Specific Aims

Prioritizing noncoding variants is a subject ripe for exploration with the new noncoding annotation from the ENCODE project as well as the many new population-scale functional genomics datasets (e.g. Geuvadis RNA-seq data). Most of the prioritization up until this point has focused on GWAS SNPs. Here we focus on a different type of variant, rare ones not in linkage with other variants, which often have much stronger effects than GWAS SNPs. In particular, we look at rare, germline SNPs (and some deletions and insertions) associated with cancer, trying to prioritize the non-coding variants most associated with disease. This work will be carried out by a team of a computational biologist and an experimental cancer genomicist that have worked together for the past 10 years.

Aim 1. Our first aim is to convert the FunSeq pipeline we previously constructed for prioritizing somatic variants into one for rare germline variants and then to significantly extend its functionality. The FunSeq pipeline defines the notion of a mutationally "sensitive" region based on population-genetic analysis. It also prioritizes hubs in the regulatory network and variants that disrupt transcription-factor binding sites. Here we will add new features to FunSeq. (1) We will elaborate its analysis of binding sites, now including gain-of-function mutations as well as disruptive loss-of-function ones. (2) We will connect all the binding sites, including those in distal enhancers, to target genes and then prioritize these sites based on their target's network connectivity (e.g. hubbiness or bottleneckness) and differential expression in cancer. (3) In addition to binding sites, we will add noncoding RNA into the pipeline and prioritize it similarly -- based on defining sensitive elements, structure-disrupting mutations and network centrality. (4) Next, we will prioritize both ncRNAs and binding sites based on their allelic activity, how sensitive their activity is to sequence differences, between maternal and paternal alleles. (5) Finally we will develop weighting schemes to combine all of the features coherently together.

Aim 2. In the second aim we will develop a large pool of rare variants and then run our elaborated FunSeq pipeline to prioritize them. Our large pool of rare variants will result from calling all the available germline variants in TCGA and ICGC whole genome sequences (est. to >2000 during the grant). We will develop a practical and efficient implementation of FunSeq to do such a large-scale compute. Our implementation will allow us to modularize the complex-to-regenerate data context (the annotation from many sources), separating it from the actual production runs on variant sets. We will also develop a special recurrence module (LARVA) to look at the degree to which the rare variants tend to recur within the same element (compared to a whole-genome background model) as well as their tendency to be in the same element that has somatic mutations in different individuals. Running the elaborated pipeline on the germline variants will allow us to develop lists of prioritized variants for aim 3.

Aim 3. In the final aim we will validate ~100 prioritized variants. We will first validate by association studies. In a separate large-scale cohort (of >5000 individuals), we will look at how these rare variants segregate with cancerous individuals versus a control. We will also look at how they are associated with downstream differential expression in large-scale RNA-seq. Then we will select a smaller pool of germline variants from this first stage of validation (~25) and subject them to detailed functional characterization. This will involve the use of reporter assays (e.g. luciferase) and also the use of the CRISPR/Cas system to generate endogenous mutations and determine their effect on biological functions.

B Significance

In this proposal we aim to prioritize rare, non-coding variants associated with cancer. Here we explain why this is significant. This work represents a collaboration between a computational scientist (Mark Gerstein) and an experimental cancer genomicist (Mark Rubin). Gerstein and Rubin have worked together for most of the last decade, co-publishing many papers during that period.

B-1 Much recent progress in annotating the non-coding genome, making it ripe for variant annotation

The Encyclopedia of DNA Elements (ENCODE) Project provides a comprehensive catalogue covering much of the entire human genome, which is further utilized to understand the genetic landscape of human diseases \cite{22955616}. In addition, the model organism ENCODE (modENCODE) Project presents an extensive

genomic annotation of *Drosophila* \cite{21177974} and *C. elegans* \cite{21177976} by systematically mapping chromatin organization, transcriptome profile and nucleosomal properties across their developmental cycle. In addition, the Geuvadis project applies large-scale mRNA and microRNA sequencing to a subset of 1000 Genome Project datasets to decipher the functional landscape of regulatory variations in the human genome \cite{24037378}. Similar efforts have been also directed toward annotating human epigenomic data to investigate underlying mechanism of various diseases \cite{23482391}.

B-2 Non-coding variants, most of which are regulatory, are significant to the study of diseases such as cancer

Noncoding variants are important in cancer. Numerous studies have been conducted on mutations to coding portions of the genome and their relation to cancer but relatively little effort has been invested into the noncoding portions of the genome. Studies of noncoding cancer disruptions can reveal mutation in regulatory features such as promoters, enhancers, and suppressors, leading to the discovery of new targets for potential cancer therapies. 99% of somatic mutations are in non-coding regions (funseq ref + other XXX refs). Much of the non-coding variation is also contributed by regulatory variants, where cis- and trans-acting variation in the human genome can modulate gene expression \cite{19636342}. Many variants implicated in cancer and diseases have been found to be related to changes in gene expression \cite{23348506,23348503,7663520,19165925,18971308}. As such, there is immense relevance and utility in integrating them into an annotation pipeline designed to prioritize disease-causing variants in the human non-coding genome.

B-3 Rare variants are significant to study of cancer & disease in general

Genome-wide association studies (GWAS) have provided only limited insight into causal variants associated with specific common diseases \cite{18987709}. Moreover, growing evidence suggests that rare variants in particular may act as primary drivers of various human diseases, including cancers \cite{11404818}. For instance, bioinformatic and biochemical analyses indicate that rare germline variants in the CHEK2 gene \cite{16982735} and PALB2 gene enhance the risk of breast cancer \cite{22241545}. In addition, a rare variant (rs138212197) on the HBOX gene \cite{22236224} and a rare SNP (rs188140481) in the telomeric region of 8q24 locus were found to be associated with prostate cancer \cite{23104005}. Similarly, a rare germline variant (rs78378222) was also implicated in the basal cell carcinoma (BCC) \cite{21946351}. In addition, somatic variations among several genes were identified as playing a key role in uterine serous carcinoma (USC) \cite{23359684}.

B-4 Rare variants in cancer patients with similar elements to somatic ones may be associated with disease risk

Prior studies have emphasized primarily on identification of somatic variants compared to germline variants in cancer - e.g. the current TCGA callset does not even contain official germline calls. However, rare germline and somatic variants have been often observed in the same genetic element across multiple individuals. We propose that the reciprocity between germline and somatic variants may increase the risk of cancer in such individuals and we plan to identify these elements in dataset taken from large population. There are multiple experimental observations that support our proposed hypothesis. Germline and somatic mutations in the promoter region of the telomerase reverse transcriptase (TERT) gene have already been observed in the cutaneous melanoma \cite{23348503}. Similarly, many somatic and germline mutations in the T53 gene and GALNT12 coding exons were implicated in Sonic-Hedgehog medulloblastoma (SHH-MB) tumors \cite{22265402} and colon cancers \cite{19617566}, respectively. The interplay between somatic and germline variants in hMSH6 and hMSH3 genes has been shown to be associated with gastrointestinal cancer \cite{11470537}. A similar association was discovered between two germline SNPs and somatic mutations in the EGFR signaling pathway in colorectal cancer \cite{24152305}.

