round of tuning & then assessment based on experiments performed in this project [[YF]]**

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 **[[MG(14apr): ,,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/Cas 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]]**

**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**

To set up our Clone-seq pipeline, we first focused on 27 interactions that can be detected by our version of Y2H, are represented in co-crystal structures, and have known missense disease mutations on their corresponding proteins in HGMD. Of these 27 chosen interactions, 24 have disease mutations on the corresponding interaction interfaces and 15 have mutations away from the interfaces. For interactions that have more than one mutation on and/or away from the interfaces, we randomly picked one for each interaction. To generate these 39 mutant alleles, we picked 4 colonies for each mutation. As a reference, we also pooled together all the WT alleles in our human ORFeome library to be sequenced together with the 4 pools of the mutagenesis colonies. In total, there are 40.1 million Illumina HiSeq 1×100 bp reads for our Clone-seq sample.

These reads were then de-multiplexed and mapped to the genes of interest using the BWA aln algorithm. There is an average of > 2,500× coverage at all desired mutation sites. For each allele of interest, we identified all reads that map to the position of the mutation of interest (Rall) and those that actually contain the desired mutation (Rmut). We then calculated a normalized score that quantifies the fraction of reads that contain the desired mutation:

S = Rmut/Rall x 1/k

where k is the number of different mutations for the same gene.

Out of 156 colonies containing the 39 mutations, 125 of them were successful. Thus, our overall PCR-mutagenesis success rate is 80%. In fact, we were able to pick correct clones for all 39 mutant alleles using only the first two pools in Clone-seq. All 78 clones from the first two pools, from which the correct ones used in subsequent steps were selected, were Sanger sequenced for verification. All 55 Clone-seq positive results with S > 0.8 were confirmed, and there is a clear separation in the S scores between the successful and failed clones (Fig. 4). One major advantage of our Clone-seq pipeline is that we can now carefully examine whether there are other unwanted mutations introduced during the PCR process. We found that there are on average 4-5 additional mutations introduced in each pool of the 39 colonies. This corresponds to a 0.013% error rate, in agreement with previous studies. The detection of additional mutations, especially those far away from the mutation of interest, cannot be achieved with the traditional site-directed mutagenesis pipeline using Sanger sequencing. These unintended mutations could lead to erroneous downstream results.

Table 1. Cost comparison between Sanger and Illumina sequencing1.

|  |  |
| --- | --- |
| Traditional Sanger sequencing | Clone-seq |
| Unique mutations | 3,047 | NEBNext Multiplex Oligos (E7335S) | $19.80 |
| Colonies per mutation | 4 |
| Total number of samples | 3,047x4=12,188 | NEBNext DNA Library Prep Master (E6040S) | $105 |
| Re-sequencing needed2 | 5% |
| Number of 96-well plates needed | 137 | Illumina HiSeq, single-end, 100 bp sequencing lane | $1,175 |
| Cost per plate | $300 |
| Minimum cost3 | 43x300=$12,900 | Total cost | $1,299.80 |
| Total cost | 137x$300=$41,100 |

1All costs are based on internal Cornell pricing.

2Sanger sequencing has an average failure rate of 5%.

3The minimum cost is the least amount of money spent in Sanger sequencing the expected number of samples needed to obtain one correct clone for each mutation of interest.

For our Clone-seq samples, we obtained only 40.1 million reads out of a total of 125 million reads in a single lane of a 1×100 bp HiSeq run with >2,500× coverage. However, to determine S to a least count of 1%, we only need 100× coverage. Since the separation between a successful mutagenesis attempt with the lowest S and an unsuccessful mutagenesis attempt with the highest S is 0.28, 100× coverage makes this separation >25 times our least count. We further increase this separation to >60 times our least count by requiring S > 0.8 for a mutagenesis attempt to be considered successful. 100× coverage is also sufficient for a conservative variant calling pipeline to identify additional unwanted mutations with high confidence35,36. Thus, we can obtain 39×(125/40)×(2,500/100) = 3,047 alleles with a single lane of a 1×100 bp HiSeq run using the Clone-seq pipeline. Overall, our Clone-seq approach will drastically improve the throughput of site-directed mutagenesis and decrease the total cost by at least 10-fold (Table 1).

To further test Clone-seq, we identified a set of 446 SNVs from the published ESP6500 dataset36 that are at the interface of protein interactions and are amenable to testing using our high-throughput Y2H approach. Using our Clone-seq pipeline, we performed large-scale, site-directed mutagenesis to generate clones for these 446 SNVs. We sequenced 4 colonies each for the 446 alleles of interest using one full 1×100 bp MiSeq run. We obtained ~14 million reads and aligned them to the reference sequence database using BWA79. For each allele of interest, we identified all reads that map to the position of the mutation of interest (*Rall*) and those that contain the desired mutation (*Rmut*). The read coverage surrounding the mutation of interest was ~300× per allele. Using a threshold of *S* > 0.8, approximately 75% of the colonies contain the desired mutation. We were able to choose at least one colony that contains only the desired mutation (without additional unwanted ones) for 437 of the 446 mutagenesis attempts, a success rate of 98.0%.

