B. Significance

B-1 Non-coding variants are significant for disease but less well-studied than coding ones

Numerous studies have been conducted on the mutations that lie in coding regions[5-8]. Not as much has been done on non-coding ones. However, several initial studies suggest that variants in non-coding regions can significantly influence an organism's phenotype[9], and they are often implicated in diseases[10,11]. Many non-coding variants impact regulatory elements. Such variation in the human genome can modulate gene expression[12], and changes in this expression have been implicated in cancer and other diseases[13-18].

B-2 Rare variants are important for disease but have received less attention than common ones

There have been a large number of GWAS studies[19-23], 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[24]. Increased disease susceptibility is often attributed to the cumulative effect produced by multiple rare variants[25] – e.g., rare germline variants in the CHEK2 and HBOX genes were associated with breast and prostate cancer, respectively[26, 27].

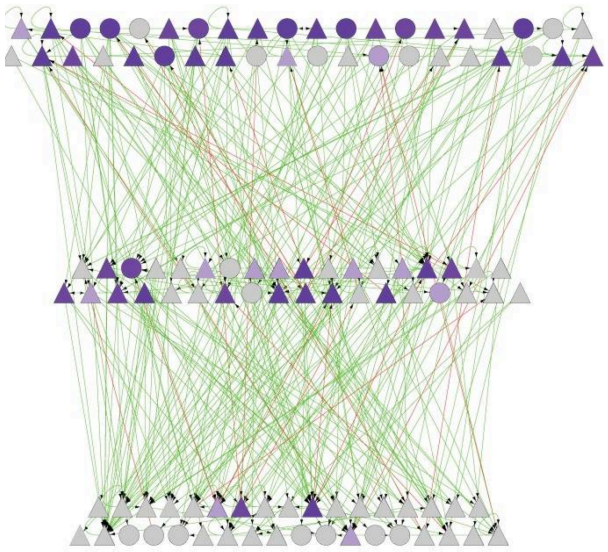

Fig 1: ENCODE Regulatory Network Analysis

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

Annotating non-coding regions is essential for investigating genome evolution[28], understanding important biological functions (including gene regulation and RNA processing)[29], and for elucidating how SNPs and structural variation may influence disease[30]. The Encyclopedia of DNA Elements (ENCODE) and the model organism ENCODE (modENCODE) Project provide extensive comparative genomic annotation of human, mouse, fly and worm genomes[31-33]. Furthermore, regulatory variations in the human genome have been investigated by large-scale mRNA and miRNA sequencing[34-37]. Recently, large-scale efforts (e.g., the Epigenome Roadmap and GTEx projects) have also been directed toward annotating human epigenomic data[38-42], as well as understanding the influence of genomic variation on the gene expression profiles[43-47] . These Expression Quantitative Trait Loci (eQTL) can further be utilized to investigate disease mechanisms[48].

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 in 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 Interpreting the impact of rare non-coding variants, consistently for TF binding & ncRNAs, using population-scale polymorphism data

While common variants associated with disease are contained within the GWAS catalog, very few studies attempt to identify rare disease-associated variants. Currently, no standard methods exist to functionally interpret such variants, especially in non-coding regions. Herein, we propose a consistent approach to prioritize rare variants associated with disease over all the non-coding regions in the genome, i.e, rare variants in the regulatory regions and ncRNAs will be prioritized using a consistent scoring scheme that utilizes the natural polymorphism data within healthy humans.

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

Previous studies have identified specific variants with allele-specific activity[36, 49]. However, there has not been a scheme that allows us to prioritize variants based on this, 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.

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

An innovative aspect of this proposal is a parameter weighting scheme for variant prioritization and iterative tuning of it. 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 score based on the weights 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: massively-parallel site-directed mutagenesis pipeline leveraging next-gen. 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]. Clone-seq is entirely different from previously described random mutagenesis approaches[50-53]: 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 past experience in non-coding annotation, as part of our 10-year history with the ENCODE and modENCODE projects. Our TF work includes the developing new machine learning techniques to define the binding peaks of TFs and predict TF target genes[54-56]. Furthermore, we developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers[57], which we have partially validated[58]. We also constructed linear and non-linear models that utilize TF binding and histone modification signals to accurately predict the transcriptional output of a gene in different cell types of several organisms including yeast, worm, fly, and human[33, 59-62]. We have also constructed regulatory networks for human and model organisms[63, 64], and completed many analyses on them (Fig 1)[33, 58, 63, 65-77]. Furthermore, we have conducted large-scale multi-organism regulatory and co-expression network comparisons, along with transcriptome and pseudogene lineage analyses[77-81]. Finally, we have experience conducting integrated analyses of RNA-Seq datasets generated by the ENCODE, modENCODE, BrainSpan and exRNA consortia[31, 33, 82-84]. In particular, we developed RSEQtools and IQseq for gene model creation and transcript quantification[85, 86]. We also developed tools that specifically analyze features of ncRNAs, including incRNA and ncVAR for finding and characterizing these elements[87, 88].

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

A specific class of regulatory variants are those associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and with allele-specific expression (ASE)[89, 90]. We have previously developed a tool, AlleleSeq[76], for the detection of candidate variants associated with ASB and ASE. Using this we have generated comprehensive lists of allelic variants for ENCODE and 1000 Genomes and found that allelic variants are under differential selection from non-allelic ones[63, 63, 74, 82]. 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[63]. Finally, we have constructed a personal diploid genome and transcriptome of NA12878 [91].