B-**5 Rare variants at promoter and enhancer regions of the genome impose selectivity on transcription factor binding**- It has been shown that a significant fraction (26%-35%) of inter-individual differences in transcription factor binding regions coincides with genetic variation loci and that about 5% of transcripts levels are associated with inherited variant states [1]. Genotype-transcript associations have been reported at large for multiple types of inherited variants [2-6], however experimental evidence of inherited variants allele-specific effect on enhancer/promoter activities and transcriptional influence (short and long range) are lacking. In order to study the potential role of inherited genetic variants within regulatory elements in the context of hormone dependent human, the our laboratory performed an unbiased computational search for AR/ERα bound

enhancers elements containing SNPs followed by *in vitro* characterization of selected variants (Garritano S, Demichelis F, *unpublished*).

C Innovation

Our method will combine various large-scale genomics data to interpret rare non-coding variants associated with increased cancer risk. Large-scale consortia, such as the 1000 Genomes and ENCODE, have produced data that can be used to interpret other genomic studies. However, these resources have not been fully exploited to understand the functional implications of variants associated with increased cancer risk. The specific innovative components of our approach are listed below.

C-1 Identification and interpretation of non-coding rare variants associated with increased disease risk

The GWAS catalog contains many common variants associated with diseases. However, as discussed in Section XXX, many rare variants increase cancer susceptibility. Currently, no standard methods exist to functionally interpret such variants, especially in non-coding regions. Thus, our approach will be the amongst the first for functional interpretation of these variants.

C-2 Using population-scale genetic variants

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

C-3 Using non-coding annotations to understand likely biological role of non-coding variants

The ENCODE consortium has annotated non-coding regions of the genome. One of the major aims of these annotations is to help understand genetic variants that cause disease by misregulation of gene expression. Our approach will be innovative since it will be amongst the first methods that use ENCODE data to interpret variants that increase cancer susceptibility.

C-4 **Functional validation of rare variants**

Rare variations in regulatory regions of genome can have a paramount influence on biologic processes and might function as primer for recurrent somatic mutations in adjacent genomic regions or might contribute to long range changes in chromatin regulation. We will use highly innovative approach to validate functional implications of rare variants and effects thereof on biologic properties.

D Approach

D-1 Approach Aim 1 - Convert the Prototype FunSeq non-coding Somatic Variant Pipeline to Prioritize Germline Variants and Elaborate it with Many New Features

D-1-a Preliminary Results for Aim 1

D-1-a-i We have extensive experience annotating non-coding regulatory regions of the genome

We have made extensive contributions in the analysis of the noncoding genome. To more effectively utilize the ChIP-seq data for network construction, we developed a method called PeakSeq \cite{19122651} to define the binding peaks of TFs. In addition, we have also proposed a probabilistic model, referred to as target identification from profiles (TIP), that identifies a given TF's target genes based on ChIP-seq data \cite{22039215}. Furthermore, we have developed machine-learning methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal regulatory regions (DRM) \cite{20126643}. As published in \cite{22950945}, we have validated some of our DRM results by experiments, which show a fairly high predictive accuracy.

Using the machine-learning approaches we developed for identifying individual proximal and distal edges together with miRNA target prediction (and other) algorithms, we have completed the highly ambitious goal of constructing highly integrated regulatory networks for humans and model organisms based on the ENCODE \cite{22955619} and modENCODE datasets \cite{21430782}. In addition to analyzing the topology of

gene-regulatory networks, we developed methods to determine the hierarchical organization of regulatory networks and applied them to analyze the regulatory networks of a variety of species from yeast to human, including networks constructed from ENCODE, modENCODE and MCF7 data \cite{22125477,22955619,21177976,20439753}.

We have developed statistical models of open chromatin associated with gene-expression \cite{21926158,22060676}We have also conducted extensive studies of the relationship between ChIP-Seq data for localization of transcription factors and histone modifications and gene expression through RNA-Seq \cite{22955619,22955978}.

D-1-a-ii We have extensive experience processing RNA-Seq data and annotating ncRNAs

We have developed a ncRNA-finder \cite{21177971}. For general RNA-Seq analysis, we have developed RSEQtools, a computational package that enables expression quantification of annotated RNAs and identification of splice sites and gene models \cite{21134889}. In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts \cite{22238592}. Comparisons between RNA-Seq samples, and to other genome-wide data, are facilitated in part with our Aggregation and Correlation Toolbox (ACT), which is a general-purpose tool for comparing genome signal tracks \cite{21349863}. We also have extensive experience conducting integrated analyses of large sets of RNA-Seq data, such as through the ENCODE project \cite{22955616}. We played a leading role in the analysis of model organisms (C.Elegans) and human transcriptome studies within the ENCODE consortium, two of the largest RNA-Seq studies to date \cite{22955620;21177976}.

D-1-a-iii We have extensive experience in Allelic Analysis

A specific class of regulatory variants is one that is related to allele-specific events. These are cis-regulatory variants that are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins and allele-specific expression (ASE). \cite{ 20567245,20846943} We have previously developed a tool, AlleleSeq, \cite{21811232} exclusively for the detection of candidate variants associated with ASB and ASE events. The tool takes in as input: (1) the variants from an individual and (2) sequence reads from the ChIP-seq and RNA-seq experiments performed on the same individual. It first constructs a diploid personal genome using the individual's variants and phases it into its haplotypes. Subsequently, it maps the ChIP-seq and RNA-seq reads to this newly constructed personal genome. Detection of variants with ASB and ASE behavior will only be apparent at heterozygous loci, i.e. sites with two different alleles. Hence, AlleleSeq statistically infers whether there is a differential read count (from ChIP-seq or RNA-seq data) between the two alleles at each heterozygous variant.

We have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project. \cite{22955620, 22955619, 24092746} We showed that allele-specific variants are more likely to be rare variants \cite{24092746}. We have also shown that there is a substantial number of genomic elements associated with ASB and ASE in the human genome - for example, about 18% of genes exhibit ASE. \cite{22955620} We demonstrated that many transcription factors (TFs) act in a parent-of-origin-specific manner and those exhibiting such allele-specific binding behavior are more likely to have more target genes, i.e. more promiscuous in their activities. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and expression. \cite{22955619} These analyses are performed only on a cell line derived from an individual of northwestern Caucasian ancestry, GM12878, which has a deeply sequenced diploid genome and extensive RNA-seq and ChIP-seq experiments performed on this same cell line at the time of the publications.

Furthermore, we have provided the AlleleSeq tool, lists of detected AS variants from the publications and the constructed personal genome and transcriptome of NA12878 on our website (alleleseq.gersteinlab.org). Since then, we updated our AlleleSeq tool, and the resource has been used a bit in the scientific community, as exemplified by the number of citations and publications using our data as references. \cite{23569280,24371156}.

D-1-a-iv We have extensive experience in relating annotation to variation and based on this experience have developed the prototype FunSeq pipeline for Somatic Variants

We have extensively analyzed patterns of variants in non-coding regions along with their coding targets \cite{21596777,22950945,22955619}. We used metrics such as diversity and fraction of rare variants to characterize selection pressures of various classes and subclasses of functional annotations \cite{21596777}. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region in

ENCODE \cite{22955616}. Further studies by our group showed relations between selection pressures and protein network structures, e.g. hubs and periphery \cite{18077332,23505346}.

