Title:

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

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

# 

# 

# 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

[[MG(28dec)2SK: resp. for this sect. Target overall is ~5 para & 1.5pg]]

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.

[( MG-to-Mark-R: What do you think of “This grant marks a practical collaboration between them attempting to prioritize the millions of rare variants that we can encountering in whole genome sequencing of individuals with cancer. Variants further prioritized by this grant may be followed up by the Rubin lab in greater detail in a more clinical setting.” )]

## B-1 There has been much recent progress in annotating the non-coding genome, thus making it ripe for variant annotation

[[MG(28dec)2SK: short para of 4 sentences… summarizing encode & modencode progress … also lots of RNAseq datasets (Geuvadis project) + chromatin ]]

The Encyclopedia of DNA Elements (ENCODE) Project provides a comprehensive catalogue covering 80.4% of the entire human genome, which is further utilized to understand the genetic landscape of human diseases including cancers \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 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 the mutations to coding portions of the genome. However, relatively little effort has been invested in the investigation of cancer-related disruptions to noncoding portions of the genome. Studies of noncoding cancer disruptions could reveal mutation in regulatory features such as promoters, enhancers, and suppressors, leading to the discovery of new targets for potential cancer therapies. 99% of the somatic mutations are in non-coding regs (funseq ref + other XXX refs).

[[MG2JC PMIDs]]

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 [PMID: 19636342]. Many variants implicated in cancer and diseases have been found to be related to changes in gene expression [PMID: 23348506, 23348503, 7663520, 19165925, 18971308]. As such, there is immense relevance and utility in integrating them into an annotation pipeline that is meant to prioritize disease-causing variants in the human non-coding genome. Mainly, we incorporate regulatory variants into the pipeline in two ways: (1) assigning greater weights to these variants per se, since they are potentially of functional significance, and (2) assigning greater weights if they fall into a type of genomic element or region known to be “highly associated” with these regulatory variants, since rare disease-causing variants will probably not recur in individuals.

## B-3 Rare Variants Are Significant to Study of Cancer & Disease in General

[[MG2(28dec)2SK: can you write up a para on why rare variants ? use text you have. Use Lifton refs + the refs below:

Rare Germline Mutations in PALB2 and Breast Cancer Risk: A Population-Based Study

Rare Germ Line CHEK2 Variants Identified in Breast Cancer Families Encode Proteins That Show Impaired Activation

Germline Mutations in HOXB13 and Prostate-Cancer Risk

Landscape of somatic single-nucleotide and copy-number mutations in uterine serous carcinoma

]]

Genome-wide association studies (GWAS) have provided only a limited insight into causal variants associated with specific common diseases \cite{18987709}. Moreover, growing evidence suggest that rare genetic variants in particular may act as primary drivers of various human diseases, including cancers \cite{11404818}. Based upon their mode of origin, rare variants are often categorized as germline and somatic variants. 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}.

Lots of people have looked at gwas in noncoding [refs]

but not many in rare var

Rare var are very important [refs] but mostly studied in genes!

If you want to study disease common variants which are probed by GWAS really only tell part of hte picture… rare variants are vary improtant ….

Recent high-throughput genomic studies have emphasized the key role of rare genetic variants in various cancers. Based on allele frequency, rare genetic variants are often classified as low frequency polymorphic variants (allele frequency < 5%), sub polymorphic variants (allele frequency 0.1-1.0%) and ‘private’ variants (allele frequency < 0.1%) [Bodmer 2010]. Furthermore, depending upon their mode of origin, they are also categorized as germline and somatic variants.

