Introduction (response to reviewer criticism of the previous submission)

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

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

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

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

· We have also added a new experimental collaborator, Dr. Haiyuan Yu. Dr. Yu is part of the original FunSeq collaboration between Yale and Cornell and brings significant expertise to the grant in terms of medium-scale enhancer validation. Dr. Yu will implement his newly-established massively-parallel site-directed mutagenesis pipeline, Clone-seq[1], to generate ~1200 specific mutations in ~600 enhancers in three rounds to tune the parameters of the our variant prioritization pipeline and comprehensively evaluate the performance of the pipeline at the end.

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

We were also criticized in the original submission regarding the large “computational load for handling all germline variants from TCGA and ICGC whole genome sequences.” We addressed this by collaborating with the PCAWG germline variant calling group (See letter of collaboration form J Korbel, head of PCAWG-8), enabling us to cut almost all 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[2, 3].

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 these strengthes of the previous submission; however, 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[4], which is still a paper focused on somatic variants in cancer but which incorporates some of the ideas proposed in the original grant. Consequently, we have moved some of our originally proposed ideas now into preliminary results and elaborated on new ideas focused on allelic variants in our proposal. Overall, our revised proposal provides a strengthened computational framework for rare variant prioritization, with more specifically targeted validation that will enable us to tune and assess our pipeline for prostate cancer as a prototype targeted disease system.

A. Specific aims

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 focused on particular disease, such as prostate cancer.

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

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

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

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

We will perform a detailed in-depth experimental validation on 10 representative positive variants from eleVAR after medium-scale validation. **(a)** We will use TaqMan assays to genotype our top 6 variants in 4,000 samples from a cohort prostate cancer patients. **(b)** We will further evaluate these 6 variants in cancer samples from the cohort studies for biochemical validation by introducing them into their endogenous loci using the CRISPR-Cas9 system.We will then assay their downstream effects on gene-expression using real-time quantitative PCR analysis, as well as cell viability, migratory potential (for metastasis), and transcription factor binding (ChIP and EMSA).

B. Significance

B-1 Non-coding variants are significant in the context of human diseases but are less well-studied than coding ones

Numerous studies have been conducted on the mutations that lie in coding regions. However, several preliminary studies suggest that variants in non-coding regions can significantly influence an organism's phenotype[5], and they are often implicated in various diseases[6, 7]. Many non-coding variants impact regulatory elements. Such variation in the human genome can modulate gene expression[8], and changes in this expression have been implicated in cancer and other diseases[9, 10, 11, 12, 13, 14].

B-2 Rare variants play important roles in human disease, though they have received less attention than common/GWAS variants

There have been a large number of GWAS[15],which have primarily focused on associating common genetic variants with diseases. However, growing evidence suggests that rare genetic variants may have strong effects in many human diseases, including cancers[16]. Increased disease susceptibility is often attributed to the cumulative effect produced by multiple rare variants[17]. For instance, rare germline variants in the CHEK2 gene[18] and in the HBOX gene[19] were associated with breast cancer and prostate cancer, respectively.

B-3 Recent progress in annotating non-coding regions of the genome provides new opportunities for variant annotation

Annotating non-coding regions is essential for investigating genome evolution[20], understanding important biological functions (including gene regulation and RNA processing)[21], and for elucidating how SNPs and structural variation may influence disease[22]. The Encyclopedia of DNA Elements (ENCODE) and the model organism ENCODE (modENCODE) Project provide extensive genomic annotation of human[23], drosophila[24] and *C. elegans[25]* genomes. Furthermore, the functional landscape of regulatory variations in the human genome has been investigated by large-scale mRNA and miRNA sequencing[26, 27, 28, 29]. Similar efforts have also been directed toward annotating human epigenomic data, as well as understanding the influence of genomic variation on the gene expression profiles[30]. These Expression Quantitative Trait Loci (eQTL) can further be utilized to investigate underlying disease mechanisms[31].

C. Innovation

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

C-1 Identifying and interpreting rare non-coding variants, consistently for both TF binding & ncRNAs, using population-scale polymorphism data The GWAS catalog contains many common variants associated with disease. However, as discussed above, many rare variants may increase susceptibility to various diseases. Currently, no standard methods exist to functionally interpret such variants, especially in non-coding regions. Furthermore, prior studies aimed at 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 the impact of variants in ncRNAs, and will thus be one of the first comprehensive approaches to decipher the functional interpretation of such variants.

C-2 Prioritizing variants based on elements enriched in allelic activity

Previous studies have identified regulatory variants using allele-specific gene expression[28, 32]. However, there has not been a scheme that allows us to prioritize variants based on allelic activity, especially rare variants that do not usually overlap with identified variants. In the proposed work, we will prioritize variants based on their presence within allelic elements or regions of the genome. To define an allelic element, we will assign an 'allelicity' score, which will be dependent on the enrichment as well as recurrence of allelic variants in that genomic element across multiple individuals. This element-based strategy will allow us to prioritize even very rare variants from different genomes.

C-3 Developing a weighting system for prioritization and a plan for tuning its parameters by multiple rounds of high-throughput experimental characterization

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 the weight of individual features associated with that particular variant. In the second iteration of this workflow, we will apply a Bayesian learning strategy to tune weights based on experimental observations. Subsequently, these updated weights will be assigned to prioritize rare variants.

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

Current protocols for site-directed mutagenesis require the selection of individual colonies and subsequent sequencing of each colony using Sanger sequencing, which makes them labor intensive, expensive and unscalable for genome-wide surveys. Using Clone-Seq, we can generate clones for ~3,000 mutations in one lane of an Illumina HiSeq run and decrease the cost by more than 10-fold[1] (see D-3-a-iv). Please note that Clone-seq is entirely different from previously described random mutagenesis approaches [33, 34, 35, 36]. In Clone-Seq, each mutant clone has a separate stock. Different clones can therefore be used separately for completely different downstream assays.