In a recent FunSeq study \cite{24092746}, we further extended and integrated these methods to develop the notion of sensitive and ultra-sensitive regions, i.e. those annotations under strong selection pressure as determined by human population variation, network connectivity, and disruptive non-coding mutations. In particular, by contrasting patterns of inherited polymorphisms from 1092 humans with somatic variants from cancer patients, we developed a scheme and a software tool (FunSeq) for identification of candidate non-coding driver mutations \cite{24092746},. In this study, we integrated large-scale data from various resources, including ENCODE and 1000 Genomes Project, with cancer genomics data. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples. It identifies deleterious variants in various non-coding functional elements, including: transcription-factor (TF) binding sites, their higher resolution motifs, regions of active chromatin corresponding to enhancer elements and regions of open chromatin corresponding to DNase I hypersensitivity sites.

D-1-b Research Plan for Aim 1

We plan to extend the current FunSeq scheme to be more comprehensive for identification of rare variants associated with complex traits. We will do some simple improvements (i.e. incorporating GERP scores and ultra-conserved regions for identifying conserved regions between species) while some major changes are outlined below.

D-1-b-i Loss-of-function and gain-of-function for transcription factor binding sites

Loss-of-function variants are more likely to cause deleterious impact \cite{24092746, 21177971}(Kheradpour, et al., 2013). When variants occur in transcription factor binding motifs, the change in position-weight matrix (PWM) can be calculated. Variants decreasing the PWM scores could potentially alter the binding strength of transcription factors, or even cause loss-of-motif events. Many studies have shown that gain of new binding sites caused by somatic mutations can constitute driver events (Horn, et al., 2013; Huang, et al., 2013; Killela, et al., 2013; Vinagre, et al., 2013). However, an automated tool to detect such events in whole genomes is not available. Such events in germline genomes might also be associated with increased disease risk. We will create a gain-of-motif scheme to scan and statistically evaluate (Touzet and Varre, 2007) all possible motifs created by variants compared to human reference genome. Gain-of-motif events are identified as those that give sequence score with mutated allele in PWMs significantly higher than the background ($p < 4e-8$). Note that in these analyses, determining the ancestral allele of the variant is essential to resolving between loss-of-function or gain-of-function since the functional impact of the variant reflects the historical event when the polymorphism was first introduced in the human population.

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

To interpret likely functional consequences of noncoding variants, we will comprehensively define associations between regulatory elements and genes through correlating various epigenetic modifications (e.g. H3K27ac and DNA methylation) with expression levels of genes. We will consider the enhancer marks H3K4me1 and H3K27ac as two types of activity signals, and DNA methylation as an inactivity signal. We will collect all bisulfite sequencing, ChIP-seq and RNA-seq data from the Roadmap Epigenomics project. For each regulatory element-candidate target pair, we will compute the correlations of their activity/inactivity and expression levels across different tissue types. To incorporate the ever-increasing amounts of genomic data, we will offer a flexible framework for users to extend the scheme with their own data. as well as integrate gene expression studies in cases vs controls to increase predictive power for identification of functional variants.

Interpretation of the functional impact of variants can be greatly enhanced if the function of its target protein-coding gene is known. We will incorporate prior knowledge of genes, such as known cancer-driver genes (Futreal, et al., 2004) and actionable genes ('druggable' genes) (Wagle, et al., 2012) into our annotation scheme. Differential expression of target genes is indicative of potential effect of noncoding variants. We will provide a "differential gene expression analysis" module to detect differentially expressed genes in cancer samples from RNA-Seq data.

We will also use the target genes to connect the non-coding elements into a variety of networks. In particular, for all the noncoding variants passing other filters, we will examine the network centralities of the associated genes in various networks, since disruption of highly connected genes or their regulatory elements is more likely to be deleterious (Khurana, et al., 2013; Kim, et al., 2007). We will make the scheme flexible so it can integrate user networks in addition to the pre-collected networks such as protein-protein interaction,

regulatory and phosphorylation networks. In addition to hubs, we will also prioritize based on bottlenecks and positions at the top of heirarchies [PMID XXX & XXX]

D-1-b-iii Detailed Variant Prioritization for ncRNAs

The original FunSeq focused on TF binding sites. Here, we will expand FunSeq to better prioritze variants in ncRNAs, in parallel to as we have for binding sites. We will first prioritize ncRNAs based on their within-human selection pressure and conservation across multiple species, and identify sensitive regions. For within-human selection, we will prioritize annotations showing higher nucleotide diversity and fraction of rare variants \cite{21596777}. We will look at GERP scores \cite{20354513} for inter-species conservation.

We will further divide ncRNA annotations according to their subcategories, expression levels, and specificity of expression in cell lines. We will take into account ncRNA subcategories including transfer RNAs, miRNAs, 5S ribosomal RNAs, small nucleolar RNAs, small nuclear RNAs, and long non-coding RNA. Expression levels of ncRNAs will be obtained from ENCODE (http://genome.crg.es/encode_RNA_dashboard/hg19/) where RNAseq was performed on dozens of cell lines. We will prioritize ncRNAs that have higher expression levels and those that are ubiquitously expressed in multiple cell lines.

Furthermore, we will annotate genomic variants against secondary structures of ncRNAs. Our preliminary data have shown that more rigid structures, such as stem regions, are under stronger selection pressure, and that those variants that incur a larger free energy change of the structures tend to be rarer in the human population. To better characterize variants in the context of structures, we define variants that disrupt secondary structures of ncRNAs as those that no longer form a complementary base-pairing or a wobble base-pairing when mutated. RNA secondary structures will be predicted using RNAshapes \cite{16357029}. We will also quantify the effect a mutation stabilizes or destabilizes the RNA structure by computing the difference in folding free energy changes of the RNA before and after the introduction of the mutation. (Again the correct identification of the ancestral allele will be important here.)

Finally, we will explore the relationship of ncRNAs with network connectivity by associating ncRNAs with canonical genes through expression levels, sequence complementarity, etc. miRNA, for instance, are known to regulate the expression level of its target genes. We will identify coding genes associated with miRNAs by correlating their expression levels based on RNAseq data. In addition, we will also search for potential miRNA binding target by examining sequence complementarity in 3'UTR regions of coding genes to the seed regions, i.e. the first 2-7 bp of the mature miRNAs, using TargetScan \cite{17612493}. We will then examine the selection pressure in ncRNAs that are associated with genes in network hubs vs. periphery.

D-1-b-iv Variant Prioritization Based on Allelic Activity and Association with EQTLs (AlleleDB module)

The direct interrogation of allelic imbalance of reads (either ChIP- or RNA-seq) at each heterozygous locus in the genome forms the basis of allele-specific (AS) analyses. Even though epigenetic factors such as imprinting \cite{18308616}, histone marks \cite{21812971} and random monoallelic expression events \cite{18006746} can confound the causality between genotype and the allele-specific behavior, this still allows us to identify specifically which variants are associated with allele-specific events. Their regulatory roles assert their usefulness in the prioritization of functional variants. However, currently, there is no prioritization scheme that integrates both allele-specific binding (ASB) and expression (ASE) regulatory variants. Further, as mentioned, most ASB and ASE analyses are focused on a single individual, GM12878. Thus, an enrichment of rare variants among AS variants implies that there are many uncovered variants associated with ASB and ASE. A direct overlap of variants in a prioritization pipeline will not be applicable here.

