Specific Aims

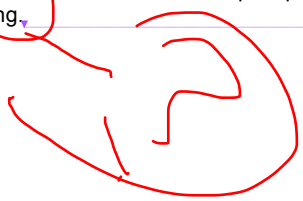
Prioritizing noncoding variants is a subject ripe for exploration with the availability of new noncoding functional annotations from the ENCODE project as well as the many new population-scale functional genomics datasets (e.g. GEUVADIS RNA-seq data). Most of the prioritization up until this point has 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. This work will be carried out by a team comprising of a computational biologist and an experimental cancer genomicist who have worked together for the past decade.

Aim 1. Our first aim is to adapt 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) 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. In the second aim we will develop a large pool of rare variants and then run our elaborated FunSeq pipeline to prioritize them. Our pool of rare variants will result from calling all the germline variants in TCGA and ICGC whole genome sequences (estimated >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 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. Finally we will develop weighting schemes to combine all of the FunSeq features coherently together. Running the elaborated and optimized pipeline on the germline variants will allow us to develop lists of prioritized non-coding elements and variants in them for aim 3.

Aim 3. In this aim we will perform medium throughput validation assays to examine ~300 prioritized variants and tune the parameters of FunSeq [[I am using FunSeq as a placeholder here until we have a new name]] for best predictive performance. Towards these goals, we will perform iterative learning in this aim. We will first clone ~150 variants using our newly-developed massively-parallel Clone-seq pipeline and verify their impact on gene regulation using high-throughput luciferase reporter assays. Based on our functional assay results, we will further tune the parameters of FunSeq to improve its performance and re-run the pipeline on all germline variants. Top ~150 candidate variants will then be cloned and functionally characterized using Clone-seq and luciferase assays to comprehensively validate the performance of FunSeq experimentally.

Aim 4. [[Dimple please edit this aim]] We will then validate by association studies: from 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 look at which of 100 highest ranked FunSeq non-coding elements has rare variants independently recurring in a representative 400 cases by targeted sequencing. Then we will select ~100 recurrent variants for further follow up on 4000 individuals by Taqman assays. Based on recurrence we will select the top third of the variants (~33) to follow up for detailed functional analysis. We will look at how they are associated with downstream differential expression by large-scale RNA-seq followed by RT-PCR validation. Next, we will carry out functional characterization using reporter assays (e.g. luciferase) and the CRISPR/Cas system to generate endogenous mutations and determine their effect on biological functions. Finally, we will use EMSA and chromatin immunoprecipitation to determine how the variant affects transcription-factor binding.



Deleted: tomorrow : new terms , ne ... [1]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [2]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [4]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [5]
Haiyuan Yu 1/6/2015 5:01 PM
Formatted ... [3]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [6]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [7]
Haiyuan Yu 1/6/2015 5:19 PM
Deleted: ... [8]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [9]
Haiyuan Yu 1/6/2015 5:19 PM
Deleted: ... [10]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [11]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [12]
Haiyuan Yu 1/6/2015 5:19 PM
Deleted: ... [13]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [14]
Haiyuan Yu 1/6/2015 5:19 PM
Deleted: ... [15]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [16]
Haiyuan Yu 1/6/2015 5:18 PM
Formatted ... [17]
Haiyuan Yu 1/6/2015 5:18 PM
Formatted ... [18]
Haiyuan Yu 1/6/2015 10:12 PM
Deleted: he
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [19]
Haiyuan Yu 1/6/2015 11:39 PM
Deleted: validate prioritized variants
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [20]
Haiyuan Yu 1/6/2015 10:11 PM
Deleted: .
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [21]
Haiyuan Yu 1/6/2015 10:12 PM
Formatted ... [22]
Haiyuan Yu 1/6/2015 10:11 PM
Deleted: first
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [23]
Haiyuan Yu 1/6/2015 5:16 PM
Deleted:

B Significance

In this proposal we aim to prioritize rare, non-coding variants associated with cancer. 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. Gerstein and Yu have collaborated closely on setting up the Funseq pipeline \cite{24092746}, where Yu led his group to comprehensively examine the impact of nsSNPs on protein interactions experimentally. Gerstein and Yu have been closely collaborating since.

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

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.

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

There have been a large number of GWAS studies \cite{19474294}, which have primarily focused on the identification of common genetic variants. They have neglected the role of rare variants (particularly in noncoding regions) in various diseases \cite{22243964}. However, growing evidence suggests that these rare genetic variants may have strong effects and can act as a primary driver of many human diseases, including cancers \cite{11404818}. Increased disease susceptibility is often attributed to the cumulative effect produced by multiple rare variants \cite{20554195}. For instance, bioinformatic and biochemical analyses indicate that rare germline variants in the CHEK2 gene \cite{16982735} and PALB2 gene increase the risk of breast cancer \cite{22241545}. In addition, a rare variant (rs138212197) in the HBOX gene \cite{22236224} and a rare SNP (rs188140481) in the telomeric region of the 8q24 locus were found to be associated with prostate cancer \cite{23104005}.

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

Haiyuan Yu 1/6/2015 5:16 PM
Deleted: [[intro bullet points what would we change we change : SK]] ... [24]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [25]

Haiyuan Yu 1/6/2015 4:53 PM
Deleted: [[intro bullet points what would we change : SK]] ... [26]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [27]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Space Before: 12 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [28]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Space Before: 12 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Font:(Default) Arial, 11 pt, Font color: Black

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [29]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Space Before: 12 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Font:(Default) Arial, 11 pt, Font color: Black

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [30]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Space Before: 12 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Font:(Default) Arial, 11 pt, Font color: Black

In cancer studies, particularly related to tumor sequencing, prior studies have primarily emphasized the identification of somatic over germline variants. For instance, the current TCGA call sets do not even contain "official" germline calls. However, somatic variants and rare germline have often been observed in the same genetic element in different individuals with the same cancer. The germline variants may increase the risk of cancer in such individuals, and we plan to identify these elements using data on large populations. Multiple experimental studies support this point of view. Germline and somatic mutations in the promoter region of the telomerase reverse transcriptase (TERT) gene have been observed in 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}. In recent years, there has been a growing interest in understanding the contribution of germline and somatic variants in tumor expression \cite{23374354}. Similar studies have been proposed to investigate whether these associations augment the risk of triple-negative breast cancer and prostate cancer among African American populations. \cite{CA165862,CA161032}.