[[MG(30dec)2SK: Shorten the next bit]]

A population-based study of high-risk women indicated that germline mutations in PALB2 gene enhance the risk of contralateral breast cancer [Tischkowitz 2012]. Bioinformatic and biochemical analysis indicated that rare germline variants delE161 and R117G in the CHEK2 gene lead to an increased risk of breast cancer [Sodha 2006]. Similarly, a targeted sequencing strategy on 17q21-22 region identified rare variant rs138212197 on the HBOX gene associated with prostate cancer [Ewing 2012]. In addition, analysis of whole genome sequencing data of 200 Icelanders identified 252 variants having significant correlation with prostate cancer [Gudmundsson 2012]. Subsequent analysis identified a rare SNP (rs188140481) in the telomeric region of 8q24 locus. Similarly, a genome-wide association study was performed to investigate 16 million SNPS observed among 457 Icelanders. This study implicated rare germline variant rs78378222 in the basal cell carcinoma (BCC) [Stacey 2011].

Rare variants are also observed in the non-coding region of the genome. Multipoint linkage analysis along with high-throughput sequencing was performed to identify germline mutations in the promoter region of the telomerase reverse transcriptase (TERT) gene [Horn 2013] responsible for cutaneous melanoma. A detailed analysis of cutaneous melanoma cell line also indicated presence of somatic mutations in the promoter region of the TERT gene among 74% of sporadic melanoma samples, 85 % of metastatic tissues and 33% of primary melanoma samples. Similarly, whole-genome sequencing was performed to analyze Sonic-Hedgehog medulloblastoma (SHH-MB) tumors among patients containing germline T53 mutation [Rausch 2012]. Further investigations identified 11-24 nonsynonymous somatic mutations in the T53 gene. These somatic and germline mutations in the T53 gene were also associated with the process of chromothripsis (large amount of genomic rearrangement) in the SHH-MB tumor formation. In addition, sequencing of GALNT12 coding exons in colon cancer patients implicated two somatic missense along with six germline mutations, which influenced the GALNT12 enzymatic activity [Guda 2009].

## B-4 Rare variants in Cancer Patients in Similar Elements to Somatic Ones May be Associated with Disease Risk

There has been many indications that for cancer rare variants may be related to the key somatic variants one sees in cancer . The basic idea is that one can have a rare variant affecting the same element as a somatic one and what they want to have a rare variant in one individual and same location as a somatic thing to another individual not rare variant can give first individual and additional proclivity to protect a cancer this has been seen to some degree in with relation to genes and only a look and only to some degree with relation to somatic events…

As X, Y & SK said in their papers, this is a very reasonable thing to do . However, only X did this on non coding variants. (X=tert)

There has been much emphasis on somatic variants in cancer but little on the germline variants. In particular, people haven’t studied the germline variants much in cancer (e.g. the TCGA doesn’t even have an official germline callset). They have mostly focussed on somatic variants.

[[MG(28dec)2SK: fix up above somatic v germline text & integ below refs into the above discussion rel. germline to somatic. (the interplay between rare germline mutations and somatic ones, i.e. germline mutations that occur in nearly the same location as somatic ones might increase a persons risk for cancer. ) Think this should be 1 para. Also look for more refs. ]]

Prior studies have emphasized primarily on identification of somatic variants compared to germline variants in cancer. 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}.

[[MG(28dec)2SK: use 4 refs below + more refs

1) TERT Promoter Mutations in Familial and Sporadic Melanoma

http://www.sciencemag.org/content/339/6122/959.full

Germline DNA from 140 sporadic melanoma cases and 165 sporadic cases were screened to ensure the absence of germline variants in TER genes.

Melanoma cell lines were screened to identify somatic mutations in the promoter region of TERT genes among 74% of sporadic melanoma samples, 85 % of metastatic tissues and 33% of primary melanoma samples.

2) Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations

Presence of T53 germline mutations among LFS-associated SHH-MBs indicates their role in chromothripsis observed for SHH-MBs.

Further analysis to correlate T53 mutation with chromothripsis in other categories of tumor indicated role of T53 somatic mutations and chromothripsis in acute myeloid leukemia.

3) Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers

· In this study, sequencing of GALNT12 coding exons in cohorts of 30-microsatellite stable colon cancer implicated two somatic missense mutations. Expression and biochemical studies indicated that both these somatic mutations completely inactivated the GALNT12 enzymatic activity.

· Furthermore, 6 additional germ-line mutations were also identified to influence the GALNT12 enzyme activity and were implicated in the rise of colon cancer.

4) Association of candidate single nucleotide polymorphisms with somatic mutation of the epidermal growth factor receptor pathway

- Based on their association analysis, they identified two germline SNPs (rs7736074 & rs4975596) at 15p5.33 that associate with somatic mutations of EGFR signaling pathway.

]]

# C Innovation

[[MG(28dec)2EK: Can you write the innovation section? We need 0.5 page]]

# 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

[[MG(28Dec)2LL: can you summarize our ENCODE experience in 2 para or less… use <http://info.gersteinlab.org/Documents> Yip et al. encode enhancers… use what’s on wiki … use PMIDs for refs]]

We have made extensive contributions in the analysis of genomics data. To more effectively utilize the ChIP-seq data for network construction, we developed a method called PeakSeq [PMID: 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 [PMID: 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) [PMID: 20126643]. As published in [PMID: 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 [PMID: 22955619] and modENCODE datasets [PMID: 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 ([PMID: 22125477];[PMID: 22955619]; [PMID: 21177976]; [PMID: 20439753]).

We have developed statistical models of open chromatin associated with gene-expression (Cheng et al., *Nucleic Acids Res*., 2011; Cheng et al., *Genome Biol.,* 2011) .

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

We have developed a ncRNA-finder (Lu et al., *Genome Res.*, 2011). 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 (Habegger et al., *Bioinformatics,* 2011). In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts (Du et al., *PLoS ONE*, 2012). Comparisons between RNA-Seq samples, and to other genomewide data, are be facilitated in part by our Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genome signal tracks (Jee et al., *Bioinformatics*, 2011). We also have extensive experience conducting integrated analyses of large sets of RNA-seq data, such as through the ENCODE project (ENCODE Consortium, *Nature*, 2012). One of such integrative efforts is to characterize genomic variants and selection pressure within ncRNAs (Mu et al.,  *Nucleic Acids Res.*, 2011). We played a lead 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 (Djebali et al., *Nature,* 2012; Gerstein et al., *Science,* 2010). 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 (Gerstein et al., *Nature,* 2012; Cheng et al., *Genome Res.,* 2012).

[[MG(28dec)2XJM: descr incRNA, rseqtools, &c.

1 para or less - 3 sentences

look at

<http://info.gersteinlab.org/Documents>

in particular, exprofile

Preliminary analysis:

\* Selection pressure in ncRNAs, pseudogenes, relative to coding regions

\* Selection pressure in subclasses of ncRNAs and pseudogenes

\* Selection pressure within miRNA structures

]]

#### 

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

[[MG(28dec)2JC: pls flesh out allele prelim. refer to all encode results by pmid]]

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) [PMID: 20567245, 20846943].

We have previously developed a tool, AlleleSeq (PMID: 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 [PMID: 22955620, 22955619, 24092746]. We showed that allele-specific variants are more likely to be rare variants [PMID: 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 [PMID: 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 [PMID: 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 (<http://alleleseq.gersteinlab.org/index.html>). Since then, the resource has been well-received by the scientific community, as exemplified by the number of citations and publications using our data as references (some examples are PMID: 23569280, 24371156). We also constantly update our AlleleSeq tool, with feedback garnered from the scientific community and construct a new GM12878’s personal genome, whenever we could obtain a set of variant calls from the Broad Institute.