Therefore, to further enable the incorporation of allele-specific variants into the annotation pipeline, we define what we term 'allelic' genomic elements.. 'Allelicity'is defined as the degree of how allele-specific a particular genomic element or a category of elements is, as opposed to 'allele-specific' being relatively more defined of having evidence of allelic imbalance. For example, an 'allelic' class of transcription factor binding site (TFBS) might possess more allele-specific binding (ASB) variants, or a particular class of elements such as enhancers and promoters might be more allelic than another.

To enable us to define an 'allelic' element, we plan to extend AS analyses to a diversity of individuals. There has been an increasing number of large-scale ChIP-seq and RNA-seq experiments performed on the genomes from the 1000 Genomes Project \cite{23128226}, in which we are active members. These datasets are found in various publications \cite{20220756, 20220758, 21173033, 24136359, 24136358, 24136355} and consortia, notably ENCODE \cite{22955616} and gEUVADIS \cite{24037378}. Our intent is to amass these datasets and detect allele-specific variations found in these ~500 individuals via the AlleleSeq tool, based on their variants found in the 1000 Genomes Project and corresponding functional genomics assays. The

hypothesis is that allelic elements (that exhibit allele-specific behavior) should be similar in most, if not all, individuals. So, even though each individual will have a private set of AS variants, more AS variants (aggregated from across the individuals) should be found in these allelic elements. Consequently, within the context of the prioritization scheme, we can capture and up-weight variants found in a more allelic element, as it will more likely have a functional role.

The aggregation of allele-specific variations in multiple individuals has several advantages. First, it endows the statistical power to detect categories of genomic elements such as enhancers and promoters that are highly enriched in allele-specific variants. Since many of the rare disease variants are not expected to overlap, defining “allele-specific” categories allows us to annotate the possible functionality of the variant within the context of a genomic element. Second, it facilitates the survey of the characteristics of allele-specific variants in various dimensions: analyses can be variant-based, element-based or based on a class of categories; analyses can also be population-based since the ancestries of the various individuals are provided. In addition, analyses can be based on transcription-factor-of-origin for ASB-associated variants.Third, the availability of trio families will allow the investigation of Mendelian inheritance of allele-specific events, both on a per-variation and per-element basis. Fourth, it allows the study of degree of coordination of ASB and ASE, whether such is consistent across multiple individuals.

The results will be housed in a central repository, which we called the AlleleDB, which will provide a catalog of allele-specific variations and the elements they are found in. This is especially useful for researchers interested in the genetics of gene expression variation in complex diseases, e.g. in genome-wide association studies (GWAS) or cancer. For example, loci identified for differential gene expression profiles in a case-control study can simply be queried in AlleleDB, thereby providing a potential link between a phenotype and a genetic variant or region associated with allele-specific behavior, and ultimately a molecular mechanism. Additionally, the database can act as a benchmarking reference for allele-specific tool development. For the purpose of a prioritization pipeline, the lists of allelic genomic elements can be used in the weighting scheme to rank variants according to their existence in allelic elements.

D-2 Approach Aim 2 - Develop Efficient, Extensible and Easy to Use Pipeline and Run on all the Germline Variants in TCGA

D-2-a Preliminary Results in Pipeline Development & Variant Calling

The Gerstein lab has much experience in developing tools for bioinformatics research. Our tools take the form of open source programs, or databases and web applications which are hosted on Amazon Web Services Elastic Compute Cloud (AWS-EC2). For instance, to extract knowledge from high-throughput genomic experiments we have developed RSEQtools \cite{21134889} to quantify RNA expression, IQSeq \cite{563456243} to identify alternative splicing, Fusion-seq \cite{363456856} to identify fusion transcripts, and BreakSeq \cite{234645647} to identify copy-number variation. To provide insights in the field of structural genomics we have developed the Database of Macromolecular Movements \cite{23414235} to catalog protein dynamics, HingeMaster \cite{4142563436} for normal mode hinge prediction, and RigidFinder \cite{23452653} for finding ridge blocks in macromolecules. To use networks for mining functional genomics experiments we have developed TopNet-like Yale Network Analyzer (tYNA) \cite{35234263} for managing, comparing and mining multiple networks, and YeastHub \cite{124235364} to apply semantic web technologies to more efficiently query life sciences data and meta-data.

We have much experience in large-scale germline variant calling through being active members of the 1000 Genomes Consortium, especially the Analysis Group and SV subgroup where majority of the variant calling tools are developed \cite{21787423;21293372;20981092;23128226}[[refs XXX & XXX]]. We will use GATK for variant calling, which we have already extensively used previouly [[FunSeq ref XXX]]. For rare variants, we will define them as those not in 1000 Genomes (phase 1 or pilot) -- the “outersect” with 1000G -- as was done previously [[refs to XXX ENCODE production & Funseq]]. Also, we will consider SVs, which are important contributors to human polymorphism. [[SVs are common \cite{17122850} and usually have high impact.]] Moreover, many SVs have high functional impacts and are associated with diseases [[inherent diseases or disease susceptibility]] \cite{23311762,22621759,8822366}. Characterizing SVs, especially accurately locating breakpoints, is a critical yet challenging task. [[We have demonstrated that precise SV location information provides insightful understanding of the genome \cite{20037582}.]] We have developed a number of novel SV calling algorithms, including BreakSeq [[, to facilitate the detecting of SVs]] by comparing raw reads with a breakpoints library (junction mapping) \cite{20037582}, CNVnator by measuring read depths \cite{21324876}, AGE by refined local alignment \cite{21233167}, array-based approaches \cite{19037015} and a sequencing-based bayesian model \cite{21034510}.

D-2-b Research Plan for Aim 2

D-2-b-i Do SNP & a limited amount of SV calling for all Germline Variants in TCGA and ICGC

We currently have access to a combined ~600 whole genome (WG) sequences from the Sanger Institute, and from prostate cancer sequencing projects. We anticipate access to another ~2000 WG sequences from the International Cancer Genome Consortium (ICGC)\cite{20393554} and The Cancer Genome Atlas (TCGA) \cite{24084870}. We plan to call variants from this data for use in our variant pipelines.

We have already developed a prototype pipeline for calling germline and somatic changes using the Broad's Genome Analysis Toolkit (GATK)\cite{21478889}. Basically, we will run GATK with standard parameters and then filter the results. We will use this pipeline to call the variants in the ICGC and TCGA cancers. We estimate that this will take 3 months. Additionally, we will employ our CNVnator software on this data to identify copy number variants. We will filter against 1000 Genomes Phase 1 to define rare variants. We plan to generate a pool of rare, diseased variants to prioritize and validate the disease-causing variants. We are considering a couple of strategies for generating this variant pool. We estimate that within a single whole genome sequence's total variant set of roughly 3 million, there will be some 10,000 somatic variants, and some 100,000 rare variants.

D-2-b-ii Analysis of Recurrence of Germline & Somatic Variants (LARVA module)

We will develop a model to study the recurrence of both germline variants and somatic mutations across multiple cancer patients. We will aim to see if there are prioritized germline variants that affect the same element as somatic ones (in different individuals).