D-1-a-iii We have experience in relating annotation to variation: the FunSeq pipeline

We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets[58,63,88]. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations[88]. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region[31]. Further studies showed relationships between selection and protein network topology (eg, quantifying selection in hubs relative to proteins on the network periphery[73, 75]). In recent studies[4, 74], we have integrated and extended these methods to develop a prioritization pipeline called FunSeq (Fig 2). It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). FunSeq links each non-coding mutation to target genes, and prioritizes such variants based on scaled network connectivity. 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.

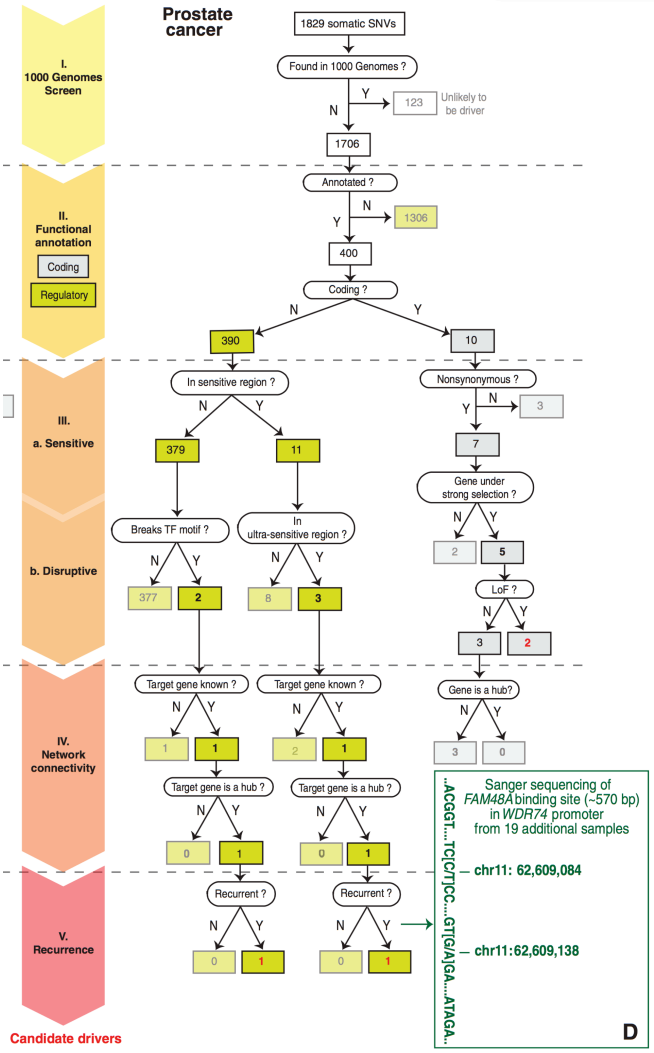

Fig 2: Filtering of somatic variants from a prostate cancer sample leading to identification of candidate drivers

It also detects their disruptiveness in TF binding sites (both loss-of and gain-of function events). FunSeq was developed as part of the 1000 Genomes Functional Interpretation Group (FIG) and represents a collaboration between that group and that of a cancer genomics research (Dr Rubin). In particular, integrating large-scale data from various resources (including ENCODE and 1000 Genomes) with cancer data, FunSeq is able to prioritize the known TERT promoter driver mutations, and it scores somatic recurrent mutations higher than those that are non-recurrent. Moreover, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples [74]. We have also applied our method to investigate non-coding mutation patterns in subtypes of gastric cancer[92]. Drawing on this experience, we are currently co-leading ICGC PCAWG-2 (Pan-cancer Analysis Working Group #2: Analysis of Mutations in Regulatory Regions).

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 (Fig 3). Our new pipeline is called eleVar. It will have several key features: 1) identifying functional sites among the conserved regions of the human genome and ncRNA regulatory elements; 2) investigating the allelic elements; and 3) taking into account network connectivity.

D-1-b-i Consistently prioritizing non-coding elements from 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 the GERP score[93]).

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[94-96], and more recently from ENCODE. In particular, we will develop a new machine-learning framework that utilizes pattern recognition within the signal of various epigenomic features and the transcription of enhancer RNAs (eRNAs) to predict active enhancers across different tissues.

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 publicly available data, such as CLIP-Seq, CLASH and computational predictions (TargetScan) to create a compendium of biochemical interactions with RNA[97-101]. Our initial analyses indicate that some binding sites are even more sensitive to variation than coding sequences. In addition, we will incorporate aspects of RNA 3D-structure. 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 cause a larger free energy change in terms of structure are rarer in human populations. We will define sensitive regions based on folding free energy and folding z-score cutoffs that are enriched for rare genetic variants.

D-1-b-ii Identifying high-impact mutations: breaking & creating motifs

For impactful events at TF binding sites, we will use motif breakers and formers to define loss-of- and gain-of-function events, respectively, as these events are more likely to have deleterious consequences[14, 15, 74, 88, 102-104]. 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. A key improvement is to employ ancestral alleles to get a more accurate determination of these events.

