Title:

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

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

#

#  SV Related points

* Previous studies identified many rare and de-novo structural variants have been found to be associated with many diseases like diabetes [25424174] and neurodegenerative diseases like autism, tourette syndrome, bipolar disorder, and schizophrenia [20531469, 22169095, 24098143, 21658582, 20368508, 21358714]
* We developed CNVnator for identification of copy number variation from individual and trio sequencing datasets. [21324876] and PEMer [2688268] for identifying the complex rearrangements. We will use these to identify the complex rearrangements.

[[MG(2Mar): svpgene too!]]

* We will utilize 1000 genomes to filter for the most likely causal variants.
* We will annotate the structural variants using the ENCODE, REMC and other functional genomics datasets.
* We will develop a novel weighted burden test for incorporating the regulatory (non-coding) effects of the identified structural variants, in addition to the effects on protein coding genes.
* We will use the novel strategy to prioritize the most likely causal variants.

**Specific Aims**

Prioritizing noncoding variants is a subject ripe for exploration with the availability of new noncoding functional annotations 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 rare variants, often not strongly linked to other variants, which may have stronger effects than GWAS SNPs. In particular, we look at rare, germline SNVs (and some deletions and insertions) associated with cancer, trying to prioritize the non-coding variants most associated with disease.

**Aim 1.** Our first aim is to adapt [[& merge]] the pipeline we previously constructed for prioritizing somatic variants (FunSeq) into one for rare germline variants and then to significantly extend its functionality. The existing 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)

[[MG we should call this "regulatry apparatus"]]

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 in many networks (e.g. hubbiness or bottleneckness in both protein-protein interaction and regulatory networks) and differential expression in cancer. (3) In addition to binding sites, we will add noncoding RNA into the pipeline and prioritize it similarly to binding sites -- 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.

**Aim 2.** 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. Finally we will develop weighting schemes to combine all of the FunSeq features coherently together.

#

# **[[SK++: who should edit]]**

# **B Significance**

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

Annotating non-coding regions is essential for investigating genome evolution \cite{16987880}, for understanding important biological functions (including gene regulation and RNA processing) \cite{19148191}, and for elucidating how SNPs and structural variations may influence disease \cite{15549674}. Many projects related to annotating the noncoding genome have recently come to completion. The Encyclopedia of DNA Elements (ENCODE) Project recently provided a comprehensive catalogue covering much of the entire human genome \cite{22955616}. In addition, the model organism ENCODE (modENCODE) Project presents an extensive genomic annotation of drosophila \cite{21177974} and *C. elegans* \cite{21177976} and a way to relate this to human. Furthermore, large-scale mRNA and miRNA sequencing have been applied to elucidate the functional landscape of regulatory variations in the human genome \cite{24037378,20220756,20220758,24092820}. Similar efforts have been directed toward annotating human epigenomic data to investigate underlying disease mechanisms \cite{23482391}. Moreover, the important role of regulatory variants in various diseases have generated a great deal of interest in identifying and annotating the expression of Quantitative Loci linked to specific genes \cite{18597885,20369019}.

## B-2 Non-coding variants, most of which are regulatory, are significant to the study of diseases but less well studied than coding variants

[[MG(10dec)-to-SK: Shorten]]Numerous studies have been conducted on the mutations to coding portions of the genome. However, comparatively less effort has been invested in the investigation of disease-related disruptions to noncoding portions of the genome. Nevertheless, a few[1] initial studies indicate that variants in non-coding regions of genome significantly influence the associated phenotype \cite{17185560} and are often implicated in various diseases\cite{23138309,16728641}. Much of the non-coding variation is contributed by regulatory variants, where cis- and trans-acting variation in the human genome can modulate gene expression \cite{19636342} and this gene expression variation has been implicated in cancer and other diseases \cite{23374354,23348506,23348503,7663520,19165925,18971308}. Specific examples are expression quantitative trait loci (eQTLs) and variants associated with allele-specific behavior. It has been shown that a significant fraction (26%-35%) of inter-individual differences in transcription-factor (TF) binding regions coincides with genetic variation loci and that about 5% of transcripts levels are associated with inherited variant states \cite{20299548}. Genotype-transcript associations have been reported at large for multiple types of inherited variants \cite{21479260,20220756,20220758,21862627,1728997}, however experimental evidence of inherited variants, allele-specific effect on enhancer/promoter activities and transcriptional influence (short and long range) are lacking.