D. Approach

D-1 Approach Aim 1 - Convert & extend the FunSeq somatic variant pipeline for germline prioritization

D-1-a Preliminary results for Aim 1

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

Our proposed work is based on our experience in non-coding annotation. As part of our 10-year history with the ENCODE and modENCODE projects, we have made a number of contributions in the analysis of the non-coding genome. Our TF work includes the development of a method called PeakSeq to define the binding peaks of TFs[37], target identification from profiles (TIP) to identify a TF's target genes[38], as well as new machine learning techniques[39]. Furthermore, we have developed machine-learning methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers[40], which we have partially validated[41]. We have constructed linear and non-linear models that utilize TF binding and histone modification signals as input to predict the transcriptional output of a gene [42]. Using these methods on a diverse set of model organisms (from yeast to human genomes [25, 43, 44, 45]), we have achieved high predictive expression levels in the K652 cell-line [42].We have also constructed regulatory networks for human and model organisms[46], and completed many analyses on them[25, 41, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58]. Furthermore, we have conducted large-scale multi-organism regulatory and gene coexpression network comparisons, along with transcriptome and pseudogene lineage analyses for the modENCODE project[58, 59, 60, 61]. We also have extensive experience conducting integrated analyses of RNA-Seq datasets, such as those generated by the ENCODE, modENCODE, BrainSpan and exRNA consortia[23, 25, 62, 63, 64]. In particular, for general RNA-Seq analysis, we have developed RSEQtools and IQseq for splice site identification, gene model creation and transcript quantification[65, 66]. 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[67].

D-1-a-ii We have extensive experience in allelic analyses

A specific class of regulatory variants is one that is related to allele-specific events. These are variants that are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and allele-specific expression (ASE)[68, 69]. We have previously developed a tool, AlleleSeq,[57] 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[46, 55, 62]. Overall, we found that these allelic variants are under differential selection from non-allelic variants[46, 55]. 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[46]. Furthermore, we have implemented the AlleleSeq tool, generated lists of detected allelic variants, and constructed a personal diploid genome and transcriptome of NA12878 on[70].

D-1-a-iii We have extensive experience in relating annotation to variation, and we have developed the FunSeq pipeline for somatic variants

We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets[41, 46, 71]. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations[71]. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region[23]. Further studies by our group showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery[54, 56]).

In recent studies[4, 55], 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 selective pressure, as determined using genomes from many individuals from diverse populations). FunSeq links each non-coding mutation to target genes, and prioritizes such variants based on scaled network connectivity (we compute the percentile rank of a particular gene after ordering centralities of all genes in a particular network). It identifies deleterious variants in many non-coding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. It also detects their disruptiveness in 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 enables the identification of candidate non-coding driver mutations[55]. Our method is able to prioritize the known TERT promoter driver mutations, and it scores somatic recurrent mutations higher than those that are non-recurrent. In this study, we integrated large-scale data from various resources (including data from ENCODE and The 1000 Genomes Project) with cancer genomics data. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples. We have also submitted a paper applying our method to investigate non-coding mutation patterns in subtypes of gastric cancer. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

D-1-b Research plan for Aim 1

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

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

In order to define rare variants with highly impactful events, we will use both intra-human variation data (from The 1000 Genomes Project) as well as cross-species evolutionary conservation (using classical measures such as GERP score).

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

Second, RNA regulatory elements will be added as prioritization features in a way that is consistent with the approach taken for TF-binding sites. Specifically, we will mine RNA interactions with proteins/miRNAs from publically available data, such as CLIP-Seq, CLASH and computational predictions (TargetScan) to create a compendium of biochemical interactions with RNA [72, 73, 74, 75]. Our initial analyses indicate that some binding sites are comparable with or even more sensitive to variation than are coding sequences. In addition, we will incorporate RNA structural elements. Our initial survey indicates that more rigid RNA structures, such as stems, are under higher selective pressure than other RNA regions, and that those variants that incur a larger free energy change of the structures tend to be more rare 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 high-impact mutations: breaking & creating motifs

For impactful events at TF binding sites, we will use motif breakers and motif formers to define loss-of- and gain-of-function events, respectively, as these events are more likely to have deleterious consequences [10, 11, 55, 71, 76, 77, 78]. 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 is to employ ancestral alleles to get a more accurate determination of these events.

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

Studies of RNA processing and function have identified key motifs associated with events ranging from RNA splicing to chemical RNA base modifications[79]. 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. Thus, variants that occur in regulatory motif regions will be scored based on the degree to which they break motifs.

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

To interpret the likely functional consequences of non-coding variants, we will comprehensively define associations between many non-coding regulatory elements and their 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[40]. 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[54, 56]. 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.

D-1-b-iv Variant prioritization based on allelic activity

The evident regulatory roles of the allele-specific variants assert that they will be useful for identifying functional variants. For example, if we can associate the differential binding effect of a particular transcription factor with different alleles of an SNV, we can identify loci that have potential functional impacts in regulation. However, because allelic variants are enriched for rare variants[26] that occur only in a few individuals, it will be difficult to match the specific rare variants to prioritize in an individual against those earlier determined to be allelic in a functional genomics experiment on a cell line. Hence, instead of prioritizing by the direct overlap of allelic variants, we need to prioritize by the presence of allelic variants within '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 The 1000 Genomes Project[80]. We will match them with their corresponding RNA-Seq and ChIP-seq experiments from multiple disparate studies, such as gEUVADIS[26] and ENCODE[23]. Because these separate studies typically have various inconsistencies in terms of tools and parameters used in processing their data, we have to reprocess and harmonize the heterogeneous data and detect allelic variants in a uniform fashion. Also, while the conventional way to detect allelic variants is using the binomial test, previous studies have found that the distributions of the allelic ratios in ChIP-seq and RNA-seq experiments have been empirically observed to give a broader, or an ‘overdispersed’, distribution than a binomial distribution[81, 82, 83]. To identify and remove problematic "outlier" datasets (to harmonize the data corpus) and to account for overdispersion of read distributions, we will extend our detection pipeline (AlleleSeq) to include the calculation of an overdispersion parameter for each ChIP-seq and RNA-seq dataset; the beta-binomial test (which parameterizes the overdispersion) will be used to detect allelic variants instead of the binomial test.