On a simple level, recurrence would be a variant at exactly the same position in two individuals. However, this is exceedingly unlikely for rare variants \cite{20981092}. Thus, we will consider variant burden spread over elements. These elements can be single annotations, such as exons, pseudogenes, noncoding RNA, and regulatory features like promoters and enhancers. On a more complex level, we will consider groups of genes related through a common pathway, or through a protein interaction subnetwork, as a single element, where variants from multiple patients that map anywhere in the gene group represent a recurrence.

Recurrences in the somatic variants are indicative of the elements that are disrupted in cancer, and are important in driving cancer progression. If these elements are also recurrently mutated in the set of rare germline variants, then the rare germline variants that overlap these elements may have a role in cancer as well. These germline variants may serve as a precursor to the development of cancer characteristics, or indicate a predisposition to cancer.

We're going to do the recurrence in staged fashion: (1) find somatic recur. (2) find rare germline (3) inter-relate these 2 and then also see if there's common variants . The most deleterious case would be rare variants in some indiv. that overlap in the same element with somatic varisants in another indiv & there is no common variants in the element at all.

We have developed a computational framework for identifying these types of recurrent variation, named Large-scale Analysis of Recurrent Variants and Annotations (LARVA). Given a set of cancer patient whole genome variant calls, and a set of genome annotations, LARVA will pick out the recurrent variants and recurrently mutated annotations that result from overlapping the variants with the annotations.

In the extreme, studying recurrent variation in the context of metabolic and signalling pathways would increase our understanding of the systems-level disruption of cancer. This line of thinking can be extended to the investigation of subnetworks of interacting proteins: their corresponding genes may contain mutations spread across multiple samples in an arrangement that would not implicate individual genes for recurrent variation, but would indicate the subnetwork as a whole is recurrently mutated.

LARVA also has a module for computing the statistical significance of its results by simulating the creation of whole genome variant calls with randomized variant positions. These random datasets, which otherwise contain the same number of samples and variants, are used to determine the null distribution of variants across the annotation set for comparison with the actual variant data. LARVA also features a number of utility scripts for postprocessing the results.

LARVA determines the positions of variants for its random variant datasets using a null mutation model designed to reflect the expected differences in the neutral mutation rates of different genome regions. The expected neutral mutation rate of the exome is influenced by a number of factors \cite{23770567}. Genes with higher expression undergo more transcription-coupled repair, resulting in lower mutation rates. DNA replication timing during S phase also influences mutation rates: later replicating genes have fewer free nucleotides to draw on, which leads to higher error rates. Finally, the state of a gene's chromatin can make it more or less susceptible to mutation processes. We use these factors in a weight function defined over genes, where higher

weight is assigned to genes with higher mutation rates. This weight function is defined as follows for each gene g:

weight(g) = log(1-CDF(g.expression)) + log(CDF(g.replication_timing)) + log(1-CDF(g.chromatin_state)) + log(CDF(g.length))

Individual variant positions are selected by first choosing a gene according to this weight function, then picking a position within that gene with uniform probability.

LARVA's whole genome null mutation model uses the genome-wide DNA replication timings, and histone marks for H3K4me1 and H3K4me3, which are anti-correlated with SNV density \cite{22820252}. It also includes whole genome RNA-seq data from the ENCODE project \cite{22955616} and SNV density data from the 1000 Genomes Project \cite{20981092}. The whole genome weight function is defined over discrete regions of the genome, rather than genes, and is defined as follows for each region r:

weight(r) = log(CDF(r.replication_timing)) + log(1-CDF(r.H3K4me1)) + log(1-CDF(r.H3K4me3)) + log(1-CDF(r.expression)) + log(CDF(r.SNV_density))

D-2-b-iii We will implement FunSeq on a large scale & then run on all the variants to produce a shortlist of prioritized variants

D-2-b-iii-1 - split it into data context and production runs and then further subsplit into modules

We will develop a practical implementation of all of the FunSeq modules and then integrate them within FunSeq. Some of the modules may be useful as stand alone programs.We're going to make funseq into a useful tool. It'll downloadable s/f, website, cloud inst. We're going to try to split it into data context and production runs and then further subsplit into modules… For instance, one submodule might be.For instance, for AlleleDB, the results will both be integrated into the pipeline and also housed in the AlleleDB database. AlleleDB can be navigated via a user-friendly interface for data mining and the casual user. It will also generate flat files for their queries and can be subsequently downloaded by the users for further analyses. Hence, AlleleDB will serve as a valuable resource for both bioinformatics and non-informatics users who are interested in allele-specific variants or regulatory variants in general.

We will integrate large-scale publicly available data resources to build the data context. It will contain polymorphisms from 1000 Genomes project (Genomes Project, et al., 2012), conservation data from (Bejerano, et al., 2004; Cooper, et al., 2005), functional genomics data from ENCODE (Maher, 2012) and REMC (Bernstein, et al., 2010), gene function (Futreal, et al., 2004; Wagle, et al., 2012) and networks data.

D-2-b-iii-2 - we're going to develop a uniform weighting system for all the features and modules (including recurrence)

An integral part of the modular nature of FunSeq, will be the weighted scoring system….We will develop a scoring scheme where each continuous feature (e.g. motif-breaking scores) will be normalized on a scale of 0 to 1 using different methods (e.g. sigmoid transformation). We will also develop a weighted-sum scoring scheme, based on the mutation pattern observed in 1000 Genomes polymorphisms. The frequencies of observing each feature in polymorphisms data would be calculated. We would then use the entropy measure to define the weight, where features with higher entropy would be assigned a lower weight. We will also consider the dependency structure of features when calculating the scores.

D-2-b-iii-3 3 - we're going to let it rip -- ie we'll run it and generate shortlist

Our first possible strategy involves collecting the rare variant set of a single personal genome. We will run these variants through our Funseq pipeline, identify the top 20 variants for further investigation, and pass these variants along to the lab of Prof. Mark Rubin for validation. Our second possible strategy involves collecting the rare variant set of 500 WGS cancer patients, producing a pool of 50 million rare variants, which we would then run through the same validation pipeline. However, this variant pool is significantly larger, and therefore would be more compute-intensive to analyze.

D-2-b-iv We will run FunSeq & Larva on all the variants & prioritize them

Given the 100K germline +10K somatic var per person for 2K indiv We'll run funseq. to get a prioritized list… We're expecting that the most time consuming step will be the LARVA recurrence analysis against the null model … this will take some 50 CPU years. However, we are working on strategies to process this computation more efficiently. In addition to algorithmic optimizations, we will make use of cloud computing resources and parallelization to speed up LARVA's computations.

The prioritization of the variants in aim 2 will generate a number of long list prioritized variants, which will fall into three basic groups. 1. A general list of prioritized rare variants based on all of the variants from the

individuals with cancer. 2. A subset of that list where the prioritized rare variant also occurs within the same element as a recurrent somatic variation. 3. A further subset of those variants, which are associated with prostate cancer.