# **D Approach**

## [[MG(2Mar): We have lots of experience on prioritize genes… ALOFT + pgenes + netsnp]]

##

## D-1 Approach Aim 1 - Convert the prototype FunSeq non-coding somatic variant pipeline to prioritize germline variants and elaborate it with new features

### D-1-a Preliminary Results for Aim 1

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

Our proposed work is based on our experience in non-coding annotation. We have made a number of contributions in the analysis of the noncoding genome, as part of our extensive 10-year history with the ENCODE and modENCODE projects. Our TF work includes the development of a method called PeakSeq to define the binding peaks of TFs \cite{19122651}, as well as new machine learning techniques \cite{19015141}. In addition, we have also proposed a probabilistic model, referred to as target identification from profiles (TIP), that identifies a given TF’s target genes based on ChIP-seq data \cite{22039215}. Furthermore, we have developed machine-learning methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers \cite{20126643}, which we have partially validated \cite{22950945}. We have also constructed regulatory networks for human and model organisms based on the ENCODE \cite{22955619} and modENCODE datasets \cite{21430782}, and completed many analyses on them \cite{22125477,21177976,20439753,15145574,14724320,17447836,15372033,19164758,16455753,22955619,22950945,18077332,24092746,23505346,21811232,2160691,21253555}

[[MG(10dec)-to-SK add in cmptxn cmpreg cmppgene]]

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

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

We also have extensive experience conducting integrated analyses of large sets of RNA-seq data, such as through the ENCODE, modENCODE, BrainSpan and exRNA consortia \cite{22955616,22955620,21177976,0000001,0000002}. In particular, for general RNA-Seq analysis, we have developed RSEQtools, a computational package that enables expression quantification of annotated RNAs and identification of splice sites and gene models \cite{21134889}. In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts \cite{22238592}. Comparisons between RNA-Seq samples, and to other genome-wide data, will be facilitated in part by our Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks \cite{21349863}. We have also developed a ncRNA-finder \cite{21177971}. Finally, we have developed statistical models relating gene expression levels to chromatin marks and TF binding \cite{22955619,22955978,22060676,21926158}.

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

[[dec12 JC]]

A specific class of regulatory variants is one that is related to allele-specific events. These are cis-regulatory variants that are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and allele-specific expression (ASE) \cite{20567245,20846943}. We have previously developed a tool, AlleleSeq, \cite{21811232} for the detection of candidate variants associated with ASB and ASE. Using AlleleSeq, we have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project \cite{22955620,22955619,24092746}. Overall, we found that these allelic variants are under differential selection from non-allelic ones \cite{22955619,24092746}. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and expression \cite{22955619}. Furthermore, we have provided the AlleleSeq tool, lists of detected allelic variants, and the constructed personal diploid genome and transcriptome of NA12878 on \cite{0000003}.

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

We have extensively analyzed patterns of variation in non-coding regions along with their coding targets \cite{21596777,22950945,22955619}. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations \cite{21596777}. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region \cite{22955616}. Further studies by our group showed relations between selection and protein network structure, e.g. hubs vs periphery \cite{18077332,23505346}.

In recent studies \cite{24092746,25273974}, we have integrated and extended these methods to develop a prioritization pipeline called FunSeq. FunSeq identifies sensitive and ultra-sensitive regions, i.e. those annotations under strong selection pressure as determined by human population variation. It links each noncoding mutation to target genes and prioritizes them based on scaled network connectivity (compute the percentile after ordering centralities of all genes in a particular network). It identifies deleterious variants in many non-coding functional elements, including transcription-factor (TF) binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitivity sites and detects their disruptiveness of TF binding sites (both loss-of and gain-of function events). It also develops a scoring scheme, taking into account the relative importance of various features, to prioritize mutations. By contrasting patterns of inherited polymorphisms from 1092 humans with somatic variants from cancer patients, FunSeq allows identification of candidate non-coding driver mutations \cite{24092746}. Our method is able to prioritize the known *TERT* promoter driver mutations and scores somatic recurrent mutations higher than non-recurrent ones. In this study, we integrated large-scale data from various resources, including ENCODE and 1000 Genomes Project, with cancer genomics data. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples.[[MG(10dec)-to-YF add funseq2 summary]]

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

We plan to convert the current FunSeq prototype from its focus on somatic variants to allow the identification of rare variants associated with high functional impact. We will do some simple improvements (i.e. incorporating GERP scores and ultra-conserved regions for identifying conserved regions between species) and some major changes outlined below.