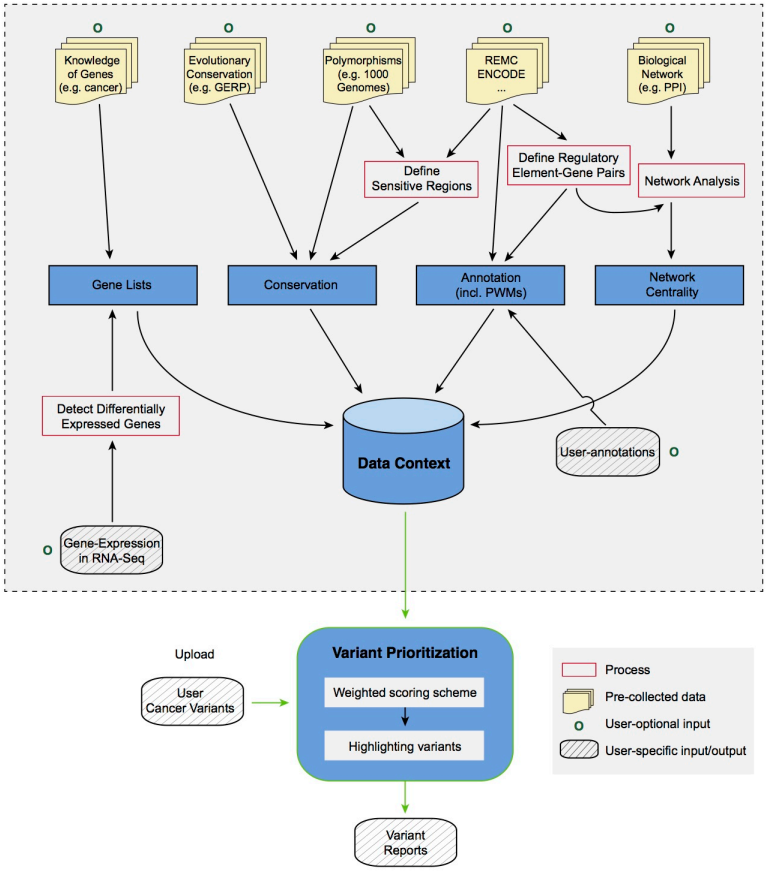

Fig 3: Description of eleVAR workflow & data context

In a way that is consistent with our means of searching for motif-breaking variants in TF binding sites, we will identify motif-breakers in specific RNA binding motifs. Studies of RNA processing and function have identified key motifs associated with events ranging from RNA splicing to chemical RNA base modifications[105]. 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. 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.

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

Allele-specific variants potentially provide a most direct readout of the functional impact of a variant. For example, if we can associate the differential binding effect of a particular transcription factor with different alleles, then we can identify loci that have potential functional impacts in regulation. However, because allelic variants are enriched for rare variants[34], it will be difficult to match the specific variants in a personal genome of interest to prioritize 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 (Fig 4).

We derive allelic elements by first identifying allelic variants from hundreds of individuals. These individuals will be amassed from The 1000 Genomes Project[106]. We will match them with their corresponding RNA-Seq and ChIP-seq experiments from multiple disparate studies, such as gEUVADIS[34] and ENCODE[31]. Because these separate studies typically have 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[107-109]. To identify and remove problematic "outlier" datasets 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 parametrizes 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 (in comparison to accessible variants within the elements and having sufficient coverage to make an allelic activity call), but also across the number of individuals having allelic variants in a consistent allelic direction. The scoring system by element is useful in two ways: (1) it allows continuous ranking of genomic elements based on its allelic impact across multiple individuals (as opposed to defining a threshold to make a binary decision of whether an element is ‘allelic’) and (2) it enables incorporation of ASE and ASB into the main prioritization scheme; input variants (even those which are rare, but lie in highly-ranked allelic genomic elements) will be up-weighted according to their scores.

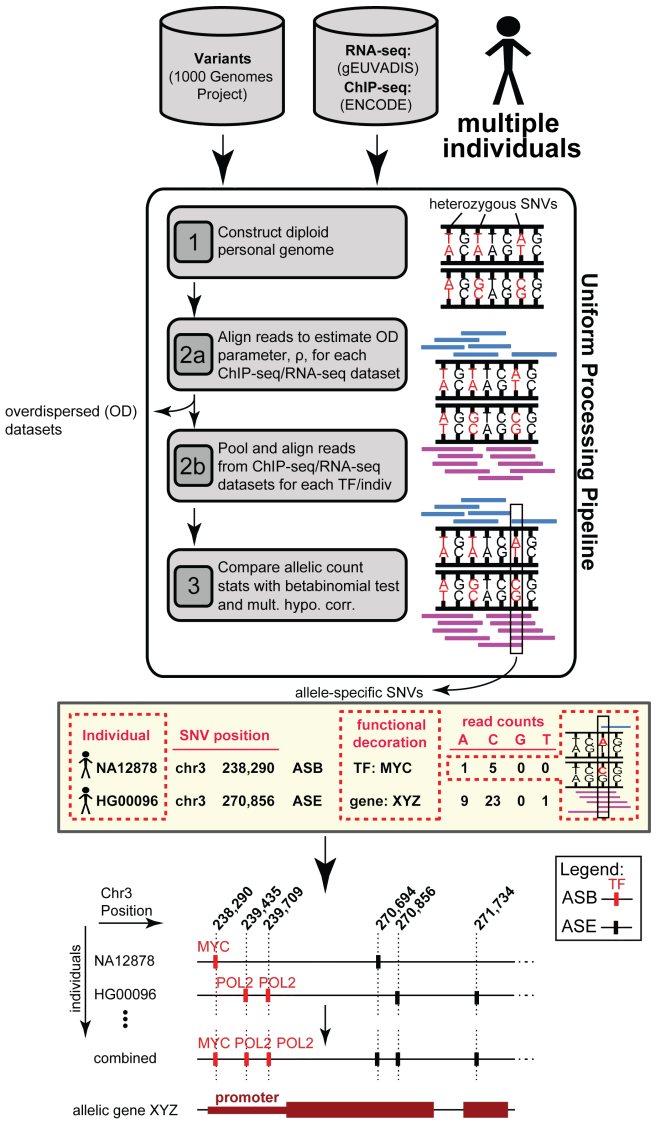

Fig. 4: Workflow for generating allelic variants and elements

D-1-b-iv Identifying likely target genes for distal regulatory elements & 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 by Yip et al[57]. A single regulatory variant may affect the expression of multiple genes, either because it directly them 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 (eg hubs, bottlenecks and hierarchy tops), as we know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious[73,75]. 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 RNAs that are bound by many proteins.