C Innovation

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

C-1 Identifying and interpreting rare non-coding variants associated with increased cancer risk using population-scale polymorphism data

The GWAS catalog contains many common variants associated with diseases. However, as discussed above, many rare variants may increase cancer susceptibility. Currently, no standard methods exist to functionally interpret such variants, especially in non-coding regions. Thus, our approach will be the among the first for functional interpretation of these variants. The 1000 Genomes consortium has created a deep catalog of genetic variation across many populations. Our approach will use the allele frequencies of variants in ~2,500 individuals from 1000 Genomes data to understand which genomic regions are tolerant to common mutations without conferring disease risk. We will then use this knowledge to identify rare variants that may be associated with increased disease risk.

C-2 Using non-coding annotations to understand the likely biological role of non-coding variants

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

C-3 Using knowledge of somatic cancer-causing variants to identify germline variants associated with increase cancer risk

We will use knowledge of somatic variants that potentially function as cancer driver events to identify germline variants associated with increased disease susceptibility. Thus, our approach will be innovative in analyzing somatic and germline variants in an integrative fashion.

C-4 Analyzing variants in ncRNAs

Most previous studies for functional interpretation of noncoding GWAS variants have primarily focused on regulatory regions associated with transcription factor binding sites or regions of open chromatin. Our approach will also analyze impact of variants in ncRNAs and thus this will form another major innovative component of our approach.

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:03 PM
Formatted: Space Before: 12 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Space Before: 12 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted ... [31]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [32]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [33]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [34]

Haiyuan Yu 1/6/2015 5:04 PM
Formatted ... [35]

Haiyuan Yu 1/6/2015 5:04 PM
Formatted ... [36]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [37]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [38]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [39]

Haiyuan Yu 1/6/2015 5:04 PM
Formatted ... [40]

Haiyuan Yu 1/6/2015 5:04 PM
Formatted ... [41]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [42]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [43]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [44]

Haiyuan Yu 1/6/2015 5:04 PM
Formatted ... [45]

Haiyuan Yu 1/6/2015 5:04 PM
Formatted ... [46]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [47]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [48]

C-5 Clone-seq: a massively-parallel site-directed mutagenesis pipeline leveraging next-generation sequencing

Current protocols for site-directed mutagenesis require the selection of individual colonies and subsequent sequencing of each colony using Sanger sequencing to find the correct clone. This standard approach is both labor intensive and expensive, and does not scale well to genome-wide surveys. In Clone-seq, we put single colonies of each mutagenesis attempt into one pool (Fig. xxx) and combine multiple pools through multiplexing for one Illumina sequencing run. Even colonies for generating different mutations of the same gene can be put into the same pool, since they can be easily distinguished computationally when processing the sequencing results. Because our overall PCR-mutagenesis success rate is 80% (**D-3-a-iv**), if we pick 4 colonies for each mutagenesis attempt, the probability of obtaining all desired clones in one pool for 3 instances of the same mutation with identical surrounding sequences is 0.94, rendering the whole pipeline successful. As described in **D-3-a-iv**, we can identify correct clones of ~3,000 mutations in one lane of an Illumina HiSeq run and decrease the cost by more than 10-fold.

C-6 Functionally validating rare variants

Rare variations in regulatory regions of genome can have a paramount influence on biological processes and might function as primer for recurrent somatic mutations in adjacent genomic regions or might contribute to long range changes in chromatin regulation. Using a comprehensive panel of cell lines and genome editing tools like the CRISPR-CAS system we will introduce the rare variations in cell lines and study the effect on cellular physiology. This innovative approach will allow us to generate a catalogue of biological outcomes that can be attributed to a rare variation in a physiological setting.

D Approach

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}

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.

Haiyuan Yu 1/6/2015 5:03 PM
Deleted:

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Font:(Default) Arial, 11 pt, Bold, Font color: Black

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [49]

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Left, Space Before: 12 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [50]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Font:(Default) Arial, 11 pt, Font color: Black

Haiyuan Yu 1/6/2015 5:04 PM
Formatted: Space Before: 12 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [51]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:05 PM
Formatted: Space Before: 12 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:06 PM
Formatted ... [52]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [53]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:07 PM
Formatted ... [54]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [55]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted: Font:1 pt

Haiyuan Yu 1/6/2015 5:06 PM
Deleted:

Haiyuan Yu 1/6/2015 5:06 PM
Formatted: Font:(Default) Arial, 1 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:06 PM
Deleted:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

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

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.

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

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 5:06 PM
Formatted: Space After: 4 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:06 PM
Formatted: Font:(Default) Arial, 11 pt, Font color: Black

Haiyuan Yu 1/6/2015 5:06 PM
Formatted: Space After: 4 pt

Haiyuan Yu 1/6/2015 5:11 PM
Formatted: Font:1 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:06 PM
Formatted: Space After: 4 pt

Haiyuan Yu 1/6/2015 5:06 PM
Formatted: Font:(Default) Arial, 11 pt, Font color: Black

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:11 PM
Formatted: Font:1 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

GERP scores and ultra-conserved regions for identifying conserved regions between species) and some major changes outlined below.

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…

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}. 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

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

Formatted ... [56]
Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [57]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [58]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [59]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [60]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [61]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [62]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [63]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [64]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [65]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [66]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [67]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [68]
Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [69]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [70]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [71]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [72]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [73]
Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [74]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [75]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [76]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [77]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [78]
Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [79]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [80]
Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [81]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [82]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [83]

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.
(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.

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

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

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [84]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [85]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [86]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [87]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [88]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [89]

Haiyuan Yu 1/6/2015 5:13 PM
Deleted:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [90]

Haiyuan Yu 1/6/2015 5:13 PM
Deleted:

Haiyuan Yu 1/6/2015 5:13 PM
Formatted: Space After: 12 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [91]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:13 PM
Deleted:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 5:13 PM
Formatted: Space After: 12 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [92]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

measure evolutionary conservation between species. Where possible, we will also assess the degree to which mutations disrupt the functional features in which they occur.

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.

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 , whether they interact with ribosomes, using ribosome-profiling data, and by biotype of ncRNA.