#### **[[MG(10dec)-to-YF: shrink GOF]]**

#### **D-1-b-i Identifying gain-of- and loss-of-function mutations for TF binding sites**

Loss-of- and gain-of-function variants are more likely to cause deleterious impact \cite{23512712,24092746,21596777,23348503,23348506,23530248,23887589}. When variants occur in TF 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. Gain-of-motif events are identified as those that give a sequence score with mutated allele in the PWM significantly higher than the background. 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 & then assessing impact of variants on network connectivity**

To interpret likely functional consequences of non-coding variants, we will define associations comprehensively between many non-coding regulatory elements and target protein-coding genes.

As part of ENCODE enhancer prediction group, we are working on predicting confident sets of enhancers in human…

[[Dec-2-ANS]]

We have applied machine-learning methods that integrate multiple genomics features to classify human regulatory regions from ENCODE data of more than 100 transcription factor binding sites. A computational pipeline was developed to identify potential enhancers from regions classified as gene-distal regulatory modules. We are currently developing a new machine learning framework that utilizes epigenomic features and transcription of enhancer RNA (eRNA) to predict active enhancers across different tissues. The effect of sequence variations in these enhancers and eRNAs will be prioritized and their functional impact will be validated experimentally.

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 \cite{20944595}. Then we will identify significant associations between regulatory elements and candidate target genes through computing the correlations of active signals and anti-correlations of inactive signals with gene expression levels across different tissue types.

We will use the regulatory element - target gene pairs to connect the non-coding variants into a variety of networks -- e.g. regulatory network, metabolic pathways, etc. We will examine their network centralities, such as hubs, bottlenecks and hierarchies, as we know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious \cite{23505346,18077332}. [[MG(10dec)2YF: move centrallity more to summary of funseq2]] Moreover, the interpretation of the functional impact of variants can be enhanced if the function of its target protein-coding genes is known. We will incorporate prior knowledge of genes, such as known cancer-driver genes \cite{14993899} and actionable genes (‘druggable’ genes) \cite{22585170} into our annotation scheme. We will also make the scheme flexible so it can integrate gene expression studies in cancer cases vs controls to increase predictive power for identification of functional variants (e.g. using DESeq\cite{20979621}).

#### **D-1-b-iii Detailed variant prioritization for ncRNAs and UTRs – MRS 24Nov2014**

[[MG(2Mar): Move up ncRNA - make seem analogous to genes]]

[[MG(10dec)-to-MRS: rewrite sect]] [[MRS 16Dec – Rewritten]]

To build upon our efforts to prioritize rare variants in noncoding DNA, we will also develop a pipeline for variant prioritization in noncoding RNAs and the untranslated regions (UTRs) of protein-coding genes. We will proceed by (1) Using functional annotations to identify both subregions and short motif features within RNA that are sensitive to mutation and/or evolutionarily conserved; (2) Predicting the ability of variants in sensitive regions to disrupt biochemical activity; (3) Considering features of the whole RNAs within which the RNAs reside; and (4) Using RNA-protein and RNA-miRNA interactions to prioritize variants by their network context. We will integrate the above information to generate predictive scores for the deleteriousness of variants occurring in noncoding RNA regions.

To find key subregions within RNAs, we will focus on footprints of RNA-binding proteins from CLIP-Seq experiments and regions of stable predicted RNA secondary structure. Our preliminary analyses of publicly available CLIP-Seq data indicate that the binding sites of many RNA-binding proteins are more sensitive to mutation than coding sequences, as measured by the proportion of rare variants with low derived allele frequency. Similarly, our secondary structure predictions using RNAShapes have shown that more rigid RNA structures, such as stems, are under higher selection pressure than other RNA regions, and that those variants that incur a larger free energy change of the structures tend to be rarer in human populations.