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 above, we plan to elaborate the weighting system in FunSeq.[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 weight these two sets of features with different strategies.

For each discrete feature d, we calculate the probability p_d that it overlaps with common polymorphisms. We then calculate the information content to denote its weighted value $w_d = 1 + p_d * log_2 p_d + (1 - p_d) * log_2(1 - p_d)$.

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 v_c, we will calculate feature weights for each v_c. In particular, we discretize at each value and compute $w_c^{v_c}$. Then we fit a smooth curve for all v_c to obtain continuous $w_c^{v_c} = 1 + p_c^{\geq v_c} * log_2 p_c^{\geq v_c} + (1 - p_c^{\geq v_c}) * log_2(1 - p_c^{\geq v_c})$. When we evaluate the continuous feature for a particular variant, we calculate its weighted value using the fitted function.

We score each variant by summing up the weighted values of all its features $s = \sum_d w_d + \sum_c w_c^{v_c}$. We will also consider the feature dependency structure when calculating the scores (e.g., removing redundant features or performing dimension reduction techniques).

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

D-2-a Overall workflow for the project in relation to Aim 2

As shown in the timeline (Fig 6), we will take our features and weighting scheme (from Aim 1) and construct a practical software pipeline that can be applied on many genomic variants in a high-throughput fashion. We will then collect genomic variants from the existing cancer genomics data. We will run the pipeline on these

variants to 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 high-throughput experimental characterization. Finally, we will perform an unbiased testing and pick a number of variants for in-depth validation.

D-2-b Research plan for Aim 2

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

The initial feature weights $\boldsymbol{W}$ $(w_1, w_2, \dots, w_m)$ (given m number of features) 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. For a variant v, given feature values $\boldsymbol{F}_v$ $(f_{v,1}, f_{v,2}, \dots, f_{v,m})$, $\boldsymbol{W}$ can be rewritten as $(t_1(f_{v,1}), t_2(f_{v,2}), \dots, t_m(f_{v,m}))$, where functions $\boldsymbol{T}$ depict the relationship between $\boldsymbol{W}$ and $\boldsymbol{F}$. This could, for instance, be a simple linear relationship with a single proportionality parameter or a more complex non-linear relationship with multiple parameters. In any event, all parameters in $\boldsymbol{T}$ are the same for all variants. Given the eleVAR score s (equation 3 in D-1-b-v), the probability that v is functional ($y_v = 1$ designates a positive result, whereas $y_v = 0$ denotes a negative result) follows a logistic function $P(y_v = 1|s) = \frac{1}{1 + exp(-k * (s-a))}$ (k, a are scaling parameters). To update $\boldsymbol{W}$ (more specifically, the parameters in functions $\boldsymbol{T}$) with training data $\boldsymbol{Y}$, we implement Bayes' rule: $P(\boldsymbol{T}|\boldsymbol{Y}, \boldsymbol{F_V}) \propto P(\boldsymbol{Y}|\boldsymbol{T}, \boldsymbol{F_V})P(\boldsymbol{T})$. The probability of observing $\boldsymbol{T}$ (given $\boldsymbol{Y}$ and feature values $\boldsymbol{F_V}$ corresponding to variants in $\boldsymbol{Y}$) is proportional to the probability of observing $\boldsymbol{Y}$ given $\boldsymbol{T}$ and $\boldsymbol{F_V}$, multiplied by the prior probability of $\boldsymbol{T}$. Assuming independency between data points in $\boldsymbol{Y}$, which can be achieved by proper training data construction,
$P(\boldsymbol{Y}|\boldsymbol{T}, \boldsymbol{F_V})P(\boldsymbol{T}) = \prod_{i=1}^{n} P(y_i|t_1, t_2, \dots, t_m, f_{i,1}, f_{i,2}, \dots, f_{i,m})P(t_1, t_2, \dots, t_m)$, given n observations in $\boldsymbol{Y}$.

Using the training data, we will maximize this function to find the most probable functions $\boldsymbol{T}$, and these will be used as our updated parameters. The updated $\boldsymbol{T}$ will then be used as tuned parameters in eleVAR to prioritize variants. The procedure will be iterated in several rounds. In the first round of tuning, feature weights obtained in D-1-b-v will be used to construct priors $P(\boldsymbol{T})$. In subsequent rounds, the updated weights will be set as new priors.

D-2-b-ii Software implementation using an explicit data context & dependency graph

We will develop an efficient, robust and yet flexible software suite for eleVAR for users to parameterize and customize for their own research projects. 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 analyze and optimize data flow in our model, aiming to eliminate data dependencies and to modularize the calculating process. We will recognize critical inter-procedural 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.

We will host our software 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, aiming to distribute the software to the entire research community and ensure the reproducibility of our results. Finally, from our planned project website (elevar.gersteinlab.org) we will also make available the results of all the validation experiments (described below), so users can re-tune the eleVAR parameters as they want.