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) it enables incorporation of ASE and ASB into the main prioritization scheme; input variants (even those which are rare, but lie in highly-ranked allelic genomic elements) will be upweighted according to their scores.

D-1-b-v We will use a unified weighted scoring scheme for combining all eleVAR features to 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[4]. Constrained by selective pressure, common variations tend to arise in functionally unimportant regions. Thus, features that are enriched with common polymorphisms are less likely to contribute to the deleteriousness of variants and are weighted less. In general, features can be classified into two classes: discrete (e.g., within or outside of a given functional annotation) and continuous (e.g., the PWM change in 'motif-breaking'). We will weigh these two sets of features with different strategies.

For each discrete feature **d**, we calculate the probability $\mathbf{p_d}$ that it overlaps with a common polymorphism. We then calculated the information content to denote its weighted value $\mathbf{w_d}$.

$$w_d = 1 + p_d * \log_2 p_d + (1 - p_d) * \log_2 (1 - p_d) \quad (1)$$

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 $\mathbf{v_c}$, we will calculate feature weights for each $\mathbf{v_c}$. In particular, we discretize at each value and compute $\mathbf{w_c^{v_c}}$. When we then evaluate the continuous feature **c** for a particular variant, we calculate its weighted value using the corresponding $\mathbf{v_c}$.

$$w_c^{v_c} = 1 + p_c^{\geq v_c} * \log_2 p_c^{\geq v_c} + (1 - p_c^{\geq v_c}) * \log_2 (1 - p_c^{\geq v_c}) \quad (2)$$

Feature f_x ($x \in \{d, c\}$) is binary, and is given a value of 1 if if that feature is observed. We score each variant by summing up the weighted values of all its features (3). We will also consider the feature dependency structure when calculating the scores (e.g., removing redundant features or performing dimension reduction techniques).

$$s = \sum_d w_d f_d + \sum_c w_c^{v_c} f_c \quad (3)$$

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

D-2-a Workflow for Aim 2

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

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

D-2-b Research plan for Aim 2

D-2-b-i Statistical framework for parameter tuning using Bayesian updates

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

The probability that a variant **v** is functional, given the eleVAR score **S** (equation 3 in D-1-b-v), follows a logistic function $\square(\square = 1|\square) = \frac{1}{1+\square\square\square(-\square * (\square-\square))}$, with **F** denoting the feature vector of **v** and **W** denoting the vector of feature weights. To update **W**, we implement Bayes' rule: $\square(\square|\square,\square) \propto \square(\square|\square,\square)\square(\square)$. The probability of observing **W** (given training data **Y**) is proportional to the probability of observing **Y** given **W** and **F**, multiplied by the prior probability of **W**. Assuming independence between data points in **Y**, which can be achieved by proper training data construction, $\square(\square|\square,\square)\square(\square) = \prod_{\square=1}^{\square} \square(\square_{\square}|\square_{1},\square_{2},\ldots,\square_{\square},\square_{1},\square_{2},\ldots,\square_{\square})\square(\square_{1},\square_{2},\ldots,\square_{\square})$, with $\square_{\square} = 1$designating a positive result, whereas $\square_{\square} = 0$ denotes a negative result, and $\square_{\square}$is the weight of feature $\square_{\square}$, given **m** total features.

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

D-2-b-ii Software implementation using the data context and dependency graph

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 will also be able to download this software and install it on their local machines or deploy it on the cloud. We will provide a downloadable version that has been configured in a Docker container to minimize portability issues. 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 results.

As our software uses features coming from large-scale genomic datasets, calculating scores is very time-consuming, space-inefficient and probably computationally intractable for some researchers. To address this problem, we will first provide pre-calculated 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 to modularize the calculating process. We will recognize critical interprocedural interfaces (e.g., intersections in which multiple flows merge) that are likely to get updated and save intermediate data files to facilitate fast rebuilding and recovery. After updating some data sources or partial corruption of runtime data files, our software will use a data flow map to identify the flow paths that require rebuilding. All other unperturbed paths will use the nearest intermediate data files and do minimal recalculation. By carefully removing data dependencies, mapping data flow paths and localizing the rebuilding after updating, we will give users the ability to customize and constantly update our model and software at minimal cost. We will also use NoSQL databases, such as MongoDB to maximize our data model flexibility. In particular, users will be able slightly perturb the data context with the addition of a single targeted functional genomics experiment or even a new data type.

D-2-b-iii Generating an initial list of prioritized variants & then running eleVAR on them

The PCAWG-8 group will be generating high-quality germline call sets (comprising SNPs, Indels, and SVs) for relatively high-coverage whole-genome datasets. The germline SNP call sets will be generated by four of the most state-of-the-art variant callers, including the GATK HaplotypeCaller[86], which is run by the Broad Institute, and Caveman[87], which is run by the Wellcome Trust Sanger Institute. These call sets will then be integrated as priors into the tool FreeBayes, which will generate the final call set for further downstream analyses.As we will be focussing on prostate cancer, we will add to this list a number of other whole-genome sequences of prostate cancers (tumor and normal) [88, 89]. We will call the variants in these genomes in a way that is consistent with what is done in PCAWG. At the start of the project, we estimate that we will have at least a total of 250 prostate cancer genomes. We will call this set of genomes and variants the "prostate compendium."