D-3 Approach Aim 3 - Validate the Prioritized Variants

[(MG-to-Mark-R: Your validation sect. should be 3.5pg)]]

D-3-a Preliminary Results Related to Validation

Low-frequency functionally active intronic and intergenic inherited variants predisposing to prostate cancer

Emerging insights into the genetics of constitutional disease etiology demonstrate that germline polymorphisms are associated with a variety of diseases including Alzheimer's, Parkinson's, mental retardation, autism, schizophrenia [7] and cancer [8,9]. Relevant to this proposal our group recently performed a large scale profiling study for 2,000 individuals from the Tyrol Early Prostate Cancer Detection Program [10,11] cohort. This cohort is part of a population-based prostate cancer screening program started in 1993 and intended to evaluate the utility of intensive PSA screening in reducing prostate cancer specific death. By genotyping DNA extracted from peripheral blood samples, we annotated the cohort on more than 5,000 CNVs and 900,000 SNPs and then queried inherited low frequency deletions variants [12] for their impact in driving prostate cancer [13] and the more aggressive form of the disease [14]. We reported on coding and non-coding functionally active risk variants. Among the top hits of the case-control study, an intronic variant in the *Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C (MGAT4C)* demonstrated transcript abundance association with genotype states both in prostate and in lymphoblastoid cells, significant increase in cell and migration upon overexpression in benign and cancer prostate cell lines, and significant decrease in proliferation upon knock down of *MGAT4C* expression with siRNA. In addition, we suggested that intergenic PCA risk variants affect gene regulation through modified transcription factor binding activity of the Activator Protein 1 (AP-1) [1,5]. Altogether, we demonstrated that inherited variants may directly or indirectly modulate the transcriptome machinery of known oncogenic pathways in prostate cancer facilitating carcinogenesis.

	H3K4me1	H3K4me1 + H3K3me3	H3K27ac	H3K9ac	DNase	FAIRE	UNION
AR	373 (136)	183 (55)	283 (98)	258 (83)	127 (39)	52 (16)	418 (148)
ER	386 (102)	221 (60)	317 (82)	339 (90)	232 (56)	127 (32)	431 (113)
AR+ER	17 (7)	9 (4)	14 (7)	17 (7)	6 (2)	3 (1)	22 (8)

***Table XX**: SNPs from the human genome that intersect regulatory regions bound by AR and/or ERα.*

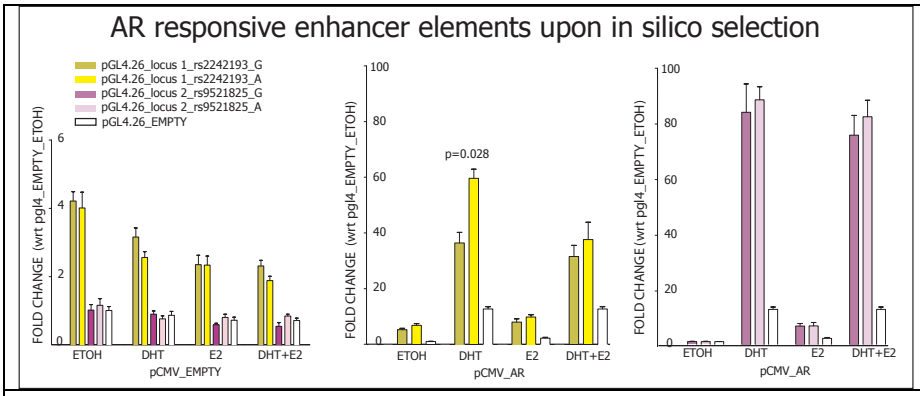


***Figure XX.** Validation AR-responsiveness and impact of SNPs within identified putative enhancer elements. Left panel) Cells cotrasfected with pCMV_EMPTY vector and the different pGl4.26 constructs. Only pGl4.26_locus 1_rs12242193_G/A reaches as much as 4 fold change. Center panel) Cells cotrasfected with pCMV_AR vector and pGl4.26_locus 1_rs12242193_G/A. Results show that the SNP has a role in the transcription regulation when cells are supplemented with 100nM DHT (p=0.028, determined by Student's t-test). Right panel) Cells cotrasfected with pCMV_AR vector and pGl4.26_locus 2_rs9521825_G/A. When cells are supplemented with 100nM DHT the construct reaches as much as 80 fold hinting at its strong enhancer role. All experiments were performed from three biological replicate, each one consisting of three technical replicates. Error bars indicates standard deviation of the mean (SD).*

The Tyrol Early Prostate Cancer Detection Program cohort The Tyrol Early Prostate Cancer Detection Program cohort is a well characterized cohort with centralized data collection that ensures proper patients' follow-up annotations and availability of well-preserved tissues and blood samples. The cohort currently includes more than 3,000 men. As part of our Trento-Innsbruck-Cornell collaboration, we further studied the genetics of prostate cancer individuals coupling serum

levels and genomics data. Specifically, we studied the impact of genetic variants relevant to the metabolism of Dihydrotestosterone [15] (DHT), the most potent form of androgen, and investigated the incidence of common genomic rearrangements with respect to PSA levels and age at diagnosis [16].

In vitro characterization of SNPs within enhancer elements bound by AR and/or ERα. It has been shown that a significant fraction (26%-35%) of inter-individual differences in transcription factor binding regions coincides with genetic variation loci and that about 5% of transcripts levels are associated with inherited variant states [1]. Genotype-transcript associations have been reported at large for multiple types of inherited variants [2-6], however experimental evidence of inherited variants allele-specific effect on enhancer activity are lacking. In order to study the potential role of inherited genetic variants within regulatory elements in the context of hormone dependent human, the Demichelis laboratory performed an unbiased computational search for AR/ERα bound enhancers elements containing SNPs followed by *in vitro* characterization of selected variants. **Table XX** shows counts of SNPs from the dbsnp137 set within AR [17] and/or ERα (D. Chakravarty, *submitted*) binding sites that intersect peak ENCODE data [18] generated from 20 cell-lines and ChIP-seq experiments for H3K4m1, H3K4me1+H3K4me3, H3K9ac, H3K27ac, Dnase-seq and FAIRE-seq. For each marker the consensus was generated as the merge of all the regions that are present in at least 2 cell lines and comply with a set of filters. **Figure XX** shows examples of AR-responsiveness and SNPs impact on putative enhancer elements in MCF7 cells (Garritano S, Demichelis F, *unpublished*).

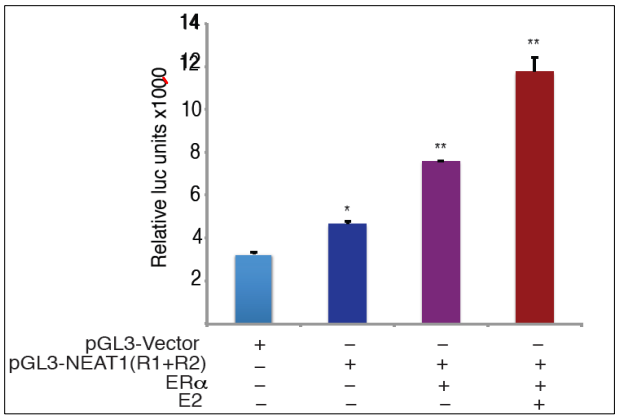


Fig 2- Promoter luciferase assay were performed using promoter reporter vectors encompassing ERα binding regions upstream of NEAT1. Luciferase assays confirmed ERα is recruited and drives transcriptional output from NEAT1 promoter

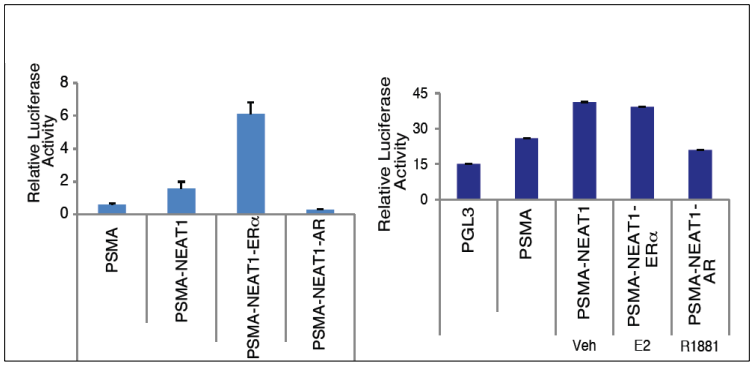


Fig 3- Promoter luciferase assay confirm NEAT1-ERα axis is involved in regulation of PSMA, a key prostate cancer gene.