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

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[110], which is run by the Broad Institute, and Caveman[111], which is run by the Wellcome Trust Sanger Institute. These call sets will then be integrated with the tool FreeBayes[92], which will generate the final call set for further downstream analyses. As we will be focusing on prostate cancer, we will add to this list a number of other whole-genome sequences of

prostate cancers (tumor and normal)[112, 113]. 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[114]. In particular, 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).

D-2-b-iv Round 1 of tuning based on publicly available datasets

To perform the initial round of performance assessment and parameter tuning, we plan to use publicly available datasets from various resources. These datasets include known disease-causing mutations from molecular studies, high-throughput reporter assays on enhancer activities and recurrence of cancer rare mutations in the region of interest involving germline and potentially somatic variants.

The Human Gene Mutation Database (HGMD)[115] and ClinVar[116] catalogue large numbers of regulatory disease-causing mutations discovered in molecular studies. Several high-throughput technologies have also been developed to test the phenotypic impacts of non-coding genomic variants. For example, Kwasnieski et al used CRE-seq[117] to assay over 1,000 single- and double-nucleotide mutations in promoter regions. Kheradpour et al[102] used MPRA to test variants affecting regulatory motifs in over 2,000 human enhancers. We will utilize these datasets to perform comparisons with other variant prioritization methods, such as CADD[118], to obtain a preliminary evaluation of method performance. We will then tune our parameters using the scheme described above.

We will further compare the germline mutation burden of healthy individuals with those suffering from cancer. Specifically, we will use over 2,500 normal samples from The 1000 Genomes Project as the control data, and run a mutation burden test using available software such as SKAT[119]. (If it is necessary to expand the controls for rarer variants, we could use deeply sequenced trios from the 1000 Genomes Project[120], 500 individuals with Complete Genomics sequencing also from 1000 Genomes[121] and healthy individual from the UK10K project[122].) In contrast to the binning process generally used in burden testing, 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 up-weighted 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 non-coding regulatory elements.

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[15, 123, 124]). In our study, we will also analyze the somatic mutation burden in our feature list. Different from the germline mutation burden test, our computational framework is used to directly evaluate the somatic mutation burden in cancer samples. It incorporates a comprehensive list of confounding covariates, which includes replication timing, histone modification marks, chromosome accessibility, and GC content, to precisely calculate a local background mutation rate for somatic burden evaluation. Accordingly, it provides a list of heavily mutated non-coding 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 upweighted in eleVAR.

D-2-b-v Round 2 of tuning using high-throughput experiments done in this project

Based on the results from 1000 Genomes, we expect ~40K rare germline variants per genome[106]. Since they rarely recur at the exact same position, we anticipate a prioritized list of ~8M variants (=40K * 250 genomes, based on the 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 expected number of variants per element will be ~4. Variants on the same element will 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 lower 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. We will

validate these variants through functional genomic screens using the Clone-seq technology coupled with high-throughput luciferase reporter assays. Overall, this refinement will be accomplished in two rounds, one round per year, as detailed in Aim 3 and the timeline (Fig 6). Finally, during the last year of the proposed work, 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 quantitatively evaluate eleVAR.

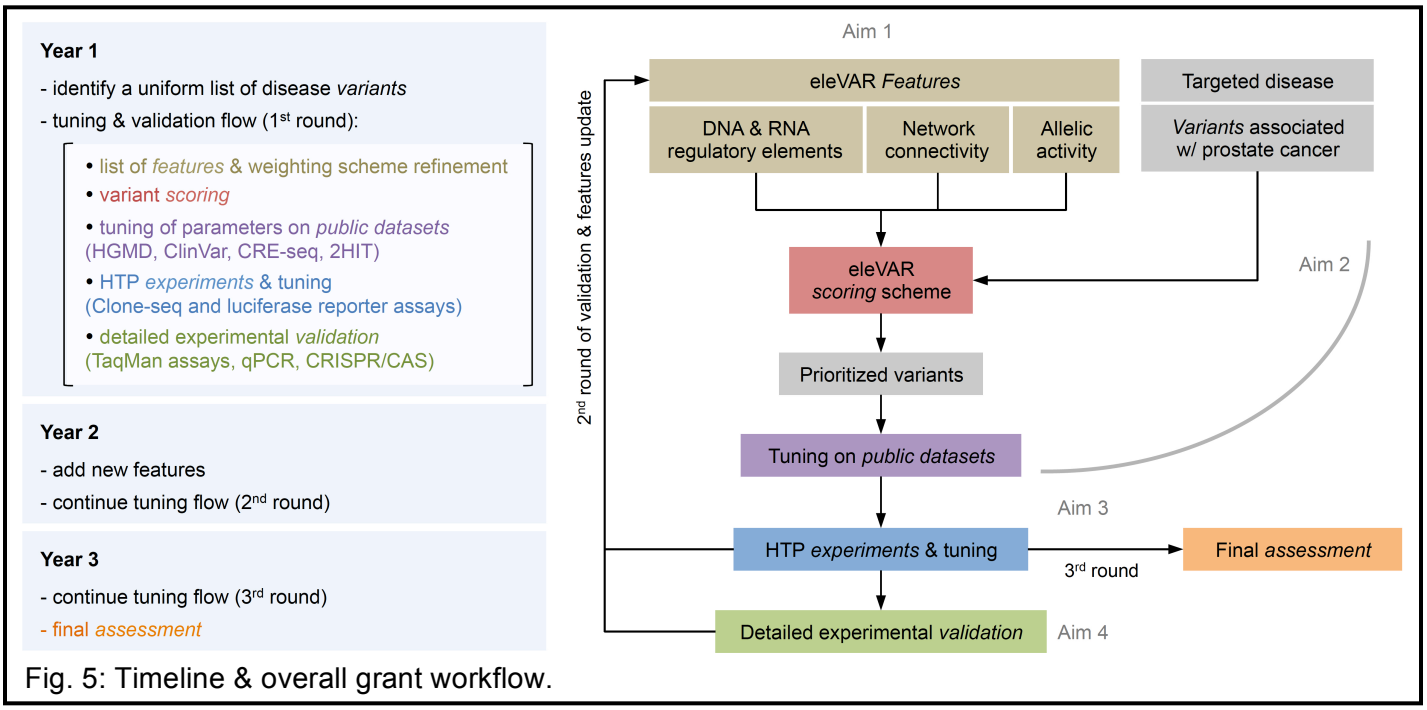

Fig. 5: Timeline & overall grant workflow.