We will run eleVAR on the rare variants resulting from our variant calling on both PCAWG and on the prostate compendium whole-genome sequences. During this process, we will add biological context to the general scoring scheme in eleVAR, as this could help prioritize variants that lead to tissue-specific phenotypic effects[90]. We will build a tissue-specific protein-protein interaction network (based on proteins that are expressed in prostate tissue), as well as a tissue-specific gene regulatory network (histone modification to define active promoters and enhancers, as well as scoring the change in PWM for motifs affecting binding sites of TFs and RNAs expressed in prostate tissue). Rare variants that affect the hubs and bottlenecks in the tissue-specific networks will be prioritized in the modified eleVAR scoring scheme.

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

To perform the 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)[91] and ClinVar[92] catalogues large numbers of regulatory disease-causing mutations. Several high-throughput technologies have been developed to test the functional impacts of non-coding genomic variants. For example, Kwasnieski et al. used CRE-seq[93] to assay over 1,000 single- and double-nucleotide mutations in promoter regions. Kheradpour et al.[76] used MPRA to test variants affecting regulatory motifs in over 2,000 human enhancers. Rare mutations close to GWAS tag SNPs are probably more deleterious than variants elsewhere. We will utilize these datasets to perform comparisons with other variant prioritization methods, such as CADD[94], to perform a preliminary evaluation of method performance. We will then tune our parameters using the scheme described in D-2-b-i.

We will further compare the germline mutation burden of healthy individuals with those suffering from cancer. Specifically, we will use over 2500 normal samples from The 1000 Genomes Project as the control data, and run a mutation burden test using available software such as SKAT. (If it is necessary to expand the controls for rarer variants, we could use deeply sequenced trios from the 1000 Genomes Project[95], 500 individuals with Complete Genomics sequencing also from 1000 Genomes[96] and healthy individual from the UK10K project[97].)

In contrast to the binning process generally used, which is relatively *ad-hoc*, we will aggregate rare mutations in each regulatory element in our updated sensitive feature list to evaluate the cumulative effects of rare variants in cancer patients. As a result, a list of heavily mutated regulatory elements in cancer patients (but missing in healthy controls) will be reported as candidate regions and would be upweighted during the tuning process. In addition, since the validation work is done in prostate cancer cell lines, we would further focus on our compendium of prostate cancer WGS (see above) to investigate the germline mutation burden on the noncoding regulatory elements. Both genome-wide and regional tests based on known loci will be analyzed to check for suspicious sites.

The interplay between germline and somatic variants may increase cancer risk, but they are not frequently analyzed 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[11, 98]). In our study, we will also analyze the somatic mutation burden in our feature list. Different from the germline mutation burden test, our novel computational framework (called LARVA) is used to directly evaluate the somatic mutation burden in cancer samples. LARVA would provide a list of heavily mutated noncoding regulatory regions, and we will compare these results with the germline mutation burden test. Regions that are heavily mutated by both germline and somatic variants should be up-weighted in eleVAR.

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

We expect an average of ~40K rare germline variants per genome[80]. Since they rarely recur at the exact same position, we anticipate a prioritized list of ~8M variants (=40K * 250 genomes, based on the the expected size of the prostate compendium). We will select 500 functional regions of appreciable size that contain highly ranked variants. Assuming ~8M variants are distributed evenly across the human genome, taking an average element size of 3kb, the number of variants per element will be ~4. Variants on the same element are expected to have different functional impacts. For each element, we will prioritize at least one of these variants to be of high impact, and the remaining variants to be of a differential impact. Specifically, we will have a total of 1000 variants (500 with a high impact and 500 with a low impact).

Subsequent tuning and refinement of the eleVAR parameters will be based on further experimental characterization of these 1000 variants (500 highly prioritized and 500 lowly, respectively). We will validate these variants through medium throughput functional genomic screens using the Clone-Seq technology coupled with luciferase reporter assays. Overall, this refinement will be accomplished in two rounds, each round per year, as detailed in 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.

D-3 Approach Aim 3: High-throughput experimental characterization of the prioritized variants

We will use our massively-parallel Clone-seq pipeline and high-throughput luciferase reporter assays to clone and examine 1200 SNVs on 600 enhancer elements to experimentally characterize their impact on gene regulation to fine tune and validate our eleVAR pipeline.

D-3-a Preliminary results related to experimental characterization

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

To set up our Clone-Seq pipeline, we attempted to generate clones for 1034 mutations on 223 genes, including 40 mutations for *MLH1*. We picked 4 colonies for each mutation (4106 in all). 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[80, 99]. 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[1].

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[100]. 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 for both 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[101, 102]. 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[103]). 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[1].

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-ii Experience with luciferase reporter assays confirming validity of *in silico* TF binding sites

We have a great amount of experience with developing reporter assays for TF binding. In particular, we have done an earlier study where we validated many binding sites for estrogen receptor (ER) related to prostate cancer. We have also done validation for FunSeq prototype pipeline through collaborations among Gerstein, Yu, and Rubin groups. This is similar to what will be done here but was for somatic rather than germline variants. The Yu group generated three mutations on WASP and examined their impact on WASP's interaction with six other proteins. The Rubin group examined a mutation in the RET promoter predicting a gain of an AP1 motif that was determined using the in silico FunSeq pipeline. Using the luciferase reporter assay, the Rubin group studied the promoter activity of the WT and mutant RET promoter in the DU145 cell line. Luciferase activity confirmed that the mutant promoter was 1.2-1.3 fold more active than the WT promoter (Fig XXX).

D-3-b Research plan related to validation

D-3-b-i Overview of validation strategy

Because of the throughput of our Clone-Seq and luciferase reporter assays, we will perform iterative learning and validation in three rounds. In each of the first two rounds, we will select and clone 250 enhancer elements and two variants on each elements that have high and low eleVAR scores, respectively (500 variants total per round). Based on the reporter assay results, we will fine-tune the parameters of the learning algorithm as described in **Aim 2**, and then perform the predictions again. In the third round, We will select and clone another 100 enhancer elements and one high scoring and one low scoring variant on each element to confirm the performance of our algorithm. Top candidate SNVs 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[104]. 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 OD_{600}. 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 . 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. 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. AR+ LnCaP cells and AR- PC3 will be seeded in 96-well plates and transfected with WT and mutant enhancer constructs. 48 hrs after transfection, enhancer activity will be measured following manufacturer's instructions (Promega E2940). Assay values will be normalized using internal renilla luciferase as control. Our expectation is that *in vitro* luciferase assays will inform us if a particular mutation had any effect on transcription.