We will further investigate shorter sequence features that affect RNA regulation, such as miRNA binding sites, polyadenylation signals, and splicing donor and acceptor sites, and chemical RNA modifications. For miRNA binding sites, we will integrate computational predictions using TargetScan, CLIP-Seq datasets for Ago proteins, and CLASH data. For polyadenylation signals, we have used RNA-PET data to show that these regions are substantially more sensitive to mutation on average than protein-coding sequences. [[MRS – need a better segue]]

(2) As a first step toward scoring the effects of variants, we will model their potential to disrupt the biochemical activity of sensitive RNA regions. We will use change in free energy for RNA structure, disruption of PWMs for motifs, and combined sequence-structure models from the tool GraphProt to predict changes in protein-binding.

(3) We will then consider features, such as expression level, breadth of expression, and RNA half-life that apply to whole RNAs. We will develop different scoring schemes for classes of ncRNA with known function, such as miRNA, tRNA, and rRNA; long noncoding RNAs; and the UTRs of protein-coding genes.

(4) Finally, we will interpret the network context of our variants, using RNA molecules as nodes and RNA-protein and miRNA-RNA interactions as edges. We will prioritize variants that are bound by multiple factors, and those within whole RNAs that are bound by many RNA-binding proteins. For UTRs, we will prioritize variants in mRNAs that encode that are essential or whose mutation can lead to disease.

We will integrate the above information to generate predictive scores for the deleteriousness of variants occurring in noncoding RNA regions. [[MRS 16Dec – Need to be more specific about scoring sheme!]]

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[[MG(2mar): discuss outline]]

RNAvar

* Functionally important sites on RNA
	+ CLIP-Seq
	+ Chemical modification
	+ miRNA target sites (predicted + experimental)
* Consensus motifs in RNA biogenesis/processing
	+ Splice sites
	+ Lariat/bridgepoint
	+ Kozak sequence for mRNA?
	+ Transcription cleavage sites (polyadenylation)
* RNA structure
* Whole RNA
	+ Expression
	+ Biotype
	+ Half life
	+ Promoters vs. enhancers
* Networks
	+ Centrality
	+ CLIP-Seq + miRNA

[[MG(2mar): below repeats]]

The original FunSeq focused on sites of TF binding to DNA. Here, we will expand FunSeq to better prioritize variants in ncRNAs and untranslated regions of mRNAs, in a parallel fashion. We will build an integrated framework that considers the wide range of functional genomics data that help characterize RNA molecules, and will investigate specific characteristics of different types of annotated non-coding RNA, e.g. rRNA, miRNA, snRNA, and lncRNA, and the UTRs of mRNA.

We will investigate RNA characteristics at 4 levels, whole gene, transcript isoform, subregions, and motifs/single nucleotides. To determine the importance of each subcategory of RNA (or combination thereof), we will use the fraction of rare variants to measure sensitivity to mutation in humans and GERP score to measure evolutionary conservation between species. Where possible, we will also assess the degree to which mutations disrupt the functional features in which they occur. [[MG ????]]

[[MG a bit rambling, incorp early JZ work - 1st subregions & structure]]

We will begin by investigating categories of sub-regions within RNAs, focusing on protein-binding sites and regions of highly stable secondary structure. The original FunSeq framework is very well suited to analysis of binding sites of RNA-binding proteins to RNA. Indeed, our preliminary data indicate that binding sites of many RNA-binding proteins, determined by CLIP-Seq, are more sensitive to mutation than other RNA regions. The binding sites of some proteins, including FMRP, whose inactivation is the most common single-gene cause of autism, are substantially more sensitive to mutation than CDS regions (Darnell 2011 Cell). To search for regions of stable secondary structure, we will predict RNA structures using RNAShapes. Our preliminary data show 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 human populations.

[[MG other rna feat - expr]]

At the transcript and whole gene level, we will leverage RNA-Seq data from the ENCODE project to categorize RNAs by their expression levels, tissue-, and where data are available, subcellular localization. We will further stratify RNAs by their half-lives in cells using 4SU-Seq data[[MG really??]], whether they interact with ribosomes, using ribosome-profiling data, and by biotype of ncRNA.