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 1,200 SNVs in 600 regulatory elements to experimentally characterize their impact on gene regulation to fine tune and validate the 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 (Fig 5), we attempted to generate clones for 1,034 mutations on 223 genes, including 40 mutations for *MLH1*. We picked 4 colonies for each mutation (4,106 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[106, 125]. The average coverage of these 1,034 alleles is > 300×. Therefore, our Clone-seq pipeline has the capacity to generate > 3,000 mutations in one full lane of a HiSeq run, drastically improving the throughput and decreasing overall sequencing costs by at least 10-fold[1].

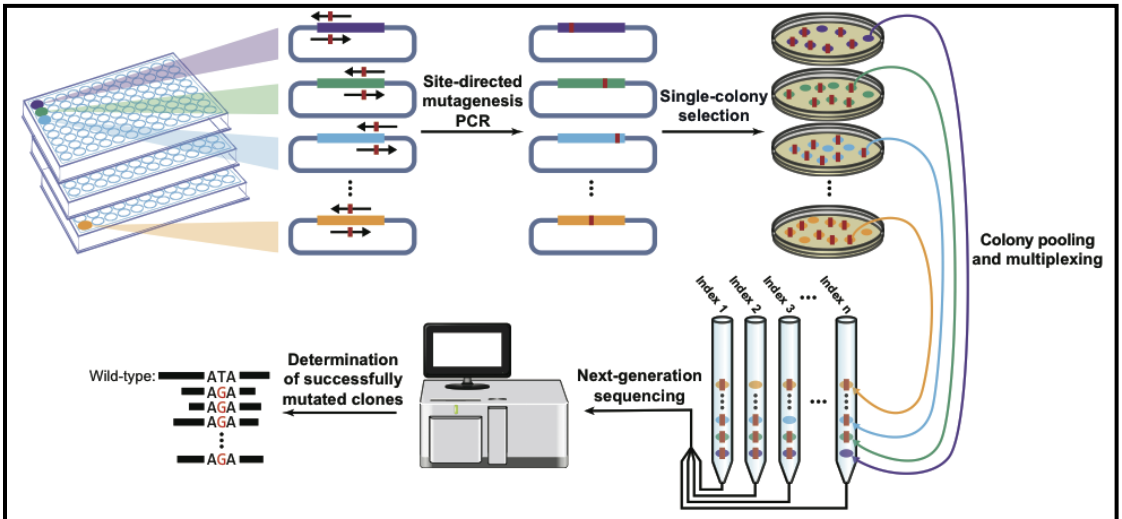

Fig. 6: Our massively parallel Clone-seq pipeline.

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[126]. 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 long genes.

In total, we have used the Clone-seq pipeline to successfully generate 1,034 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 SNVs as proposed.

D-3-a-ii Experience with luciferase reporter assays confirming validity of predicted TF binding sites

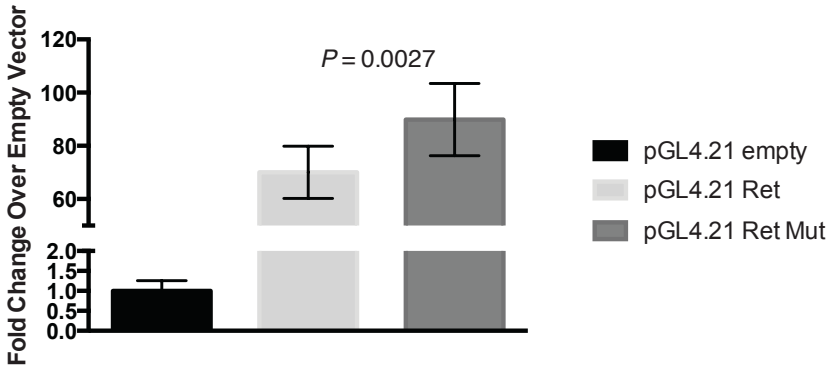

Fig. 7: Identification of a somatic mutation resulting in a motif gain in the Ret promoter by FunSeq. Luciferase reporter assays demonstrating transcriptional activity of the WT Ret promoter (-440 to +65) as compared to the same promoter containing the A291C mutation in DU145 cells. Results shown are the average of three independent experiments ±SD, and statistical significance was determined by t-test.

We have a great amount of experience with developing reporter assays for TF binding [127, 128, 129]. In particular, we have done an earlier study where we have used luciferase reporter assays to demonstrate the transcriptional regulation of several prostate cancer genes by ERα and the long non-coding RNA, NEAT1[127]. We have also done validation for the FunSeq prototype pipeline through collaborations among Gerstein, Yu, and Rubin groups. This was an outgrowth of the FunSeq development work that was part of the 1000 Genomes FIG group (see above). It 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 7).