D-4 Detailed validation of specific variants

In this aim we strive to examine in detail ~10 variants that we find as positives through the high-throughput experimental characterization. The luciferase assays in Aim 3 are often considered as in vitro characterizations. In Aim 4, the goal is to understand the molecular basis for the observed impact of the variants and how changes in gene expression caused by these variants might lead to disease. We first describe our preliminary results in screening against a large cohort for genetic validation and in applying the CRISPR/Cas technology for in vivo experiments. Then we describe how we will carry this out for the ~10 variants culled. We will choose ~10 variants to be representative positives from the 1200 tested in Aim 3. These will be variants with high eleVAR scores and also scored positive in luciferase assays. Through the detailed validation experiments in this aim, we will not only further confirm the validity of our eleVAR pipeline, but also significantly improve our understanding of cancer.

D-4-a Preliminary results related to detailed validation

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

We have much experience with prostate cancer cohorts. Relevant to this our group recently performed a large scale profiling study for 2,000 individuals from the Tyrol Early Prostate Cancer Detection Program[105, 106]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. We are also involved in the Early Detection Research Network (EDRN)[107] prostate cancer cohort. This includes men enrolled at three sites as part of the Prostate Cancer Clinical Validation Center that prospectively enrolls individuals at risk for prostate cancer at Beth Israel Deaconess Medical Center (Harvard), at the University of Michigan (Michigan) and at Weill Cornell Medical College (Cornell). Cases are defined as men diagnosed with prostate cancer and controls are men who have undergone prostate needle biopsy without any detectable prostate cancer and no prior history of prostate cancer. Together, these two cohorts provide us with samples from thousands of prostate cancer patients and normal controls.

D-4-a-ii We have experience in the detailed validation of SNVs within enhancer elements

In order to study the potential role of inherited genetic variants within regulatory intergenic elements in the context of hormone dependent human tumors, we recently performed an unbiased computational search for AR/ERα bound enhancers elements containing SNVs followed by *in vitro* characterization of two selected variants [108]. After a series of filters, two were selected for *in vitro* characterization (on 1q21.3, rs2242193 MAF=0.038 and 13q34, rs9521825 MAF=0.235), here referred to as Locus 1 and Locus 2. Selected loci were cloned in pGL4.26 plasmid (plasmid with alternative allele was also generated) and then validated and characterized *in vitro* by luciferase assay with and without DHT treatment in MCF7 cells. Both constructs reached high responsiveness to DHT treatment hinting at their strong enhancer role. Moreover, the SNP variant on 1q21.3, rs2242193, demonstrated a role in the transcriptional regulation (p=0.028, Student's t-test) (**Figure XXX**). As both loci were confirmed as transcriptionally responsive to DHT by luciferase assay, we next opted for ChIP assays with AR antibody (or with normal IgG as a control) using MCF7 cells that are heterozygous at rs2242193 in Locus 1, but homozygous for the reference allele at Locus 2. Using quantitative PCR (qPCR), we were able to detect AR binding to both selected loci in MCF7 cells transiently over-expressing AR (**Figure XXX**). Occupancy levels at KLK3, KLK2 and TMPRSS2 were measured as positive controls (**Figure XXX**). Moreover, to assess whether AR showed allele-specific DNA binding at rs2242193, we amplified AR-enriched Locus 1 region by standard PCR followed by double-strand direct DNA sequencing analysis. Quantification of the electropherograms showed that the A allele was significantly enriched ($p<10^{-22}$, Fisher test) in chromatin fragments immunoprecipitated with antibody against AR compared to input genomic DNA (**Figure XXX**). Altogether, our results show that unbiased genome-wide search for polymorphic regulatory regions (PRRs) is an efficient methodology to discover new functional cis-elements relevant to hormone driven diseases and beyond by providing experimental evidence for selected variants mapping to regulatory regions.

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

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

D-4-b Approach to detailed validation

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

We will determine if any or all 10 variants selected based on successful validation in Aim 3 are associated with cancer or cancer causing characteristics. We will achieve this by studying the specific variant in test cohorts. We will use both the Tyrol cohort and the Early Detection Research Network (EDRN)[107] prostate cancer cohort with thousands of prostate cancer individuals as well as normal controls (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. Then, we will follow up for detailed functional screening, to be discussed below. For controls, we will utilize deeply sequenced control cohorts (individuals with no cancer) that are already available (see above). Superior allelic discrimination is achieved in these assays as they utilize TaqMan minor groove-binding (MGB) probes. This technique generates a low signal to noise ratio and affords a greater flexibility. The Taqman probes are functionally tested to first ensure assay amplification and optimization for amplification conditions.

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

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

D-4-b-ii-(1) Impact on gene expression: real-time quantitative PCR

Real-time quantitative PCR analysis of the genes downstream of the 10 selected variants will be performed on individuals that have been identified as recurrent for the variants and a similar sized group of non-recurrent individuals. We will look for perturbed gene expression in the target genes. This analysis will inform us if a SNP (in promoter or enhancer regions) has any effect on transcription of the target gene. Recurrent rare SNPs will be further validated by PCR assays using primers that can amplify the genomic region encompassing the SNP. PCR will be followed by direct sequencing of the amplicon using an ABI 3730 DNA Sequence Analyzer on a subset of tumor-normal pairs to verify the individual promoter/enhancer mutations for further confirmation.