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.

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)

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.

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [93]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [94]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [95]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [96]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [97]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [98]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [99]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [100]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [101]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [102]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [103]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [104]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [105]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [106]

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.

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.

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

In this aim, we will provide an efficient implementation of FunSeq, including the development of a weighting system to bring together all its features, the calling all the rare germline variants in sequenced tumor genomes, and then running FunSeq on them to develop a prioritized variant and element list. Overall, using FunSeq functional prioritization plus screening out the common variants will allow us to identify the rare variants on a haplotype block with the greatest impact. We cautiously 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 highly likely the causal variants. Furthermore, we will analyze the element-wise recurrence of these rare variants with somatic variants.

D-2-a Preliminary results in developing efficient tools & calling variants on a large-scale

We have significant experience in developing high-throughput tools for bioinformatics research. Our tools take the forms of web services, distributed open source programs, annotation databases and distributed virtual machines. In particular, for the analysis of high-throughput genomic experiments, we have developed pipelines for analysing, RNA expression \cite{12952525,19015660,12952525}, alternative splicing \cite{22238592}, fusion transcripts \cite{20964841}, and copy-number variation \cite{18842824}. We have developed pipelines for the analysis of regulatory networks \cite{14555624,22955619,22125477,14555624} and protein–protein interaction networks \cite{17021160,14724320,22343087,22160691,21826754,21460040}. We have much experience in large-scale germline variant calling through being active members of the 1000 Genomes Consortium, especially the Analysis and Structural Variant (SV) subgroups of the Consortium where the majority of the variant calling tools are developed \cite{21787423,21293372,20981092,23128226}. We have developed a number of SV calling algorithms, including BreakSeq, by comparing raw reads with a breakpoints library (junction mapping) \cite{20037582}, CNVnator, by measuring read depths \cite{21324876}, AGE, by refined local alignment \cite{21233167}, PEMer, for paired ends \cite{19236709}, array-based approaches \cite{19037015} and a sequencing-based bayesian model \cite{21034510}.

D-2-b Research Plan for Aim 2

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

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [107]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [108]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [109]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [110]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted: Font:1 pt

Haiyuan Yu 1/6/2015 5:15 PM
Formatted: Level 2, Space Before: 18 pt, After: 4 pt

Haiyuan Yu 1/6/2015 5:15 PM
Deleted:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 5:11 PM
Formatted: Font:(Default) Arial, 1 pt

Haiyuan Yu 1/6/2015 5:15 PM
Formatted: None, Space Before: 0 pt, After: 0 pt

Haiyuan Yu 1/6/2015 5:15 PM
Deleted:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [111]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [112]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [113]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [114]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:15 PM
Deleted:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [115]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [116]

Haiyuan Yu 1/6/2015 5:11 PM
Formatted ... [117]

We are playing important roles in various pan-cancer analysis working groups (PCAWG) as part of our involvement in the ICGC Project. Within the germline variation group (PCAWG-8), we are identifying key loci associated with cancer susceptibility. The PCAWG-8 group will be generating high-quality germline call sets (comprising SNPs, Indels, and SVs) for relatively high-coverage whole-genome datasets. Furthermore, working with the somatic structural variation group (PCAWG-6), we are identifying and characterizing various balanced and unbalanced structural variations present in different cancers. We will be utilizing these germline variant calls for further downstream analyses.

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.

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.

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.

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 5:15 PM
Deleted:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 5:11 PM
Formatted: Font:1 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 5:11 PM
Formatted: Font:1 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

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

D-2-b-iii-1 We will modularize FunSeq to handle updates to a complex data context & simultaneously carry out efficient production runs

We will develop a practical implementation for all of the new FunSeq modules proposed in aim 1 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 a standalone AlleleDB database. This 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.

Our implementation will allow us to modularize FunSeq into two components: (#1) building a complex data context and (#2) an efficient and high-throughput production run. To build the data context (#1), we will integrate large-scale publicly available data resources, such as polymorphisms from 1000 Genomes project \cite{23128226}, conservation data from Bejerano *et al.* and Cooper *et al.* \cite{15131266,15965027}, functional genomics data from ENCODE \cite{22955616} and Roadmap Epigenomics Mapping Consortium \cite{20944595}. We anticipate this step will be very time-consuming, as we will process large scale genomic data into smaller summary files (e.g. associations between distal regulatory elements and likely target genes). The production run (#2) will prioritize variants from WGS based on the data context. The variant prioritization step needs to be quite efficient, so we can tackle >1000 genomes in fairly short time. The overall modularization offers a flexible framework for users to incorporate the ever-increasing amounts of genomic data to both rebuild the underlying data context and prioritize case-specific variants. We plan to make FunSeq an easy-to-use tool. It will be implemented as a downloadable tool, a web server, and a cloud instance.

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 pdthat it overlaps a natural polymorphism. Then we will compute 1-Shannon entropy as its weighted value wd. This measure ranges from 0 to 1 and is monotonically decreasing when pd is between 0 and 0.5.

$$wd=1+pd*log2pd + (1-pd)*log2(1-pd) \quad (1)$$

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 vc(e.g. PWM change), we will calculate feature weights for each vc. In particular, we will discretize at each value and compute wcvc using (2). Now, when we come to evaluate the continuous feature c for a particular variant, we calculate its weighted value using the actual vc corresponding to the variant.

$$wcvc=1+pcvc*log2pcvc + (1-pcvc)*log2(1-pcvc) \quad (2)$$

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

Haiyuan Yu 1/6/2015 5:14 PM
Deleted:

Haiyuan Yu 1/6/2015 5:11 PM
Formatted: Font:1 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [118]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [119]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [120]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [121]

Haiyuan Yu 1/6/2015 5:14 PM
Deleted:

Haiyuan Yu 1/6/2015 5:11 PM
Formatted: Font:1 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [122]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [123]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [124]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [125]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [126]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [127]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [128]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [129]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [130]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

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.

D-2-b-iii-4 We will tune the prioritization scheme using publicly available dataset and our medium scale functional validation results.