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 or promoter elements and two variants on each element 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 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 regulatory elements

WT enhancer or promoter elements will be amplified using human genomic DNA as template with forward and reverse sequence-specific primers that are combined with attB1 and attB2 sequences, respectively[130]. We will perform large-scale Gateway BP reactions to clone each PCR product into a pDONR223 vector. High-throughput *E. coli* transformation will be carried out with 5 μL of BP reaction products using the Tecan robot. The cells are then spread out in plates through vigorous shaking with glass beads. The plates are incubated overnight at 37 °C. The next day, four colonies per allele are picked for Illumina sequencing.

D-3-b-i-(2) Illumina library preparation & HiSeq sequencing

E. coli cells for all four colonies of all WT alleles are individually cultured in 96-well deepwell plates overnight 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 ~1,200 mutant elements using Clone-seq

Primers for site-directed mutagenesis are designed by our automated web tool[131]. 50 µL mutagenesis PCR reactions are set up on ice in 96-well PCR plates using Phusion polymerase. *DpnI*-digested PCR products are used for *E. coli* transformation (see above). The next day, four colonies per allele are picked for sequencing.

D-3-b-i-(4) Functional consequences evaluated by high-throughput luciferase reporter assays

Reporter assays that employ either luciferase or next-generation reporter vectors can provide direct insight into the functional relevance of SNPs on target genes. We use a Gateway-compatible version of the firefly luciferase reporter vector, pGL4.23-GW (Addgene 60323). All WT and mutant constructs will be cloned 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 cells will be seeded in 96-well plates and transfected with WT and mutant enhancer/promoter constructs. 48 hrs after transfection, element activity will be measured following the manufacturer's instructions (Promega E2940). Assay values will be normalized using internal renilla luciferase as a 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

We strive to examine in detail 6 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 CRISPR/Cas-9 for *in vivo* experiments. Then we describe how we will carry this out for the 6 variants culled. We will choose 6 representative variants from positives of the 1,200 tested in Aim 3. These will be variants with high eleVAR scores that 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

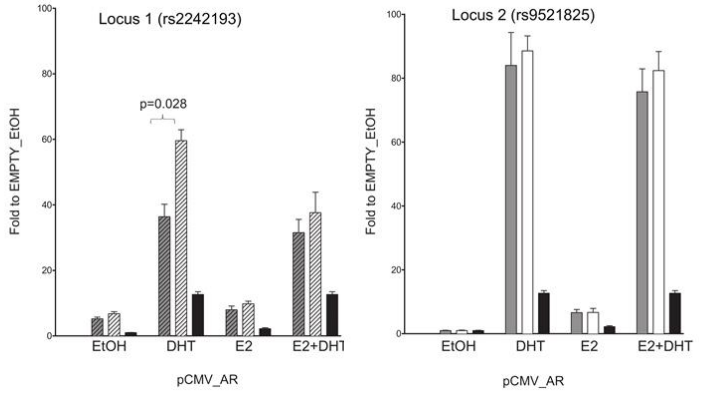

Fig. 8: The selected locus 1 and locus 2 regulatory regions act as AR-dependent enhancers in the MCF7 cell line. MCF7 cells were co-transfected with pCMV AR vector along with pGL4.26 locus 1 or locus 2 reporters. 24hr post-transfection cells were treated for 16 hrs with E2, DHT or the combination of the 2 compounds to stimulate respectively ER- or AR-dependent transcription. WT and SNPs rs2242193 or rs9521825 (both, dashed- or plain-white bars) containing constructs were tested. Indicated is the percentage value of relevant differences (t-test).

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[132, 133]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)[134] 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 and at Weill Cornell Medical College. 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 regulatory elements

In order to study the potential role of inherited genetic variants within regulatory 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 [135]. 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.

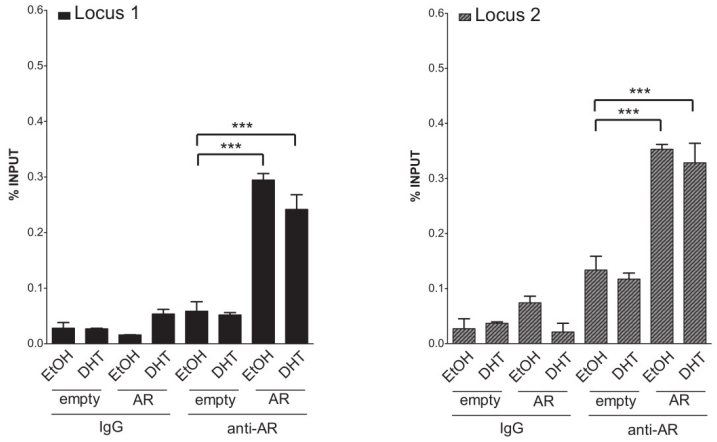

Fig. 9: Both locus 1 and locus 2 are directly bound by AR. ChIP-qPCRs were performed in MCF7 cells (heterozygous for SNP rs2242193 within locus 2) to determine AR chromatin binding at locus 1 and locus 2 regions (presented as black and grey- bars, respectively). Mean ±s.d. of 3 technical replicates were plotted. (*p<0.05, .**p<0.01, ***p<0.005, t-test).

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, t-test) (Fig 8). Further, ChIP (chromatin immuno-precipitation) 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, showed AR binding to both selected loci in MCF7 cells transiently over-expressing AR (Fig 9). 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 [135].