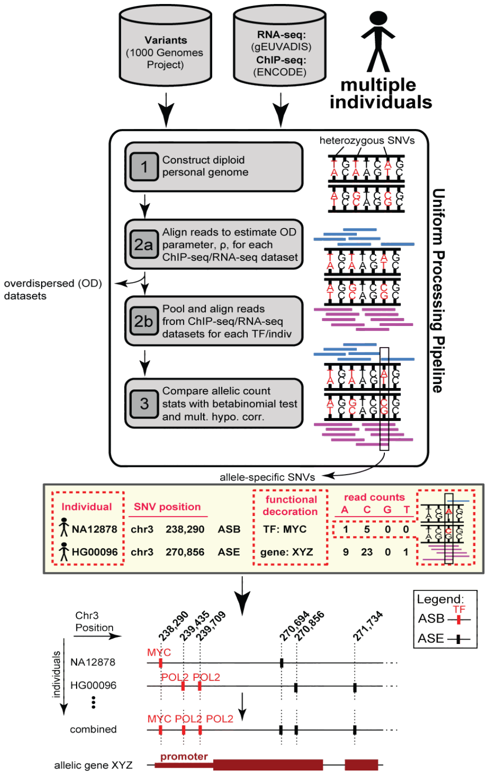
D-4-b-ii-(2) Functional consequences: CRISPR/CAS system

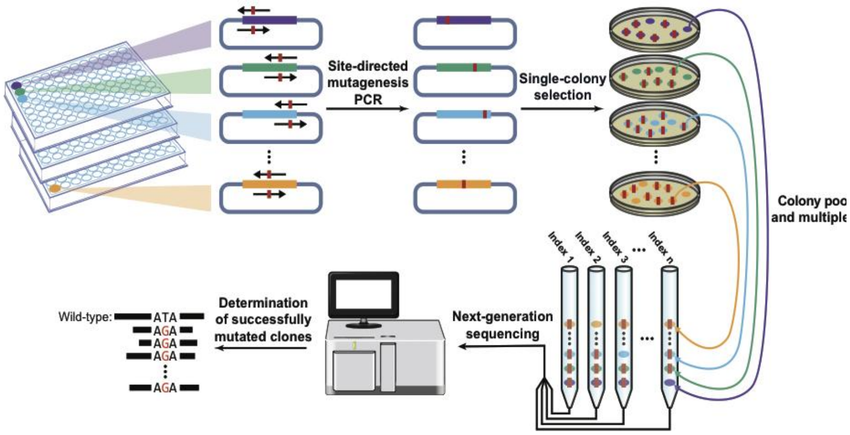
We will utilize the newly discovered CRISPR/CAS system[109] to generate endogenous mutations in TF binding site in a panel of prostate cancer cell lines (VCaP, LnCaP, DU145 and PC3). This unique system will provide us an opportunity to directly modulate endogenous genes and minimize artifacts due to the transfection based reporter assays. Using CRISPR/CAS mediated genome-engineering method[110] we will directly generate mutations within promoter/enhancers of target genes. Theoretically we will generate 10 individual SNPs in each cell line and will study functional relevance of these changes compared to WT. In case of rare mutations, which occur within both promoter and enhancer regions of the same gene, we will develop cell lines having these combinatorial mutations. Mutations within regulatory regions like promoter and enhancer regions might contribute to one or more biological effects as described in the schematic (Fig. XXX).

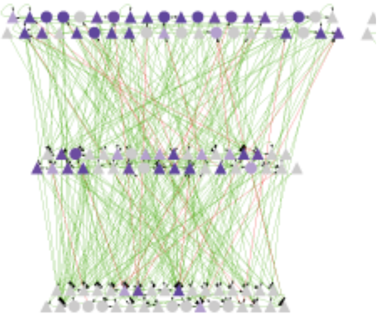
The mutant and WT cell lines generated using CRISPR/CAS system will be monitored for a) phenotypic changes by confocal microscopy and actin staining to determine effects of mutation on cytoskeletal reorganization b) Influence on proliferation by MTT and CellTiter-Glo® Luminescent Cell Viability Assay (Promega) c) Influence on invasive and migratory potential using, matrigel coated invasion and boyden chambers in 24 well format d) senescence by β-gal staining e) apoptosis by tunnel assay.

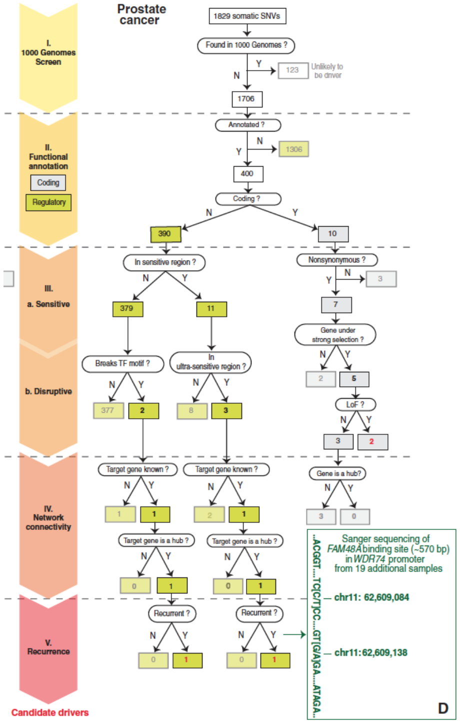
D-4-b-ii-(3) Effect of the mutation on TF binding

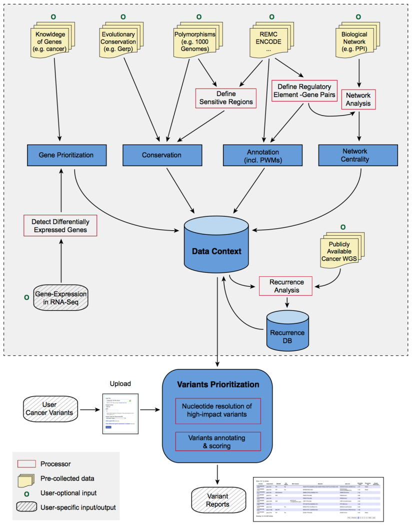
In vitro EMSAs will confirm specific binding to WT or mutant sequence by a particular transcription factor. EMSA (electrophoretic mobility shift assay) is a common technique employed to study protein-DNA interactions. We will use the WT and the MT sequences to determine binding of the transcription factor predicted to be present at the site of mutation. Chromatin immuno-precipitation (ChIP) assays for TFs overlapping the variant will be conducted to determine if the variant can distort TF binding in vivo. This would help validate the variants that are predicted to be motif breakers. Alternatively, for the SNVs predicted to create a new motif, ChIP experiments will help validate binding.