Several high-throughput technologies have been developed to test functional impact of noncoding genomic variants. For example, Kwasnieski et al. used CRE-seq \cite{23129659} to assay > 1,000 single and double nucleotide mutations in promoter regions. Kheradpour et al. \cite{23512712} used MPRA to test variants affecting regulatory motifs in >2,000 human enhancers. The number of validations are limited and not restricted to rare mutations. In addition to these datasets, we plan to validate our prioritized variants from two perspectives: Aim 3 - two-round medium-scale functional validations (~300 mutations in total) using Clone-seq and luciferase assay and Aim 4 - in-depth validation of 6 mutations using technologies, such as CRISP-CAS and mass spectrometry.

With newly available experimental data, we will tune ProVar parameters based on the validation results, as described below. 1) we will tune parameters using CRE-seq and MPRA result and then reproduce the prioritization list. 2) we will select ~150 top variants and do our first-round medium-scale validation, and then re-tune model parameters. 3) In the second round of medium-scale validation, we will select another ~150 top non-tested variants and test them using Clone-seq and luciferase assay.

We plan to tune ProVar parameters using incremental bayesian learning strategy. The probability that a variant

$$P(y=1 \mid W,F) = \frac{\exp(\sum_{i=1}^{m} w_i f_i)}{\exp(\sum_{i=1}^{m} w_i f_i)+1}$$

v is functional is , with F denotes different features of *v* and W is a vector of feature weights. To update W with newly available experimental data, we implement bayesian's rule. $P(W \mid Y,F) \propto P(Y \mid W,F)P(W)$, the probably of observing W given training data Y is proportional to the probability of observing Y given W and F times probability of W. Assuming

$$P(Y \mid W)P(W) = \prod_{i=1}^{N} P(y_i \mid w_1, w_2, ..., w_m, f_1, f_2, ..., f_m) P(w_1, w_2, ..., w_m)$$

independency, , y_i is the experiment data, with y_i=1 means positive results and y_i=0 means negative results. w_m is the weight of feature f_m. We will maximum this function to obtain updated weights W based on training data, using the initial weights W as priors in $P(W)$. The updated weights will then be used as tuned parameters in ProVar to prioritize variants.

D-3 Approach Aim 3 - Medium scale validation of the prioritized variants

D-3-a Preliminary results related to validation

D-3-a-iv Performance, throughput, and cost of our Clone-seq pipeline

To set up our Clone-seq pipeline, we first focused on 27 interactions that can be detected by our version of Y2H, are represented in co-crystal structures, and have known missense disease mutations on their corresponding proteins in HGMD. Of these 27 chosen interactions, 24 have disease mutations on the corresponding interaction interfaces and 15 have mutations away from the interfaces. For interactions that have more than one mutation on and/or away from the interfaces, we randomly picked one for each interaction. To generate these 39 mutant alleles, we picked 4 colonies for each mutation. As a reference, we also pooled together all the WT alleles in our human ORFeome library to be sequenced together with the 4 pools of the mutagenesis colonies. In total, there are 40.1 million Illumina HiSeq 1×100 bp reads for our Clone-seq sample.

These reads were then de-multiplexed and mapped to the genes of interest using the BWA aln algorithm. There is an average of > 2,500× coverage at all desired mutation sites. For each allele of interest, we identified

Haiyuan Yu 1/6/2015 5:11 PM Formatted ... [131]
Haiyuan Yu 1/6/2015 5:11 PM Formatted ... [132]
Haiyuan Yu 1/6/2015 5:14 PM Deleted:
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [133]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [134]
Haiyuan Yu 1/6/2015 5:13 PM Formatted ... [135]
Haiyuan Yu 1/6/2015 5:13 PM Deleted:
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [136]
Haiyuan Yu 1/6/2015 11:59 PM Deleted:
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [137]
Unknown Formatted ... [138]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [139]
Unknown Formatted ... [140]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [141]
Unknown Formatted ... [142]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [143]
Unknown Formatted ... [144]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [145]
Unknown Formatted ... [146]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [147]
Unknown Formatted ... [148]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [149]
Unknown Formatted ... [150]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [151]
Unknown Formatted ... [152]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [153]
Unknown Formatted ... [154]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [155]
Unknown Formatted ... [156]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [157]
Haiyuan Yu 1/6/2015 5:13 PM Formatted ... [158]
Haiyuan Yu 1/6/2015 5:13 PM
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [159]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [160]
Haiyuan Yu 1/6/2015 4:54 PM ... [161]
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [162]
Haiyuan Yu 1/6/2015 5:14 PM
Haiyuan Yu 1/6/2015 4:52 PM Formatted ... [163]

all reads that map to the position of the mutation of interest (Rall) and those that actually contain the desired mutation (Rmut). We then calculated a normalized score that quantifies the fraction of reads that contain the desired mutation:

$$S = Rmut/Rall \times 1/k$$

where k is the number of different mutations for the same gene.

Out of 156 colonies containing the 39 mutations, 125 of them were successful. Thus, our overall PCR-mutagenesis success rate is 80%. In fact, we were able to pick correct clones for all 39 mutant alleles using only the first two pools in Clone-seq. All 78 clones from the first two pools, from which the correct ones used in subsequent steps were selected, were Sanger sequenced for verification. All 55 Clone-seq positive results with S > 0.8 were confirmed, and there is a clear separation in the S scores between the successful and failed clones (Fig. 4). One major advantage of our Clone-seq pipeline is that we can now carefully examine whether there are other unwanted mutations introduced during the PCR process. We found that there are on average 4-5 additional mutations introduced in each pool of the 39 colonies. This corresponds to a 0.013% error rate, in agreement with previous studies. The detection of additional mutations, especially those far away from the mutation of interest, cannot be achieved with the traditional site-directed mutagenesis pipeline using Sanger sequencing. These unintended mutations could lead to erroneous downstream results.

Table 1. Cost comparison between Sanger and Illumina sequencing1.

Traditional Sanger sequencing		Clone-seq	
Unique mutations	3,047	NEBNext Multiplex Oligos (E7335S)	$19.80
Colonies per mutation	4		
Total number of samples	3,047x4=12,188	NEBNext DNA Library Prep Master (E6040S)	$105
Re-sequencing needed2	5%		
Number of 96-well plates needed	137	Illumina HiSeq, single-end, 100 bp sequencing lane	$1,175
Cost per plate	$300		
Minimum cost3	43x300=$12,900	Total cost	$1,299.80
Total cost	137x$300=$41,100		

1All costs are based on internal Cornell pricing.