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 the CRISPR/Cas-9 system

We have successfully used the CRISPR/Cas-9 system to generate mutations and deletions in genes. We detected a somatic mutation in the MAP3K7 gene in hypertrophic keloid patients. In order to determine the functionality of the mutation we used the CRISPR/Cas-9 system to generate the mutation in cell lines. We successfully introduced the cancer-specific MAP3K7 mutation in HEK 293 cells using the CRISPR/Cas-9 system. Sequencing of cell lines confirmed the mutation (data not shown). 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-9 system to generate FANCA deletion in prostate cancer cell line 22RV1. Briefly, the CRISPR/Cas-9 plasmid (Px459) was obtained from Addgene (Cambridge,MA). Using a previously published protocol [136] we determined a FANCA CRISPR DNA target sequence using publicly available algorithms [137]. The oligonucleotides were cloned into Px459 vector. Western blot analysis confirmed complete absence of FANCA protein. The deletion of FANCA in 22 RV1 cells leads to increased cisplatin sensitivity[138] (Fig 10).

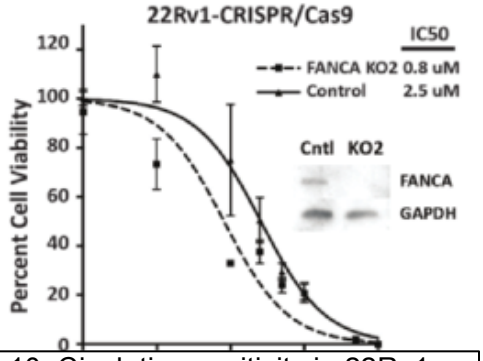

Fig. 10: Cisplatin sensitivity in 22Rv1 cells following editing of FANCA (KO1) or a control sequence by CRISPR. Inset: Western blot of FANCA & GAPDH expression in indicated cell lines.

D-4-b Approach to detailed validation

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

We will determine which of the 6 variants selected based on successful validation in Aim 3 are associated with cancer. We will achieve this by studying the specific variant in test cohorts. We will use both the Tyrol and EDRN[134] cohorts with thousands of prostate cancer individuals as well as normal controls (described above).

TaqMan assays for these 6 variants will be performed on ~4,000 cases to see if the precise variants recur in a larger cohort. Then, we will follow up for detailed functional screening, to be discussed below. For controls, we will utilize deeply sequenced control cohorts (individuals with no cancer) that are already available (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 6 selected variants will be performed on individuals that have been identified as recurrent for the variants and a similar sized group of non-recurrent individuals. We will look for perturbed gene expression in the target genes. This analysis will inform us if a SNP (in non-coding regulatory 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-9 system

We will utilize the newly discovered CRISPR/Cas-9 system[139-141] to generate endogenous mutations in TF binding sites 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-9-mediated genome-engineering method[142] we will directly generate mutations within promoter/enhancers of target genes. Theoretically we will generate 6 individual SNPs in each cell line and will study functional relevance of these changes compared to WT. Mutations within regulatory regions such as promoters and enhancers might contribute to one or more biological effects as described in Fig. 11.

The mutant and WT cell lines generated using CRISPR/Cas-9 system will be monitored for (a) perturbed expression of genes downstream of the variants using qPCR, (b) phenotypic changes by confocal microscopy and actin staining to determine effects of mutation on cytoskeletal reorganization, (c) influence on proliferation by MTT and CellTiter-Glo® Luminescent Cell Viability Assay (Promega), (d) influence on invasive and migratory potential using, matrigel coated invasion and boyden chambers in 24 well format, (e) senescence by β-gal staining and (f) 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 TF. Computational predictions for motif disruption or gain (see above, **Aim 1**) due to the variant will be validated using EMSA gel shift assays. 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 TF predicted to be present at the site of mutation. 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.

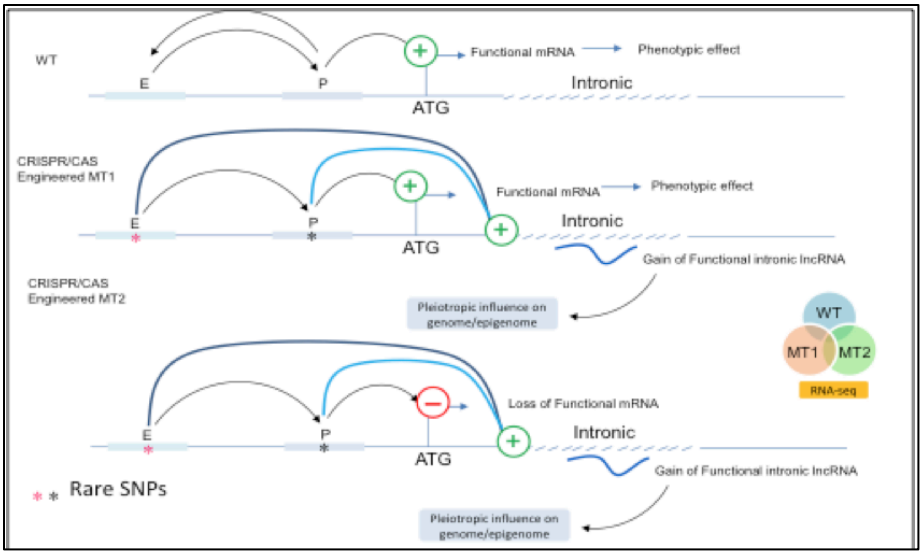

Fig. 11: *In vivo* generation of key SNPs using CRISPR/CAS-9 genome-engineering tools. WT is the parental cell line, while MT1 & MT2 are mutant cell lines harboring specific SNPs. Single or multiple effects of SNPs between the 3 cell lines will be evaluated.