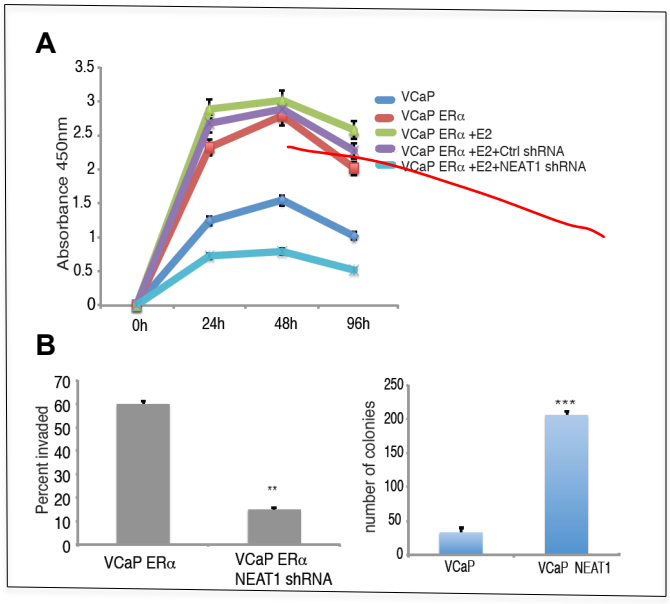


Fig 4: NEAT1 is a driver of oncogenic cascade
(a) Cell proliferation assays were performed in VCaP and VCaP ERα expressing cells transfected with control or NEAT1 siRNA and estrogen treatment (10nM). (b) (left) Quantitative bar chart for depicting the relative cell counts obtained at the completion of the invasion assay performed in VCaP ERα control and NEAT1 shRNA expressing cells, (**) $p < 0.01$. (right) Soft agar assays were performed with VCaP control and NEAT1 expressing cells. Quantitative bar-plot analysis of stained colonies at 21 days are shown), (***) $p < 0.001$.

Reporter luciferase assays confirm validity of insilico Transcription factor binding sites. Using and insilico approach we determined genome wide distribution of ERα in prostate cancer. Intriguingly, we observed a robust recruitment to non-coding genome and identified several intergenic sites that correlated with high ERα occupancy. Analysis of recruitment vs transcript profiles confirmed that ERα recruitment was associated with productive transcription of long noncoding RNA.

Recruitment of ERα upstream of NEAT1 lncRNA was addressed in greater details. Reporter assays using promoter luciferase constructs encompassing upstream regulatory regions of NEAT1 and corresponding to two ERα binding sites are described in Fig 2.
Interestingly, we discovered that NEAT1 is associated with chromatin and regulates transcription of key prostate cancer genes. Recruitment of NEAT1 was evaluated by ChIP assay and influence on key target genes like PSMA was validated using ChIP and reporter assays (Fig 3). Functional validation of NEAT1 functions revealed a predominant tumorigenic role as overexpression of NEAT1 was sufficient to augment proliferation, invasion and migratory behavior of prostate cancer cells (Fig 4).
The above results clearly establish our laboratory to validate key insilico findings in search of biologic functions.

D-3-b Research Plan Related to Validation

We will take the highest prioritized variants then subject them to validation. Overall we plan to start the validation pipeline with about a hundred variants. First we would perform an initial screen to determine whether any of the variants are associated with cancer in a different cohort of individuals or are associated with differential gene expression and RNA-seq. We will use both the Tyrol cohort (described above) and the Early Detection Research Network (EDRN) (http://edrn.nci.nih.gov/) prostate cancer cohort with thousands of prostate cancer individuals as well as normal controls. The prostate cancer cohort include men enrolled at three sites as part of the Prostate Cancer Clinical Validation Center that prospectively enroll individuals at risk for prostate cancer at Beth Israel Deaconess Medical Center (Harvard), at the University of Michigan (Michigan) and at Weill Cornell Medical College (Cornell). Cases are defined as men diagnosed with prostate cancer and controls are men who have undergone prostate needle biopsy without any detectable prostate cancer and no prior history of prostate cancer.
This will give rise to a smaller subset of variants, approximately 30, that we will follow up for detailed functional screening. This functional screening will be through various reporter assays (e.g. luciferase) looking for the effect on the target gene and also from using the CRISPR/Cas system.

We plan to integrally feed back some of the results from the validation into refining the pipeline though obviously the number of things being validated is not large enough for large-scale statistical parameterization.

D-3-b-i Genotyping

We will utilize robust Taqman genotyping assays for screening 100 nominated rare variants in a cohort of 4000 individuals. Superior allelic discrimination is achieved in these assays as they utilize TaqMan minor groove-binding (MGB) probes. This technique generates a low signal to noise ratio and affords a greater flexibility. The Taqman probes are functionally tested to first ensure assay amplification and optimization for amplification conditions.
Methods: Genomic DNA will be extracted from the blood cellular-EDTA samples in a high-throughput fashion using the QIAamp 96 DNA Blood Kit (Qiagen). All DNAs are evaluated by NanoDrop spectrophotometer (NanoDrop, Thermo Scientific) and gel electrophoresis (2% agarose). For TaqMan Real-Time Quantitative PCR, each DNA sample is diluted to 10 ng/μl with nuclease-free water.

D-3-b-ii RNA-seq

SNPs that are recurrent amongst the 100 nominated rare variants will be studied further. RNA seq analysis will inform us if SNP (in promoter or enhancer regions) has any direct effect on transcription of target gene. This analysis will provide a comprehensive list of SNPs that might correlate with loss or gain of expression. Recurrent rare SNPs will be further validated by PCR assays using primers that can amplify the genomic region encompassing the SNP. PCR will be followed by direct sequencing of amplicon using an ABI 3730 DNA Sequence Analyzer on a subset of tumor-normal pairs to verify the individual promoter/enhancer mutations for further confirmation.