Variants
(1000 Genomes
Project)
RNA-seq:
(gEUVADIS)
ChIP-seq:
(ENCODE)
multiple
individuals
heterozygous SNVs
1 Construct diploid personal genome
2a Align reads to estimate OD parameter, p, for each ChIP-seq/RNA-seq dataset
overdispersed (OD) datasets
2b Pool and align reads from ChIP-seq/RNA-seq datasets for each TF/indiv
3 Compare allelic count stats with betabinomial test and mult. hypo. corr.
Uniform Processing Pipeline
allele-specific SNVs
Individual
SNV position
functional decoration
read counts
A C G T
NA12878 chr3 238,290 ASB TF: MYC 1 5 0 0
HG00096 chr3 270,856 ASE gene: XYZ 9 23 0 1
Chr3
Position
238,290
239,435
239,709
270,694
270,856
271,734
Legend:
ASB
TF
ASE
individuals
NA12878
HG00096
combined
MYC
POL2
POL2
MYC POL2 POL2
promoter
allelic gene XYZ


Site-directed
mutagenesis
PCR
Single-colony
selection
Colony poo
and multiple
Index 1
Index 2
Index 3
Index n
Next-generation
sequencing
Determination
of successfully
mutated clones
Wild-type: ATA
AGA
AGA
AGA
AGA




Prostate
cancer
I.
1000 Genomes
Screen
II.
Functional
annotation
Coding
Regulatory
III.
a. Sensitive
b. Disruptive
IV.
Network
connectivity
V.
Recurrence
Candidate drivers
1829 somatic SNVs
Found in 1000 Genomes ?
Y
N
123
Unlikely to be driver
1706
Annotated ?
N
Y
1306
400
Coding ?
N
Y
390
10
In sensitive region ?
N
Y
379
11
Breaks TF motif ?
N
Y
377
2
In ultra-sensitive region ?
N
Y
8
3
Nonsynonymous ?
N
Y
3
7
Gene under strong selection ?
N
Y
2
5
LoF ?
N
Y
3
2
Target gene known ?
N
Y
1
1
Target gene known ?
N
Y
2
1
Gene is a hub?
N
Y
3
0
Target gene is a hub ?
N
Y
0
1
Target gene is a hub ?
N
Y
0
1
Recurrent ?
N
Y
0
1
Recurrent ?
N
Y
0
1
Sanger sequencing of FAM48A binding site (~570 bp) in WDR74 promoter from 19 additional samples
..ACGGT....TC[C/T]CC....GT[G/A]GA....ATAGA..
chr11: 62,609,084
chr11:62,609,138
D


O
Knowledge of Genes (e.g. cancer)
O
Evolutionary Conservation (e.g. Gerp)
O
Polymorphisms (e.g. 1000 Genomes)
O
REMC ENCODE ...
O
Biological Network (e.g. PPI)
Define Sensitive Regions
Define Regulatory Element -Gene Pairs
Network Analysis
Gene Prioritization
Conservation
Annotation (incl. PWMs)
Network Centrality
Detect Differentially Expressed Genes
O
Gene-Expression in RNA-Seq
Data Context
O
Publicly Available Cancer WGS
Recurrence Analysis
Recurrence DB
Variants Prioritization
Nucleotide resolution of high-impact variants
Variants annotating & scoring
User Cancer Variants
Upload
Variant Reports
Processor
Pre-collected data
O User-optional input
User-specific input/output

Year 1: The newly developed computational framework for rare variant prioritization, *eleVAR*, will score each targeted disease-related variant by integrating the weights of individual features associated with it.

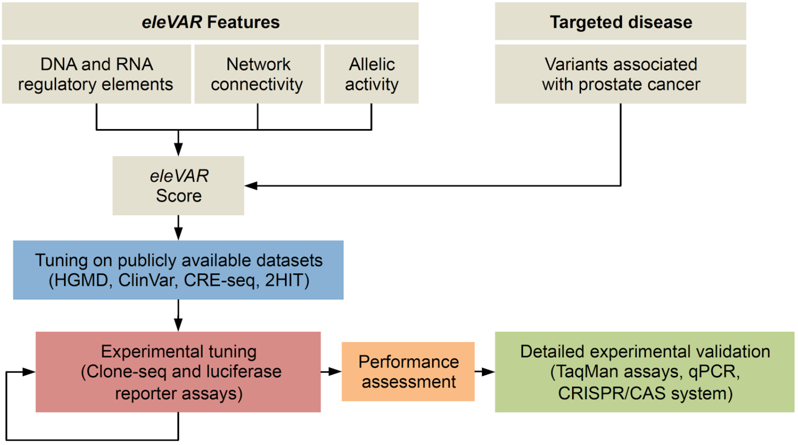
The **initial parameter tuning and performance assessment** will be carried out using Bayesian update approach utilizing publicly accessible validated variants.