2Sanger sequencing has an average failure rate of 5%.

3The minimum cost is the least amount of money spent in Sanger sequencing the expected number of samples needed to obtain one correct clone for each mutation of interest.

For our Clone-seq samples, we obtained only 40.1 million reads out of a total of 125 million reads in a single lane of a 1×100 bp HiSeq run with >2,500× coverage. However, to determine S to a least count of 1%, we only need 100× coverage. Since the separation between a successful mutagenesis attempt with the lowest S and an unsuccessful mutagenesis attempt with the highest S is 0.28, 100× coverage makes this separation >25 times our least count. We further increase this separation to >60 times our least count by requiring S > 0.8 for a mutagenesis attempt to be considered successful. 100× coverage is also sufficient for a conservative variant calling pipeline to identify additional unwanted mutations with high confidence35,36. Thus, we can obtain 39×(125/40)×(2,500/100) = 3,047 alleles with a single lane of a 1×100 bp HiSeq run using the Clone-seq pipeline. Overall, our Clone-seq approach will drastically improve the throughput of site-directed mutagenesis and decrease the total cost by at least 10-fold (Table 1).

To further test Clone-seq, we identified a set of 446 SNVs from the published ESP6500 dataset[36] that are at the interface of protein interactions and are amenable to testing using our high-throughput Y2H approach.

Haiyuan Yu 1/6/2015 5:14 PM
Deleted:
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [166]
Haiyuan Yu 1/6/2015 5:14 PM
Formatted ... [167]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [168]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [169]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [170]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [171]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [172]
Haiyuan Yu 1/6/2015 4:58 PM
Formatted ... [173]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted Table ... [174]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [175]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [176]
Haiyuan Yu 1/6/2015 4:58 PM
Formatted ... [177]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [178]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [179]
Haiyuan Yu 1/6/2015 4:58 PM
Formatted ... [180]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [181]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [182]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [183]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [184]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [185]
Haiyuan Yu 1/6/2015 4:58 PM
Formatted ... [186]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [187]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [188]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [189]
Haiyuan Yu 1/6/2015 4:58 PM
Formatted ... [190]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [191]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [192]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [193]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [194]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [195]
Haiyuan Yu 1/6/2015 4:58 PM
Formatted ... [196]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [197]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [198]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [199]
Haiyuan Yu 1/6/2015 4:58 PM
Formatted ... [200]
Haiyuan Yu 1/6/2015 4:59 PM
Formatted ... [202]

Using our Clone-seq pipeline, we performed large-scale, site-directed mutagenesis to generate clones for these 446 SNVs. We sequenced 4 colonies each for the 446 alleles of interest using one full 1×100 bp MiSeq run. We obtained ~14 million reads and aligned them to the reference sequence database using BWA[79]. For each allele of interest, we identified all reads that map to the position of the mutation of interest (R_{all}) and those that contain the desired mutation (R_{mut}). The read coverage surrounding the mutation of interest was ~300× per allele. Using a threshold of $S > 0.8$, approximately 75% of the colonies contain the desired mutation. We were able to choose at least one colony that contains only the desired mutation (without additional unwanted ones) for 437 of the 446 mutagenesis attempts, a success rate of 98.0%.

Overall, our pipeline has been significantly optimized to make it very efficient. We established a web-tool (http://www.yulab.org/Supp/MutPrimer) to design mutagenesis primers both individually and in batch. MutPrimer can design ~1,000 primers for ~500 mutations in one batch in less than one second. All primers for the 476 mutations in this study were generated by MutPrimer. All mutagenesis PCRs are performed in batch using automatic 96-well procedures. Since single colony picking after bacterial transformation of mutagenesis PCR product is a rate-limiting step, we rigorously optimized this step and found that adding 10 µL mutagenesis PCR products to 100 µL competent cells and plating 50 µL transformed cells give the best transformation yield and well-separated single colonies. Furthermore, rather than individually streaking transformed cells onto agar plates one sample at a time, we were able to significantly increase throughput by spreading colonies using glass beads onto four sector agar plates which are partitioned into four non-contacting quadrants. In this manner, a 96-well plate of transformed bacteria can be plated out onto 24 four-sector agar plates in ~15 minutes. Traditional site-directed mutagenesis pipelines require miniprepping each of the selected colonies and sequencing them separately by Sanger sequencing. To drastically improve the throughput of our Clone-seq pipeline, we pooled together the bacteria stock of a single colony for each mutagenesis attempt to perform one single maxiprep, which makes the library construction step much more efficient and amenable to high-throughput. Furthermore, existing variant calling pipelines cannot be applied to our Clone-seq results because the expected allelic ratios built into these pipelines are a function of the ploidy of the organism. However, in our Clone-seq pipeline there is no concept of ploidy. We pool together many mutations for one gene in the same pool (e.g., 40 mutations for *MLH1*) and different genes often have different numbers of mutations. Our *S* score calculation and unwanted mutation detection pipeline was designed according to our pooling strategy .

In total, we have used the Clone-seq pipeline to successfully generate 476 (39 + 437) clones with the desired mutant alleles. The results confirm the scalability, accuracy, and throughput of our Clone-seq pipeline. Through careful considerations, we are confident that this approach can be scaled up to generate the ~1000 SNVs as proposed.

D-3-a-v Reporter luciferase assays confirm validity of in silico TF binding sites

Using an *in silico* approach we determined genome wide distribution of ER in prostate cancer. Intriguingly, we observed a robust recruitment to non-coding genome and identified several intergenic sites that correlated with high ER occupancy. Analysis of recruitment vs transcript profiles confirmed that ER recruitment was associated with productive transcription of long noncoding RNA. Recruitment of ER upstream of NEAT1 lncRNA was addressed in greater details. Reporter assays using promoter luciferase constructs encompassing upstream regulatory regions of NEAT1 and corresponding to two ER binding sites are described in Fig. 9. Interestingly, we discovered that NEAT1 is associated with chromatin and regulates transcription of key prostate cancer genes. Recruitment of NEAT1 was evaluated by ChIP assay and influence on key target genes like PSMA was validated using ChIP and reporter assays (Fig. 10). Functional validation of NEAT1 functions revealed a predominant tumorigenic role as overexpression of NEAT1 was sufficient to augment proliferation, invasion and migratory behavior of prostate cancer cells (Fig. 11).