#### 

#### **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

[[MG(28dec)2XJM: expl. how ncvar flows into funseq and the encode paper… we need to elaborate slightly and explain sens & ultrasens, which are X, and disruptive mutations and prioritizing network hubs (kim et al. 2007) … 4 sentences total. ]]

We have extensively analyzed patterns of sequence variants in non-coding regions along with the likely protein-coding target genes of these regions (Mu et al., *NAR*, 2011; Yip et al., *Genome Biology*, 2012; Gerstein et al., *Nature*, 2012; Khurana et al., *PLoS Comp Bio*, 2013). We used patterns of genomic variants, such as diversity and fraction of rare variants, to characterize selection pressures of various classes and subclasses of functional annotations (Mu et al., *Nucleic Acids Res.,* 2011). In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region in the ENCODE study (The ENCODE Project Consortium, *Nature,* 2012). Another study by our group showed relations between selection pressures and protein network structures, e.g. hubs and periphery (Kim et al., *Genome Res.,* 2007). In a recent FunSeq study (Khurana et al., *Science*, 2013), 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, prioritize based on network connectivity, and disruptive non-coding mutations.

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 (Khurana et al., S*cience*, 2013). In this study, we integrated large-scale data from various resources, including ENCODE and 1000 Genomes Project data, with cancer genomics data. Using FunSeq, we were able to identify ~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 scheme to be more comprehensive for identification of rare variants associated with complex traits. We will do some simple improvements (i.e. using GERP score in addition to human polymorphisms for identifying sensitive regions) and some major changes outlined below.

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

Loss-of-function variants are more likely to cause deleterious impact (Kheradpour, et al., 2013; Khurana, et al., 2013; Mu, et al., 2011). 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.

[[MG(28dec)2XJM+YF: can we explain the ancestral allele issue in 2 sentences ?]]

#### 

#### D-1-b-ii Identifying likely target genes of distal regulatory elements

[[MG(28dec)2EK: more detail on how doing linkage - 3 sentences]]

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. 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. We will also make the scheme flexible so it can integrate gene expression studies in cases vs controls to increase predictive power for identification of functional variants.

[[MG(30dec)2EK: we should add in something here too on target gene priorization from RNAseq differential expression which is in FunSeqC]]

#### D-1-b-iii More Elaborate Network analysis of variants associated with genes

For all the noncoding variants passing other filters, as in the prototype, 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-iv Detailed Variant Prioritization for ncRNAs

The original FunSeq focused mostly on TF binding sites. Here, we will expand our treatment of ncRNAs beyond that in FunSeq to better prioritze the variants in them. We will first prioritize ncRNA annotations based on their within-human selection pressure and conservation across multiple species. For within-human selection, we will prioritize annotations that show higher levels of nucleotide diversity and higher fraction of rare variants within a broader 200 bp window of the ncRNA annotations (Mu et al., *Nucleic Acids Res.*, 2011). We will look at GERP scores (Goode et al., *Genome Res.,* 2010) for inter-species conservation in each ncRNA region. For those ncRNA broad regions that show strong signatures of selection pressure within humans and between species, we will prioritize them as sensitive regions.

We will further divide ncRNA annotations according to their subcategories, expression levels, and specificity of expression in cell lines. We will categorize ncRNAs into different gene families, including transfer RNAs (tRNAs), miRNAs, 5S ribosomal RNAs (rRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), and long non-coding RNA (lncRNAs). Expression levels of ncRNAs will be obtained from the ENCODE project (http://genome.crg.es/encode\_RNA\_dashboard/hg19/) where RNAseq was performed on dozens of cell lines. We will prioritize ncRNAs that have a higher expression level and those that are ubiquitously expressed in multiple cell lines over cell-line specifically expressed ncRNAs.

Furthermore, we will annotate genomic variants against the secondary structures of ncRNAs. 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 (Steffen et al. 2006). 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.

Finally, we will explore the relationship of ncRNAs with protein 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 searching for 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 (Grimson et al. 2007). We will then examine the selection pressure in ncRNAs that are associated with genes in network hubs vs. periphery.