Overall, our pipeline has been significantly optimized to make it very efficient. We established a web-tool (<http://www.yulab.org/Supp/MutPrimer>) to design mutagenesis primers both individually and in batch. MutPrimer can design ~1,000 primers for ~500 mutations in one batch in less than one second. All primers for the 476 mutations in this study were generated by MutPrimer. All mutagenesis PCRs are performed in batch using automatic 96-well procedures. Since single colony picking after bacterial transformation of mutagenesis PCR product is a rate-limiting step, we rigorously optimized this step and found that adding 10 µL mutagenesis PCR products to 100 µL competent cells and plating 50 µL transformed cells give the best transformation yield and well-separated single colonies. Furthermore, rather than individually streaking transformed cells onto agar plates one sample at a time, we were able to significantly increase throughput by spreading colonies using glass beads onto four sector agar plates which are partitioned into four non-contacting quadrants. In this manner, a 96-well plate of transformed bacteria can be plated out onto 24 four-sector agar plates in ~15 minutes. Traditional site-directed mutagenesis pipelines require miniprepping each of the selected colonies and sequencing them separately by Sanger sequencing. To drastically improve the throughput of our Clone-seq pipeline, we pooled together the bacteria stock of a single colony for each mutagenesis attempt to perform one single maxiprep, which makes the library construction step much more efficient and amenable to high-throughput. Furthermore, existing variant calling pipelines cannot be applied to our Clone-seq results because the expected allelic ratios built into these pipelines are a function of the ploidy of the organism. However, in our Clone-seq pipeline there is no concept of ploidy. We pool together many mutations for one gene in the same pool (e.g., 40 mutations for *MLH1*) and different genes often have different numbers of mutations. Our *S* score calculation and unwanted mutation detection pipeline was designed according to our pooling strategy .

In total, we have used the Clone-seq pipeline to successfully generate 476 (39 + 437) 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 be scaled up to generate the ~1000 SNVs as proposed.

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

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 noncoding RNA. Recruitment of ER upstream of NEAT1 lncRNA was addressed in greater details. 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**

Identification of rare variants and understanding the influence thereof on repertoire of biological responses will afford us a unique opportunity to understand causal role of these variations on other somatic mutations associated with diseased states including but not limited to cancer.

We will use Clone-seq to generate ~300 candidate non-coding variant clones identified in Aim 1 and 2. The clones will then be subjected to the downstream reporter assays. 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, 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 site-directed mutagenesis PCR and E. coli transformation.*** Primers for site-directed mutagenesis are selected based on an optimized version of the protocol accompanying the QuikChange Stratagene site-directed mutagenesis kit (200518). 50 µL mutagenesis PCR reactions are set up on ice in 96-well PCR plates using Phusion polymerase (NEB M0530) according to manufacturer’s manual. All WT clones are obtained from the Human ORFeome 8.181. 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*). *E. coli* competent cells are prepared in 96-well plates with 20 µL cells per well. 10 µL of *DpnI*-digested PCR 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.*** DNA plasmids from all four colonies of all alleles are mini prepped using our fully-automated 96-well miniprep pipeline. 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. The *S* score for each colony of each allele is calculated as described above. As shown in **Fig. 4**, we found that all clones with *S* > 0.44 are confirmed to be correct via Sanger sequencing with a clear separation between those that are correct and those that are not. However, to ensure that the clones we pick are correct, we require *S* > 0.8 for a colony to be scored as containing the desired mutation.

**D-3-b-i-(3) Functional consequences: Reporter assays**

Reporter assays that employ either LUC or next generation reporter vectors can provide direct insight to functional relevance of SNPs on target gene. GeneCopoeia offers Gaussia-luciferase (GLuc), eGFP,or mCherry based lentiviral or non-viral promoter reporter clones that can serve as efficient tools to study transcription regulation. Minimal essential promoter region for each WT target gene will be subcloned from germline DNA using TOPO cloning kit (Invitrogen). If patient sample that harbors the mutation is available, we will amplify the corresponding mutant promoter sequence from the genomic DNA of the patient. PCR products will be cloned upstream to pGL-4-LUC promoter reporter plasmid or upstream to Gluc vectors. For each WT DNA Target gene-promoter plasmid a corresponding MT DNA Target gene-promoter plasmid will be generated using site directed mutagenesis utilizing QuikChange Lightning (Agilent). In this way we will have 300 WT promoter plasmids and 300 MT promoter plasmids in both PGL-3 LUC and Gluc background. We will utilize a panel of adherent cell lines. 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.

Cells will be seeded in 6 well plates and transfected with promoter reporter WT and mutant plasmid constructs. 48 hrs after transfection promoter activity will be measured following manufacturer’s instructions. Assay values will be normalized using internal renilla luciferase as control.

Our expectation is that *in vitro* promoter LUC assays will inform us if a particular mutation had any effect on transcription.

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

**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 CAS 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 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. 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 CAS 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 evaluationof 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/Cas 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/CAS system**

We will utilize the newly discovered CRISPR/CAS 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/CAS 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/CAS system will be used to generate the mutation.

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

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.

**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 noncoding variants****Aim 2: Optimization & beginning of variant calling****Aim 3: development of validation assays** |
| **Year II** | **Aim 2: Germline variants called from ICGC/TCGA 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** |