D-3-b Research plan related to validation

D-3-b-i Overview of validation strategy

Identification of rare variants and understanding the influence thereof on repertoire of biological responses will afford us a unique opportunity to understand causal role of these variations on other somatic mutations associated with diseased states including but not limited to cancer.
We will use Clone-seq to generate ~300 candidate non-coding variant clones identified in Aim 1 and 2. The clones will then be subjected to the downstream reporter assays. Because of the throughput of our Clone-seq

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt, No underline, Font color: Auto

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:59 PM
Formatted: Space Before: 12 pt, After: 2 pt

Haiyuan Yu 1/6/2015 4:59 PM
Deleted:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:59 PM
Deleted:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

and luciferase reporter assays, we will perform iterative learning. That is, we will first clone and test ~150 candidate ncSNVs predicted by our computational learning algorithm. Based on the reporter assay results, we will fine tune the parameters of the learning algorithm, and then perform the predictions again. We will then clone and test another ~150 ncSNVs to confirm the performance of our algorithm. Top candidate ncSNVs that are shown to significantly alter gene expression will be selected for further *in vivo* validations as described in **Aim 4**.

D-3-b-i-(1) ***High-throughput site-directed mutagenesis PCR and E. coli transformation.*** Primers for site-directed mutagenesis are selected based on an optimized version of the protocol accompanying the QuikChange Stratagene site-directed mutagenesis kit (200518). 50 µL mutagenesis PCR reactions are set up on ice in 96-well PCR plates using Phusion polymerase (NEB M0530) according to manufacturer's manual. All WT clones are obtained from the Human ORFeome 8.1[81]. PCR products are digested by *DpnI* (NEB R0176L) overnight at 37 °C (30.5 µL PCR product, 3.5 µL 10× NEBuffer 4, 1 µL *DpnI*). *E. coli* competent cells are prepared in 96-well plates with 20 µL cells per well. 10 µL of *DpnI*-digested PCR products are added to the competent cells using the Tecan robot. After heat shock, 800 µL of SOC recovery medium is added to each well using the Tecan robot and the plate is incubated at 37 °C for 1 hr with vibration. A 20 µL aliquot of the cells is then spotted onto LB + Spectinomycin plates in a fully automated fashion using the Tecan robot. The cells are then spread out in the plates through vigorous shaking with glass beads, as is routinely done in the lab. The plates are incubated overnight at 37 °C. The next day, four colonies per allele are picked for Illumina sequencing. We have already carefully titrated the amount of cells plated so that almost all plates have well-separated single colonies.

D-3-b-i-(2) ***illumina library preparation and HiSeq sequencing.*** DNA plasmids from all four colonies of all alleles are mini prepped using our fully-automated 96-well miniprep pipeline. Four libraries representing one colony of each allele are generated according to Illumina protocols and labeled with distinct barcodes. These four libraries are then mixed into one pool for one 1×100 bp HiSeq run. The *S* score for each colony of each allele is calculated as described above. As shown in **Fig. 4**, we found that all clones with $S > 0.44$ are confirmed to be correct via Sanger sequencing with a clear separation between those that are correct and those that are not. However, to ensure that the clones we pick are correct, we require $S > 0.8$ for a colony to be scored as containing the desired mutation.

D-3-b-i-(3) 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 that can serve as 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-4-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 300 WT promoter plasmids and 300 MT promoter plasmids in both PGL-3 LUC and Gluc background. We will utilize a panel of adherent cell lines. We will use prostate cancer as a model for the validation but we expect that the results will be generalizable to a number of cancers.

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.

Our expectation is that *in vitro* promoter LUC assays will inform us if a particular mutation had any effect on transcription.

D-4 Approach Aim 4 - In-depth functional validation of selected variants

The functional role of prioritized targets will be evaluated using a panel of cell lines that will serve as *in vitro* model to simulate effects *in vivo.* Once tested in cell line model we expect to extend these studies further to animal. We will use prostate cancer as a model for the validation but we expect that the results will be generalizable to a number of cancers.

DON'T REALLY UNDERSTAND

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted: Font:(Default) Arial, 11 pt

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [224]

Dimple Chakravarty 12/30/2014 7:10 PM
Comment [1]: Haiyuan: will you please expand this

Haiyuan Yu 1/6/2015 5:00 PM
Deleted:

Haiyuan Yu 1/6/2015 5:00 PM
Moved down [2]:

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [225]

Haiyuan Yu 1/6/2015 5:00 PM
Moved down [1]: D-3-b-iii Targeted sequencing & Genotyping

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [226]

Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [227]

Haiyuan Yu 1/6/2015 5:15 PM
Deleted:

D-4-a Preliminary results related to validation:

D-4-a-i Capture-Seq identifies rare physiologically relevant mutations

We have applied hybrid capture technology to sequence specific regions with high coverage. Specifically, we have developed a novel targeted next-generation sequencing (NGS) assay, suitable for FFPE and frozen material. The developed protocol is as follows: DNA is extracted from 3x1.5mm FFPE cores, using the Promega Maxwell 16 system. DNA quality is determined using Agilent FFPE derived DNA quality assessment kit in a subset of cases. TruSeq library preparation is obtained using 1µg input DNA. Custom capture is performed using the NimbleGen SeqCap EZ library kit. Paired-end sequencing (2x75bp) is then performed using Illumina HiSeq 2500. Samples are multiplexed (5-7 samples per lane) to ensure a nominal coverage of ~25-40M paired-end reads per samples. Raw sequences are aligned to the human genome reference sequence (GRC37/hg19). This initial mapping is then refined following a series of computational steps to remove potential artifacts and increase the quality of the alignment. We then identify the somatic single nucleotide variants by comparing the tumor against its matching normal. In our study, we analyzed 31 cases of localized prostate cancer. We generated a total of ~1.340B paired-end reads (average per sample ~24.4M; range: 0.97M – 74M). The average coverage per sample is ~177x (range: 3x – 510x). The average capture efficiency is 61.4% (range 45.5% - 70.8%; see Fig. 6). These results suggest that it is feasible to obtain good coverage with archival material with this assay. We were able to identify the known mutations in these samples, including TP53 and SPOP (see Fig. 7), and to nominate some new ones. In this study, we were successful validating genomic alterations in samples up to 10 years old.