[[MG(28dec)2XJM:

flesh out for 2-3 para, broader regions, and then disruptive mutations, linking ncRNA to genes (expression pattern correlations or miRNA binding sites), weighting for expression levels, ubiquitous/specific expressions, variations of expression among cell lines

classes of ncRNAs - miRNA, sno, lnc, &c ; further subsplits of by expr & ubiq-v-cellline-spec

sens & ultra sens (now incl. gerp)

disruptive mutations based on struc

assoc w canonical genes (expr correl, &c) => network connectivity

ancestral allele is important (directionality)

Approach:

Functional interpretation of ncRNAs

\* SNP density in ncRNAs

\* SNP allele frequency spectrum in ncRNAs

\* conservation of ncRNA sequences (GERP)

\* Structures of ncRNAs

- predict free energy change of forming secondary structures

- more rigid structures(larger free energy change), such as stem regions, contain fewer number of mutations

- stem regions contain mutations of lower allele frequency

- stem regions are enriched for 'synonymous' mutations, i.e. those that form complementary wobble RNA pairs

- compare allele frequency spectrum of those form wobble pairs to those otherwise

\* Subclasses of ncRNAs:

- For miRNA and tRNA, compare known substructures within

- For snoRNA and snRNA, compare selection pressure between subclasses

- For lncRNA, try to look for novel subclasses or substructures.

]]

#### 

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

[[MG(28dec)2JC: convert to PMID format + shorten.]]

[[MG(28dec)2jc: pls def better allelicity & integrate. How will this exactly be used to prioritize ? We need to come up with a rationale. ]]

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 [PMID: 18308616], histone marks [PMID: 21812971] and random monoallelic expression events [PMID: 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 evident regulatory roles assert that they are useful 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 [PMID: 23128226], in which we are active members. These datasets are found in various publications [PMID: 20220756, 20220758, 21173033, 24136359, 24136358, 24136355] and consortia, notably ENCODE [PMID: 22955616] and gEUVADIS [PMID: 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.

[[JC2MG - the eQTLs part doesnt seem to strengthen this module… can we delete?]]

On a broader scale, we intend to integrate regulatory variants, or quantitative trait loci (QTL), such as Dnase I sensitivity QTLs (dsQTLs), splice QTLs (sQTLs) and expression QTLs (eQTLs). [cite pritchard, montgomery, stranger, yoav, cheung papers] The integration of these variants into the annotation pipeline can be obtained in a similar fashion: by consolidating these variants in a database. QTLs are typically detected solely from indirect statistical correlations between a quantitative trait (such as gene expression) with the allele frequency of a genetic variant in a group of individuals. Hence, the variants detected also have to be common in a population. This is unlike detection of allele-specific variants which are based on direct read count differences between alleles and thus can include rare variants. Therefore, this database will be constructed separately from AlleleDB. We will then aggregate them in a genomic-element-aware fashion for the pipeline.

#### D-1-b-v Develop Unified Weighted-sum scoring scheme to prioritize variants

[[MG(28dec)2EK+YF: we need to elaborate the scoring system here based on text in the paper]]

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 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

#### D-2-a-i We have extensive experience in tool development

The Gerstein lab has 15 years of 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).

[[Sentence #1]] 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.

[[Sentence #2]] 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.

[[Sentance #3]] 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.

[[MG(28dec)2JAS: summarize all of the below in 3 sentences ... see skel. below

To extract knowledge from high-throughput genomic experiments, such as RNA-seq and ChIP-chip the Gerstein lab has made:

-> RSEQtools pipeline to quantify RNA expression and to identify splice sites and gene models

-> IQSeq to identify alternative splicing and exon skipping events

-> Fusion-seq to identify fusion transcripts from paired-end RNA-sequencing

-> ACT to aggregate the distribution of signals in RNA-seq or ChIP-chip signal proﬁles

-> BreakSeq to identify copy-number variation, and unbalanced inversion events

the Gerstein lab has been a pioneer in applying network analysis to generate knowledge form large-scale experiments, such as:

-> TopNet-like Yale Network Analyzer (tYNA) for managing, comparing and mining multiple networks

-> YeastHub to apply semantic web technologies such as resource description framework (RDF), RDF site summary (RSS), relational-database-to-RDF mapping (D2RQ) to more efficiently query life sciences data

]]

#### D-2-a-ii We have extensive experience in calling variants from 1000G, principally SVs

[[MG(28dec)2STL: slightly elaborate this]]

Structural variants are important contributor to human genomic heterogeneity. SVs are common (Redon, 2006) and usually have high impact. Many are associated with inherent diseases or disease susceptibility (Ying, 2013, MacArthur, 2012, Iuliana Ionita-Laza, 2009). Precisely characterizing structural variants, especially accurately locating breakpoints, is a critical yet challenging task. We have demonstrated that precise SV location information provide insightful understanding of the genome (Lam, 2010). We developed a novel approach, BreakSeq, to facilitate the detecting of SVs by comparing raw reads with a breakpoints library (junction mapping). Meanwhile, BreakSeq can effectively infer SV formation mechanism. Indeed, it successfully classifies over 8,000 SVs in the Phase I of 1000 Genomes Project (unpublished data? Ryan, 2011?). Our previous work also involves developing new SV alignment algorithm (Abyzov, 2011) and analysis methods for array-based comparative genomic hybridization approaches (Wang, 2009).

#### 

### D-2-b Research Plan for Aim 2

#### 

#### D-2-b-2 Do SNP & a limited amount of SV calling on all Germline Genomes TCGA and ICGC

[[MG(28dec)2LL: No more than 2 full para in this sect… can fill in how much data, how long approx, currently 600 wGS from prostate + Sanger & the we’re expecting at least 2000 WGS ICGC+TCGA

how to do the germline calls

this a big compute… each run takes XXX time….

we have dbgap & icgc access…]]

[[MG(28dec)2LL: turn the below into text :

\* We need to a pool of rare, disease associated variants to prioritize and validate.

\* Where do we get this?

10K somatic

100K rare

3M total var per person

One idea <scenario 1>

Take the 100K rare ones in a personal genome

500 \* 100K => 50M var.

take these, do funseq & find the top 20 => Mark Rubin

Another idea <scenario 2>

500 indiv. WGS cancer seq. each has 100K rare variants

in tot. they have 50M rare variants

However, that’s a lot .

]]

We currently have access to a combined ~600 whole genome (WG) sequences from the Sanger Institute, and from prostate cancer sequencing projects. We anticipate receiving access to another ~2000 WG sequences from the International Cancer Genome Consortium (ICGC)[PMID: 20393554] and The Cancer Genome Atlas (TCGA)[PMID: 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)[PMID: 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 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. 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-iii We will adapt FunSeq to handle a more complex data context

We will develope 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.

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.

Furthermore, Larva is accessible both a downloadable open source code and as a web server for quick analysis. Each has their advantages and target audiences. For example, the downloadable code is best suited for larger analysis on dedicated servers. the target audience for this would be experienced bioinformaticians with experience installing the required dependencies and access to a large server. On the other hand, the web app is best suited for small quick analysis. the target audience for this would be a translational researcher or clinician with little command line experience and without access to large servers.

[[MG(28dec)2JAS(w/JC+LL): put in stuff from LARVA & alleleDB implementation here (**see above**)]]

[[MG(28dec)2EK: put in stuff from funseq2 about data context]]

#### 

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

[[MG(30dec)2LL: moved Larva to Aim 2. See specific aims to see how larva fits in!!!]]

[[MG(28dec)2LL: fix up larva sect]]

[[MG(28Dec)2LL: 2 more para on background model, More from LL on the background, fix up first para of larva sect. How will we inter-relate germline & somatic… maybe this can be a larva feature? w = log(CDF(g,chrom)) + log(CDF(g,expr) + log….]]