[[MG motifs - move this up]]

We will also look at short or single nucleotide annotations that share biological functions. We will investigate key motifs related to the RNA life cycle, such as miRNA binding sites, polyadenylation signals, RNA splicing donor and acceptor sites, and Kozak sequences in the UTRs mRNAs. We will also look at chemical modifications of RNA, such as N6-methylation of adenosine and pseudouridylation.

[[MG net connect]]

Finally, we will investigate the network connectivity of RNAs. We will search for ncRNAs with miRNA binding sites, that might regulate regulate mRNAs that are also targets of the same mRNA. We will also use both connections to RNA-binding proteins and expression clustering and associate ncRNAs with mRNAs of known function. We will investigate whether ncRNAs with high network connectivity are more sensitive to mutation and under higher selection.

To score the effects of specific variants, we will combine the importance of a given RNA feature, as defined by the average sensitivity to mutation and evolutionary conservation of all instances of the feature, and the potential a variant to break the biological activity of the feature. For protein binding sites in RNA, we will search for motifs and generate predictive binding models using GraphProt, and use these models to assess the potential effects of specific mutations (Maticzka 2014 Genome Biology). For structured RNA regions, we will quantify the stabilizing or destabilizing effects of mutations on the RNA structure by computing the difference in folding free energy changes of the RNA before and after the introduction of the mutation. Variants in short sequence features will be scored based on disruption of their motifs, and variants in single base features will be given flat scores. Finally, we will leverage our experience integrating overlapping DNA-binding features in FunSeq2 to develop a framework for integration that allows us to variants in regions with multiple RNA features.

The integration of available functional annotations and structure prediction tools described here will greatly aid investigations of the effects of genetic variants on ncRNA function, RNA regulation, and disease biogenesis.

#### **D-1-b-iv Variant prioritization based on Allelic activity & eQTL association (AlleleDB module)**

[[dec12 JC]]

The evident regulatory roles of the allele-specific variants assert that they will be useful in identifying functional variants. However, to our current knowledge, there is no prioritization scheme that integrates ASB and ASE regulatory variants. One of the main challenges appears to be that allelic variants are enriched for rare variants \cite{24037378}. This implies that a direct overlap of variants in a prioritization pipeline will not be applicable (that is, we would not expect any of the allelic variants to directly overlap the rare variants prioritized by FunSeq.) Moreover, previous analyses restricted by being primarily variant-specific or focused mainly on a single deeply-sequenced individual, GM12878 \cite{22955620,22955619,24092746}. Therefore, to enable the incorporation of allele-specificity into the annotation pipeline, our strategy is to (1) detect allelic variants (both ASB and ASE) from a large pool of individuals and (2) aggregate them into meaningful regions or what we term ‘allelic’ genomic elements.

To detect allelic variants over multiple individuals, we will modify the detection algorithm in the AlleleSeq tool to take into consideration the overdispersion of empirical read distributions observed in RNA-seq and ChIP-seq datasets \cite{25223782,20671027,22499706}. We will then implement the modified AlleleSeq tool on hundreds of genomes. Finally, we will aggregate allelic variants (rare and common) across these individuals into allelic genomic elements and provide an ‘allelicity’ measure for each element, where a greater enrichment of observing allelic variants will result in a higher allelicity score. Because this is also a continuous value, it provides a means for integration into the main prioritization scheme by up-weighting input variants found in allelic genomic elements with higher allelicity scores. Lists of detected ASB and ASE variants and the allelicity scores for various elements will be provided for the scientific community in a public repository, which we called the *AlleleDB*. In addition, a list of ASB variants that are found in the sequence motifs of TF binding sites will be further differentiated by the effects of the variant, i.e. whether the variant causes a loss-of-function, gain-of-function or neutral effect on the TF binding motifs, based on position-weighted matrices of the motifs for each TF.