D-4-a-ii Low-frequency functionally active intronic & intergenic inherited variants predisposing to cancer

Emerging insights into the genetics of constitutional disease etiology demonstrate that germline polymorphisms are associated with a variety of diseases including Alzheimer's, Parkinson's, mental retardation, autism, schizophrenia \cite{19715442}and cancer \cite{19536264,18685109}. Relevant to this proposal our group recently performed a large scale profiling study for 2,000 individuals from the Tyrol Early Prostate Cancer Detection Program \cite{18321314,16829552}cohort. This cohort is part of a population-based prostate cancer-screening program started in 1993 and intended to evaluate the utility of intensive PSA screening in reducing prostate cancer specific death. By genotyping DNA extracted from peripheral blood samples, we annotated the cohort on more than 5,000 CNVs and 900,000 SNPs and then queried inherited low frequency deletions variants \cite{20059347} for their impact in driving prostate cancer \cite{20479773} and the more aggressive form of the disease \cite{10351184}. We reported on coding and non-coding functionally active risk variants. Among the top hits of the case-control study, an intronic variant in the *Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C (MGAT4C)* demonstrated transcript abundance association with genotype states both in prostate and in lymphoblastoid cells, significant increase in cell and migration upon overexpression in benign and cancer prostate cell lines, and significant decrease in proliferation upon knock down of *MGAT4C* expression with siRNA. In addition, we suggested that intergenic PCA risk variants affect gene regulation through modified transcription factor binding activity of the Activator Protein 1 (AP-1) \cite{20299548,21862627}. Altogether, we demonstrated that inherited variants may directly or indirectly modulate the transcriptome machinery of known oncogenic pathways in prostate cancer facilitating carcinogenesis.

D-4-a-iii In vitro characterization of SNPs within enhancer elements bound by AR and/or ER

The Tyrol Early Prostate Cancer Detection Program cohort is a well characterized cohort with centralized data collection that ensures proper patients' follow-up annotations and availability of well-preserved tissues and blood samples. The cohort currently includes more than 3,000 men. As part of our Trento-Innsbruck-Cornell collaboration, we further studied the genetics of prostate cancer individuals coupling serum levels and genomics data. Specifically, we studied the impact of genetic variants relevant to the metabolism of Dihydrotestosterone \cite{20056642}(DHT), the most potent form of androgen, and investigated the incidence of common genomic rearrangements with respect to PSA levels and age at diagnosis \cite{23381693}.

It has been shown that a significant fraction (26%-35%) of inter-individual differences in transcription factor binding regions coincides with genetic variation loci and that about 5% of transcripts levels are associated with inherited variant states \cite{20299548}. Genotype-transcript associations have been reported at large for multiple types of inherited variants \cite{21479260,20220756,20220758,21862627,17289997}, however experimental evidence of inherited variants allele-specific effect on enhancer activity are lacking. In

order to study the potential role of inherited genetic variants within regulatory elements in the context of hormone dependent human, we have performed an unbiased computational search for AR/ER bound enhancers elements containing SNPs followed by *in vitro* characterization of selected variants. **Table 1** shows counts of SNPs from the dbsnp137 set within AR \cite{20478527} and/or ER (Chakravarty D, *submitted*) binding sites that intersect peak ENCODE data \cite{22955616} generated from 20 cell-lines and ChIP-seq experiments for H3K4m1, H3K4me1+H3K4me3, H3K9ac, H3K27ac, Dnase-seq and FAIRE-seq. For each marker the consensus was generated as the merge of all the regions that are present in at least 2 cell lines and comply with a set of filters. **Fig. 8** shows examples of AR-responsiveness and SNPs impact on putative enhancer elements in MCF7 cells (Garritano S, Demichelis F, *unpublished*).

D-4-b-iv Functional consequences: CRISPR/CAS system

We will utilize the newly discovered CRISPR/CAS system \cite{0000009} to generate endogenous mutations in target genes in a panel of prostate cancer cell lines (VCaP, LnCaP, DU145 and PC3). This unique system will provide us an opportunity to directly modulate endogenous genes and minimize artifacts due to the transfection based reporter assays. Using CRISPR/CAS mediated genome-engineering method \cite{23643243} we will directly generate mutations within promoter/enhancers of target genes. Theoretically we generate 6 individual SNPs in each cell line and will study functional relevance of these changes compared to WT. In case of rare mutations, which occur within both promoter and enhancer regions of the same gene, we will develop cell lines having these combinatorial mutations. Briefly, the CRISPR/Cas9 plasmid (Px459) was obtained from Addgene (Cambridge, MA). Using Ran *et al*(15) protocol we identified a FANCA CRISPR DNA target sequence using algorithms based on analysis in Hsu *et al*(16). The corresponding oligonucleotides were ordered (IDT Coralville, IA) and were cloned into Px459 vector. Sanger sequencing confirmed integration of the FANCA target site into the vector.

Mutations within regulatory regions like promoter and enhancer regions might contribute to one or more biological effects as described in the schematic (Fig. 12). In addition to loss or gain of cognate coding transcript, it is quite conceivable that the SNPs might alter expression of non-coding transcript. To capture the complete influence of rare nominated SNPs at genomic and transcriptomic level we will perform RNA seq. The schematic (Fig. 12) shows representative iterations of plausible genomic changes that will be captured in this validation.

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

D-4-b-(2) Functional consequences: Effect of the mutation on TF binding

In vitro EMSAs will confirm specific binding to WT or mutant sequence by a particular transcription factor. EMSA (electrophoretic mobility shift assay) is a common technique employed to study protein-DNA interaction. We will use the WT and the MT sequences to determine binding to a transcription factor predicted to be present at the site of mutation.

Chromatin immuno-precipitation (ChIP) assays for TFs overlapping the variant will be conducted to determine if the variant can distort TF binding. This would help validate the variants that are predicted to be motif breakers. Alternatively for the SNVs predicted to create a new motif, ChIP experiments will help validate binding.