D-3-b-iii Evaluation of functional consequence of variants

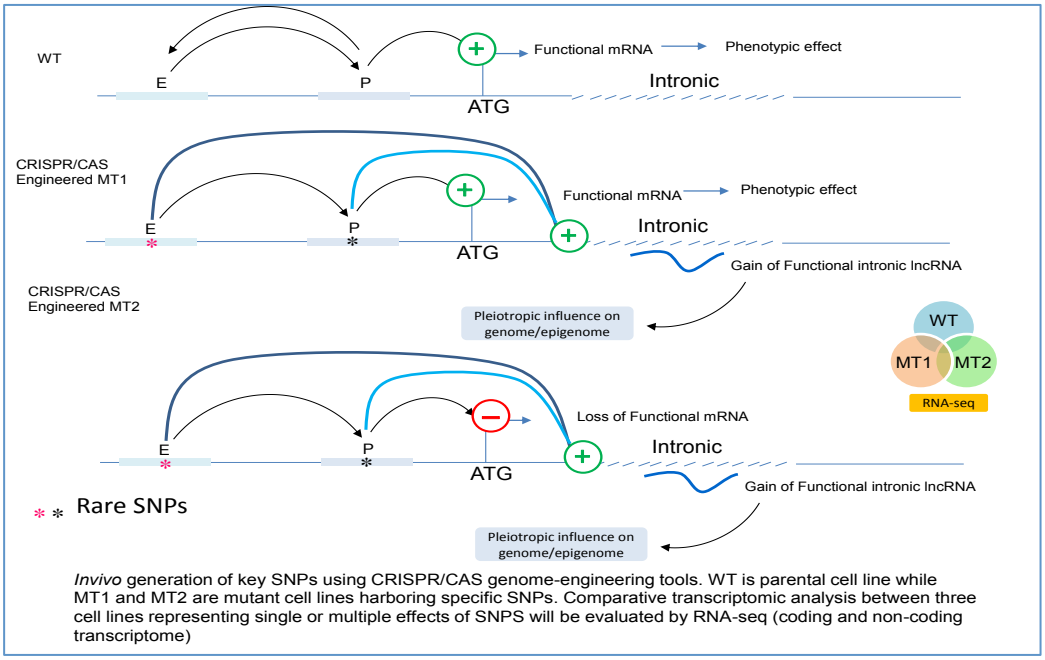
D-3-b-iii-(1) Functional consequences: Reporter Assays

Reporter assays that employ either (LUC or next generation reporter vectors can provide direct insight to functional relevance of SNPs on target gene. GeneCopoeia offers Gaussia-luciferase (GLuc), eGFP,or

mCherry based lentiviral or non-viral promoter reporter clones. In addition we can also purchase Gluc vectors that are efficient tools to study transcription regulation. Minimal essential promoter region for each WT target gene will be subcloned from germline DNA using TOPO cloning kit (Invitrogen). If patient sample that harbors the mutation is available, we will amplify the corresponding mutant promoter sequence from the genomic DNA of the patient. PCR products will be cloned upstream to pGL-3-LUC promoter reporter plasmid or upstream to Gluc vectors. For each WT DNA Target gene-promoter plasmid a corresponding MT DNA Target gene-promoter plasmid will be generated using site directed mutagenesis utilizing QuikChange Lightning (Agilent). In this way we will have 30 WT promoter plasmids and 30 MT promoter plasmids in both PGL-3 LUC and Gluc background. We will utilize a panel of adherent cell lines. Cells will be seeded in 6 well plates and transfected with promoter reporter WT and mutant plasmid constructs. 48 hrs after transfection promoter activity will be measured following manufacturer's instructions. Assay values will be normalized using internal renilla luciferase as control.

D-3-b-iii-(2) Functional consequences: using CRISPR/CAS system

We will utilize the newly discovered CRISPR/CAS system (http://www.crispr-cas.org/) to generate endogenous mutations in target genes in a panel of cell lines. 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 (http://zlab.mit.edu/assets/reprints/Wang_H_Cell_2013.pdf) we will directly generate mutations within promoter/enhancers of target genes. Theoretically we generate 30 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. In addition to loss or gain of cognate coding transcript, it is quite conceivable that the SNPs might alter expression of non-coding transcript. To capture the complete influence of rare nominated SNPs at genomic and transcriptomic level we will perform RNA seq. The schematic shown represents representative iterations of plausible genomic changes that will be captured in this validation.



Invivo generation of key SNPs using CRISPR/CAS genome-engineering tools. WT is parental cell line while MT1 and MT2 are mutant cell lines harboring specific SNPs. Comparative transcriptomic analysis between three cell lines representing single or multiple effects of SNPS will be evaluated by RNA-seq (coding and non-coding transcriptome)

Expectations for the influence of variants on biologic functions

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

2.Invitro promoter luc assays will inform us if a particular mutation had any effect on transcription.

3.Promoter/Enhancer analysis using Transfac and other database will provide a comprehensive view of transcription factors that can bind the WT and mutant sequence.

4.Invitro EMSAs will confirm specific binding to WT or mutant sequence by a particular transcription factor.

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

5.Chromatin immune precipitation assays for transcription factors overlapping the SNV will be conducted to determine if the SNV can distort TF binding. This would help validate the SNVs that are predicted to be motif breakers. Alternatively for the SNVs predicted to create a new motif ChIP experiments will help validate binding.

1 Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232-235, doi:10.1126/science.1183621 (2010).
2 Banerjee, S. *et al.* A computational framework discovers new copy number variants with functional importance. *PLoS One* **6**, e17539, doi:10.1371/journal.pone.0017539 (2011).
3 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777, doi:nature08903 [pii]
10.1038/nature08903 (2010).
4 Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772, doi:nature08872 [pii]
10.1038/nature08872 (2010).
5 Schlattl, A., Anders, S., Waszak, S. M., Huber, W. & Korbel, J. O. Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res* **21**, 2004-2013, doi:10.1101/gr.122614.111 (2011).
6 Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-853, doi:315/5813/848 [pii]
10.1126/science.1136678 (2007).
7 Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**, 451-481 (2009).
8 Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987-991 (2009).
9 Shlien, A. *et al.* Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc Natl Acad Sci U S A* **105**, 11264-11269 (2008).
10 Bartsch, G. *et al.* Tyrol Prostate Cancer Demonstration Project: early detection, treatment, outcome, incidence and mortality. *BJU Int* **101**, 809-816, doi:BJU7502 [pii]
10.1111/j.1464-410X.2008.07502.x (2008).
11 Oberaigner, W. *et al.* Reduction of prostate cancer mortality in Tyrol, Austria, after introduction of prostate-specific antigen testing. *Am J Epidemiol* **164**, 376-384, doi:kwj213 [pii]
10.1093/aje/kwj213 (2006).
12 Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**, 437-455 (2010).
13 Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415-425, doi:nrg2779 [pii]
10.1038/nrg2779 (2010).
14 D'Amico, A. V. *et al.* Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *JAMA* **280**, 969-974, doi:joc80111 [pii] (1998).
15 Setlur, S. R. *et al.* Genetic variation of genes involved in dihydrotestosterone metabolism and the risk of prostate cancer. *Cancer Epidemiol Biomarkers Prev* **19**, 229-239, doi:19/1/229 [pii]
10.1158/1055-9965.EPI-09-1018 (2010).
16 Schaefer, G. *et al.* Distinct ERG rearrangement prevalence in prostate cancer: higher frequency in young age and in low PSA prostate cancer. *Prostate cancer and prostatic diseases*, doi:10.1038/pcan.2013.4 (2013).
17 Yu, J. *et al.* An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443-454, doi:S1535-6108(10)00109-1 [pii]
10.1016/j.ccr.2010.03.018 (2010).
18 Consortium, E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247
nature11247 [pii] (2012).