[[MG(10dec)-TO-JC: one or two lines on new proc strat w over dispersed ... also rel. of allelic to motif breakers]]

[[MG(2mar): Progression: genes, ncRNAs, reg regions]]

[[MG(10dec)-to-JZ: should we incl. the eqtls this time around, maybe on mag, common v rare... 1 sent saying that we'll use upweighted sites form rare eqtl]]

In a similar vein, we also plan to extend this approach to integrate another category of regulatory variants: quantitative trait loci (QTL), such as DNase I hyper sensitivity QTLs (dsQTLs), splice QTLs (sQTLs) and expression QTLs (eQTLs). Rare variants near the common associated loci might be potentially more informative and have phenotypic consequences. For example, scientists already reported rare coding variants in disease genes identified by common variants in type II diabetes that causes familial phenotypes \cite{20581827}. Besides, the pooled burden tests for rare mutations have successfully identify genomic regions with functional rare variants \cite{15297675}. These regions needs to be upweighted during the prioritizing process.

## [[MG we have an initial vers on the web]]

[[MG-probably not this work will be carried out in collab w pcawg 2]]

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

#### **D-2-b-ii Analysis of recurrent 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 or more individuals. However, this is exceedingly unlikely for rare or somatic variants \cite{20981092}. Thus, we will consider mutational burden spread over elements, which include transcribed features, regulatory features, and groups of genes related through a common pathway or protein interaction subnetwork.

Our mutation recurrence discovery procedure has three stages. Given a cancer patient cohort, we will first identify recurrences in the somatic variants. We will then do the same for the rare, germline variants. The third step involves looking for connections between the two sets: elements that contain recurrent somatic variants and rare germline variants imply that the germline variant may be functionally connected with respect to cancer. The absence of common variants from these elements would serve as further evidence for a functional connection to cancer. We have developed a computational framework for identifying these types of recurrent variation, named Large-scale Analysis of Recurrent Variants in 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, recurrently mutated annotations, and recurrently mutated subsets of annotations.

LARVA also uses a new, more accurate model of background somatic mutation in cancer to determine which genome annotations have a significant mutational burden. Many previously developed models have assumed a constant background mutation rate, which gives rise to a binomial distribution for the spread of mutations throughout the genome \cite{25261935}. However, we have found that the mutation rate in cancers is highly heterogeneous between cancer types, between samples of the same cancer type, and between genome regions in the same sample. Hence, we propose modelling the mutation rate as a variable that follows a beta distribution, which gives rise to a beta-binomial distribution for the spread of mutations. With this model, we intend to control the false positive rate that could arise from using the binomial distribution, as we have observed that the distribution of somatic variants in cancer is overdispersed.

Furthermore, LARVA’s significance results incorporate certain regional mutation rate corrections. For example, it has been observed that later-replicating regions during the cell’s S phase are more error-prone due to the depletion of free nucleotides towards the end of replication \cite{20103589}. Hence, our model allows later-replicating regions a higher mutational burden significance threshold. We intend to incorporate other such factors in the near future, such as GC content and RNA-seq expression level.

Hence, for a genome element *i*, LARVA models the number of mutations *xi* in this element, given the number of nucleotides *ni*, the probability *p* of observing a mutation in each position of *i*, the average background mutation rate μ, the dispersion of background mutation rates σ, and the replication timing rank *R* of *i*. LARVA’s model for the expected number of mutations in *i* is:



[[MG \*\*\*\*Should discuss the creation of germ-larva which is paramterized for rare germline. This could be from training on somatic events, however. ]]

In the future, we envision the extension of LARVA’s features in a number of ways. We plan to incorporate additional factors that influence the whole genome somatic mutation rate into LARVA’s null mutation model, such as GC content, chromatin state, and histone modifications, among others. Another useful extension would be the adaptation of LARVA for general purpose GPU (graphics processing unit) computation, allowing the speedup of portions of LARVA that can be optimized for the specialized parallel computations performed on GPUs.

[[MG make this more integral.... remove mention of somatic]]

One particular area of great interest would be the use of LARVA for studying rare germline variants. Some rare germline variants correspond to phenotypes that result in increased susceptibility to certain diseases, instead of giving rise to the disease outright \cite{23011869}, \cite{24759409}. These rare germline variants would probably arise with the same frequency as disease driver mutations, making LARVA ideal for identifying these variants. The identification of both acquired and rare germline variants in a genomic element could serve as a strong indicator of important functional involvement in genetic disease.