D-4-b-iii Targeted sequencing & Genotyping

We will conduct the hybrid capture technology (as described in preliminary results) to sequence the top-ranking ~20 elements in 400 samples with high coverage. Custom capture will be performed using the NimbleGen SeqCap EZ library kit followed by paired-end sequencing (2x75bp) using Illumina HiSeq 2500.

Then we will utilize robust Taqman genotyping assays for screening ~10 nominated variants associated with the top-ranked elements in a cohort of 4000 individuals (Tyrol + EDRN, as described above). Superior allelic discrimination is achieved in these assays as they utilize TaqMan minor groove-binding (MGB) probes. This technique generates a low signal to noise ratio and affords a greater flexibility. The Taqman probes are functionally tested to first ensure assay amplification and optimization for amplification conditions.

Methods: Genomic DNA will be extracted from the blood cellular-EDTA samples in a high-throughput fashion using the QIAamp 96 DNA Blood Kit (Qiagen). All DNAs are evaluated by NanoDrop spectrophotometer

Haiyuan Yu 1/6/2015 4:53 PM
Deleted: Need to discuss with M... [228]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [229]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [230]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [231]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [232]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [233]
Haiyuan Yu 1/6/2015 4:57 PM
Deleted: (1)
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [234]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [235]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [236]
Unknown
Field Code Changed ... [237]
Unknown
Field Code Changed ... [238]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [239]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [240]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [241]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [242]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [243]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [244]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [245]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [246]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [247]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [248]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [249]
Haiyuan Yu 1/6/2015 4:55 PM
Formatted ... [250]
Haiyuan Yu 1/6/2015 4:55 PM
Deleted:
Haiyuan Yu 1/6/2015 5:00 PM
Moved (insertion) [1] ... [251]
Haiyuan Yu 1/6/2015 5:00 PM
Deleted: 3

(NanoDrop, Thermo Scientific) and gel electrophoresis (2% agarose). For TaqMan Real-Time Quantitative PCR, each DNA sample will be diluted to 10 ng/ml with nuclease-free water.

D-4-b-iv Evaluation of functional consequence of variants

Based on the Taqman results, we will pick the top ~6 variants for functional follow up.

D-4-b-iii-(1) Functional consequences: RNA-seq

First, we will use RNA-seq. We have RNA-seq data for many members of the cohort. To fill out the dataset, further RNA sequencing will be done on the cases where we see recurrent variants (on up to ~160 individuals). The RNA-seq will be done according to the protocols in \cite{21036922}. This analysis will inform us if a SNP (in promoter or enhancer regions) has any 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-4-b-ii We would determine whether any of the top 10 variants identified in **D-3-b-i** are associated with cancer in a different cohort of individuals or are associated with differential gene expression and RNA-seq. We will use both the Tyrol cohort (described above) and the Early Detection Research Network (EDRN) \cite{0000005} prostate cancer cohort with thousands of prostate cancer individuals as well as normal controls. The prostate cancer cohort include men enrolled at three sites as part of the Prostate Cancer Clinical Validation Center that prospectively enroll individuals at risk for prostate cancer at Beth Israel Deaconess Medical Center (Harvard), at the University of Michigan (Michigan) and at Weill Cornell Medical College (Cornell). Cases are defined as men diagnosed with prostate cancer and controls are men who have undergone prostate needle biopsy without any detectable prostate cancer and no prior history of prostate cancer.
We will first take the highest prioritized variants then subject them to validation. Overall we plan to start the validation pipeline with the top ~20 elements identified from the **D-3-b-i** (as described above). We will perform Hybrid capture assay (described in preliminary data), on 400 cases (patients with cancer) from the above-mentioned cohorts. From the Capture-Seq experiments, we will identify the top 10 recurring variants and subsequently perform TaqMan assays on a further 4,000 cases to see if the precise variants recur in a larger cohort. From this group, we will select top third of the variants (~6), based on recurrence, that we will follow up for detailed functional screening, to be discussed below. This functional screening will be through various reporter assays (e.g. luciferase) looking for the effect on the target gene and also from using the CRISPR/Cas system. For controls, we will utilize deeply sequenced control cohorts (individuals with no cancer) that are already available, including deeply sequenced trios from the 1000 Genomes Project \cite{0000006}, 500 individuals with Complete Genomics sequencing also from 1000 Genomes \cite{0000007} and non-cancerous individual from the UK10K project \cite{0000008}.

Timeline

Year I	**Aim 1: Development of extended Funseq pipeline for annotating noncoding variants** **Aim 2: Optimization & beginning of variant calling** **Aim 3: development of validation assays**
Year II	**Aim 2: Germline variants called from ICGC/TCGA data** **Aim 2: Prioritization of most variants for validation experiments** **Aim 3: Begin functional validation experiments**
Year III	**Aim 2: Finishing prioritization of variants** **Aim 3: Functional annotation of prioritized variants** **Aim 2: Interpreting validation results in light of prioritization**

Deleted: 3
Haiyuan Yu 1/6/2015 5:00 PM
Deleted: 3
Haiyuan Yu 1/6/2015 5:00 PM
Moved (insertion) [2] ... [252]
Haiyuan Yu 1/6/2015 5:00 PM
Deleted: 3
Haiyuan Yu 1/6/2015 5:01 PM
Deleted:
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [253]
Dimple Chakravarty 12/30/2014 7:10 PM
Comment [2]: Haiyuan: are you ... [256]
Haiyuan Yu 1/6/2015 4:55 PM
Deleted: ... [254]
Haiyuan Yu 1/6/2015 4:55 PM
Formatted ... [255]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [257]
Haiyuan Yu 1/6/2015 4:55 PM
Deleted: ... [258]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [259]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [260]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [263]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [264]
Haiyuan Yu 1/6/2015 4:56 PM
Formatted ... [261]
Haiyuan Yu 1/6/2015 4:55 PM
Formatted Table ... [262]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [265]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [266]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [267]
Haiyuan Yu 1/6/2015 4:56 PM
Formatted ... [268]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [269]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [270]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [271]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [272]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [273]
Haiyuan Yu 1/6/2015 4:56 PM
Formatted ... [274]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [275]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [276]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [277]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [278]
Haiyuan Yu 1/6/2015 4:54 PM
... [279]
Haiyuan Yu 1/6/2015 4:52 PM
Formatted ... [280]