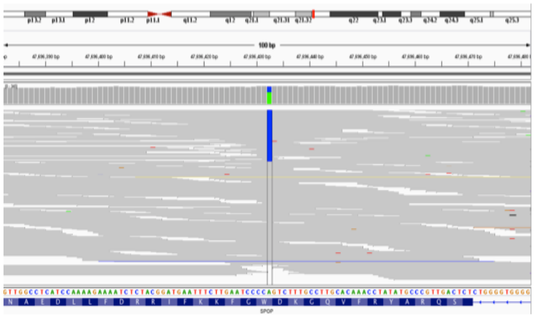
The **first round of experimental tuning** and validation using Clone-seq and luciferase reporter assay will be performed to tune the weights of the *eleVAR* parameters.

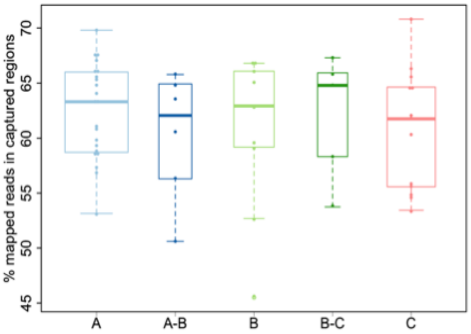
Year 2: The **second round** of experimental refinement will be performed to fine-tune the parameters.

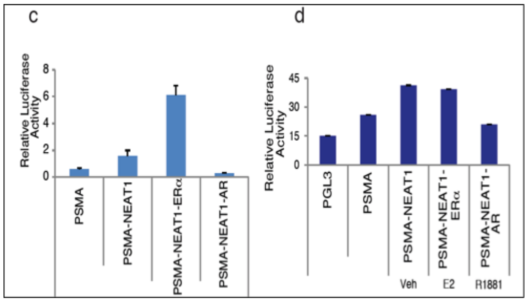
Year 3: The **third round** will be used for **assessing performance of the model**.

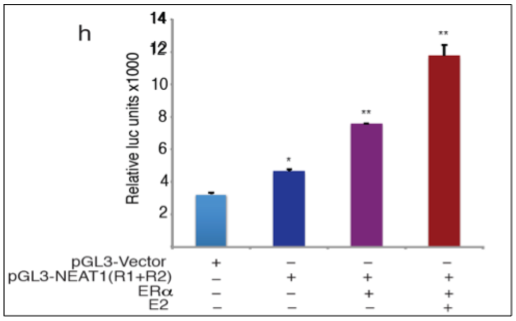
In-depth experimental validation and an investigation of the molecular basis of the observed impact will be performed for top scored variants.


eleVAR Features
DNA and RNA regulatory elements
Network connectivity
Allelic activity
Targeted disease
Variants associated with prostate cancer
eleVAR Score
Tuning on publicly available datasets (HGMD, ClinVar, CRE-seq, 2HIT)
Experimental tuning (Clone-seq and luciferase reporter assays)
Performance assessment
Detailed experimental validation (TaqMan assays, qPCR, CRISPR/CAS system)

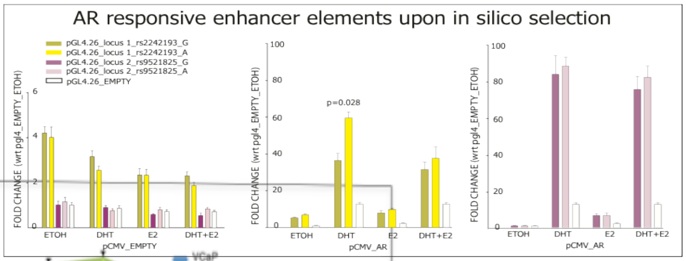

100 bp
SPOP

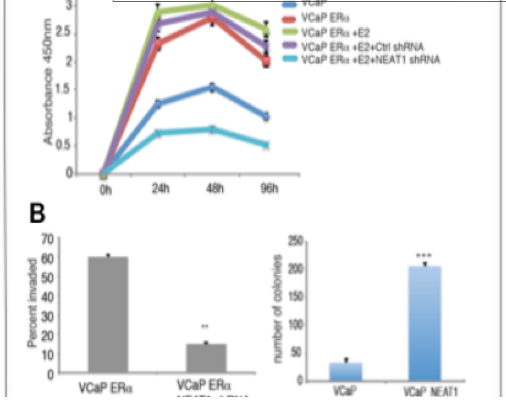

% mapped reads in captured regions
45
50
55
60
65
70
A
A-B
B
B-C
C


c
d
Relative Luciferase Activity
0
2
4
6
8
PSMA
PSMA-NEAT1
PSMA-NEAT1-ERα
PSMA-NEAT1-AR
Relative Luciferase Activity
0
15
30
45
PGL3
PSMA
PSMA-NEAT1
PSMA-NEAT1-ERα
PSMA-NEAT1-AR
Veh
E2
R1881


h
Relative luc units x1000
2
4
6
8
10
12
14
*
**
**
pGL3-Vector + – – –
pGL3-NEAT1(R1+R2) – + + +
ERα – – + +
E2 – – – +

(C).


AR responsive enhancer elements upon in silico selection
pGL4.26_locus 1_rs2242193_G
pGL4.26_locus 1_rs2242193_A
pGL4.26_locus 2_rs9521825_G
pGL4.26_locus 2_rs9521825_A
pGL4.26_EMPTY
FOLD CHANGE (wrt pgl4_EMPTY_ETOH)
0
2
4
6
ETOH
DHT
E2
DHT+E2
pCMV_EMPTY
0
20
40
60
80
100
p=0.028
ETOH
DHT
E2
DHT+E2
pCMV_AR
0
20
40
60
80
100
ETOH
DHT
E2
DHT+E2
pCMV_AR


A
Absorbance 450nm
0
0.5
1
1.5
2
2.5
3
3.5
0h
24h
48h
96h
VCaP
VCaP ERα
VCaP ERα +E2
VCaP ERα +E2+Ctrl shRNA
VCaP ERα +E2+NEAT1 shRNA
B
Percent invaded
0
10
20
30
40
50
60
70
**
VCaP ERα
VCaP ERα
number of colonies
0
50
100
150
200
250
***
VCaP
VCaP NEAT1