#### **[[MG: shorten]]**

#####

##### **D-2-b-iii-2 We will develop a unified weighted scoring scheme for combining all FunSeq modules to consistently prioritize variants**

An integral part of the modular nature of FunSeq will be a way to combine the results of all of the modules into a single variant score and obtain consistent ranking. Different features may contribute differently to the deleterious impact of variants. We will use the mutation patterns observed in the 1000 Genomes polymorphisms to assign weight values to features \cite{25273974}.

In general, features can be classified into two classes: discrete (e.g. “in a particular functional annotation or not”) and continuous (e.g. the PWM change in ‘motif-breaking’). For each discrete feature $d$, we will calculate the probability $p\_{d}$that it overlaps a natural polymorphism. Then we will compute 1-Shannon entropyas its weighted value $w\_{d}$. This measure ranges from 0 to 1 and is monotonically decreasing when $p\_{d}$ is between 0 and 0.5.

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

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

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

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

**D-2-b-iii-3 We will run FunSeq & Larva on all the variants & prioritize them**

We will run FunSeq on the rare variants resulting from our variant calling on all the TCGA/ICGC whole-genome sequences. We expect ~100K variants per genome and also that these variants will recur rarely at the exact same position. Henceforth, we will generate a prioritized list ~200M variants (=100K \* 2000 genomes). We note that unlike GWA studies, which look for association signal, our method prioritizes variants based on functional information. Thus, the variants identified by our pipeline are likely to be the causal ones.

Moreover, we will also prioritize non-coding elements (and not just variant positions), and thus any variants occurring in these functionally important regions are more likely to have an impact. From this pool of ~200M prioritized variants, we will select those in the top quartile that also recur in same element as a somatic variant in another individual, based on LARVA analysis. We will further prioritize variants with germline recurrence in the same element. Overall, this analysis will yield a list of the top 200 variants and elements associated with them. (Note this might not be exactly 200 elements, since it is possible that some of the same variants recur in the same element.) We will select 1000 unique elements from this list and move them on to validation as described below.

# **Resource Sharing Plan**

A comprehensive list of the rare germline variants which are prioritized by our FunSeq pipeline will be made available. Data on variants from 2000 genomes will be submitted to dbGaP in VCF format. Along with the variants themselves, we will provide the important accessory data from the FunSeq ranking. This will include whether the variants lie in sensitive regions, regulatory network hubs, transcription-factor binding sites, and other elements likely to impact phenotype. Also we will notate whether a given variant is a gain- or loss-of-function variant at binding sites. For variants in enhancers or binding sites, we will provide data on their associated gene(s), such as their genomic loci and ensembl gene IDs. We will also provide metrics from the weighting scheme for each variant.

Among the rare variants that we prioritize and to which we assign features, we will publically release about a thousand in flat text files on our FunSeq website (see below). This amounts to less than one variant per genome. As a result, we emphasize that this will not constitute any real concerns regarding privacy -- though we will be modify this if the NIH feels otherwise.

Among these variants, we will experimentally assay about 100, as stated in our proposal. The results gained from the experimental assays for validating variants (such as the TaqMan assays) will be released in the form of flat text files. We plan on releasing this data to the public once the work analyzing the data is published, or three months after the award period ends (whichever is shorter). Again, given that this is substantially less that one variant per genome, this should not constitute a problem as far as privacy is concerned.

A project website developed by our groups will not only host the data on prioritized variants, but will include functionality to enable researchers to query the data using criteria such as genomic region or variant frequency. We note that we have already developed and launched a preliminary version of our FunSeq website (funseq.gersteinlab.org). Any information linking the variants prioritized by our FunSeq pipeline to the genomes from which they were derived will be removed. Of course, information linking the variants to the individuals will not be included in our released variant dataset.

It is likely that investigators may like to tailor the FunSeq pipeline to their own specific needs or workloads. Thus, we plan to make the source code for FunSeq (as well as the associated modules, such allele DB and LARVA) freely available as tar files hosted at github and at the FunSeq website.

Finally, as we point out the grant, FunSeq relies on a complex data context that is derived from many genomic resources. We will make available via the FunSeq website many of these associated annotated files. These include a list of ultra-sensitive regions for TFs and ncRNAs, a list of predicted enhancers, the linkage between enhancers and their target genes, and a set of allelic SNPs and elements.

This could be strengthened by removing "a few". If appropriate.