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 [PMID: 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 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 [PMID: 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 [PMID: 22820252]. It also includes whole genome RNA-seq data from the ENCODE project [PMID: 22955616] and SNV density data from the 1000 Genomes Project [PMID: 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-v 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.

[[MG(28dec)2LL:

can fill in how how long approx… larva is the most difficult step

make up diff lists , 1-2 para

funseq prioritized without ref to somatic

funseq prioritized with ref to somatic recurr.

[1para]

MG(28dec)2LL: issues sect… what problems we foresee i nthe implementation and how we’ll get around them… do the comptue on the cloud … paralllel

We’re going to restrict to the rare variants that are in annotations that have recurrent somatic variants.

and then do funseq & find top 20 => Mark Rubin

How do we do this?

take their somatic variants

find recurrant ones in annotations

for each individual we take the ~100K rare germline var in that personal

larva intersect on these to thing

]]

## D-3 Approach Aim 3 - Validate the Prioritized Variants

### D-3-a Preliminary Results Related to Validation

[(MG-to-Mark-R: Can you fill this with ~0.5 pg)]]

### 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 a large cohort with thousands of prostate cancer individuals as controls and also large-scale RNA-seq for this. 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 (eg 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 5000 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 noice ratio and affords a greater flexibility. The Taqman probes are functionally to first ensure assay amplification and optimization for amplification conditions.

#### 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 100 WT promoter plasmids and 100 MT promoter plasmids in both PGL-3 LUC and Gluc background.

##### D-3-b-iii-(2) Functional consequences: Cell based invitro luciferase assays

{{ what’s diff from 1 & 2}}

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-(3) 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 100 SNPs in each cell line and will study functional relevance of these changes compared to WT.

#### D-3-b-iv Expectations for the influcence on variants on on biologic functions

1. Mutant and WT cell lines generated using CRISPR/CAS system will be monitored for growth, proliferation, invasion, migration, senescence, apoptosis using appropriate assay techniques.

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

3. Promoter/Enancer 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

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.

# 

# 

# References

[Genome Res.](http://www.ncbi.nlm.nih.gov/pubmed/23512712#) 2013 May;23(5):800-11. doi: 10.1101/gr.144899.112. Epub 2013 Mar 19.

# Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.

[Kheradpour P](http://www.ncbi.nlm.nih.gov/pubmed?term=Kheradpour%20P%5BAuthor%5D&cauthor=true&cauthor_uid=23512712), [Ernst J](http://www.ncbi.nlm.nih.gov/pubmed?term=Ernst%20J%5BAuthor%5D&cauthor=true&cauthor_uid=23512712), [Melnikov A](http://www.ncbi.nlm.nih.gov/pubmed?term=Melnikov%20A%5BAuthor%5D&cauthor=true&cauthor_uid=23512712), [Rogov P](http://www.ncbi.nlm.nih.gov/pubmed?term=Rogov%20P%5BAuthor%5D&cauthor=true&cauthor_uid=23512712), [Wang L](http://www.ncbi.nlm.nih.gov/pubmed?term=Wang%20L%5BAuthor%5D&cauthor=true&cauthor_uid=23512712), [Zhang X](http://www.ncbi.nlm.nih.gov/pubmed?term=Zhang%20X%5BAuthor%5D&cauthor=true&cauthor_uid=23512712), [Alston J](http://www.ncbi.nlm.nih.gov/pubmed?term=Alston%20J%5BAuthor%5D&cauthor=true&cauthor_uid=23512712), [Mikkelsen TS](http://www.ncbi.nlm.nih.gov/pubmed?term=Mikkelsen%20TS%5BAuthor%5D&cauthor=true&cauthor_uid=23512712), [Kellis M](http://www.ncbi.nlm.nih.gov/pubmed?term=Kellis%20M%5BAuthor%5D&cauthor=true&cauthor_uid=23512712).

# Other useful stuff

<http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-13-013.html>