***Gerstein Lab Experience in Noncoding Genome Annotation and Variant Prioritization***

***Non-coding genome annotation***

We have developed widely used tools to identify ChIP-Seq peaks \cite{19122651,25292436}, perform RNA-Seq quantification \cite{21134889,22238592}, and identify new non-coding transcripts and categorize them according to function \cite{21177971,25164757}. Our tool to predict enhancer regions has undergone functional validation of its predictions \cite{22950945}. We have further linked enhancers to target genes \cite{25273974} and developed tools to process HiC data \cite{28369339,yan2017mrtadfinder}. We have also built linear and nonlinear models that use epigenetic signals to predict gene expression \cite{22955978,21926158,21324173}. Moreover, we have extensive experience incorporating genomic data into networks to help explaining gene regulation and to identify key regulators \cite{22955619,25249401,27760135,25884877}.

***Experience in genomics consortia***

We have extensive experience in the ENCODE \cite{22955616,22955619,22955620}, modENCODE \cite{25164755,21177976}, 1000 Genomes \cite{20981092,26432246,24092746} and PsychENCODE \cite{26605881} consortia, where we served in a variety of leadership roles (e.g., co-lead of the AWG for modENCODE and leadership of the ENCODE & cancer workgroup)\cite{22955616,25164757,22955619,21177976}. We also have extensive experience analyzing cancer genomes through our participation in The Cancer Genome Atlas (TCGA) and Pan-cancer Analysis of Whole Genomes (PCAWG) consortium. We participated in the TCGA consortium studies of prostate \cite{26544944} and kidney \cite{26536169} cancers and recently conducted a detailed investigation of the non-coding variants in TCGA kidney papillary cancer samples \cite{28358873}. We have also developed tools for somatic variant calling \cite{26381235}. Currently, we are co-leading the PCAWG group investigating the impact of non-coding mutations.

***Non-coding variant prioritization***

We have extensively analyzed patterns of variation in non-coding regions and their coding targets \cite{21596777,22955619,22950945}, integrating these variants with biological networks and other features for prioritization \cite{23505346}. In recent projects \cite{24092746,25273974}, we integrated multiple methods into a comprehensive prioritization pipeline called FunSeq (**Fig. 1**). The pipeline identifies sensitive regions with annotations under high selective pressure, links non-coding mutations to their target genes, and prioritizes variants based on network connectivity. It also identifies deleterious variants in non-coding elements including TF binding sites, enhancers, and regions corresponding to DNase I hypersensitive sites. Using integrated data from large-scale resources (including ENCODE and 1000 Genomes Project) with cancer genomics data, FunSeq can prioritize known TERT promoter driver mutations.

We have further expanded our FunSeq tool into a second version called FunSeq2 that comprehensively defines associations between many noncoding regulatory elements and their target protein-coding genes. To integrate the various features mentioned above, FunSeq2 employs an entropy based scoring scheme \cite{24092746}. In general, features can be classified into two classes: discrete (e.g., within or outside of a given functional annotation) and continuous (e.g., the PWM change in ‘motif-breaking’). We weigh these two sets of features with different strategies.

For each discrete feature $d$, we calculate the probability $p\_{d}$ that it overlaps with common polymorphisms. We then calculate the information content to denote the value of discrete features $s\_{d}=1+p\_{d}\*log\_{2}p\_{d}+(1-p\_{d}) \*log\_{2}(1-p\_{d})$.

The situation is more complex for continuous features, as different feature values have different probabilities of being observed in natural polymorphisms. Thus, one weight cannot suffice for varied feature values. For a continuous feature $c$, which is associated with a value $v\_{c}$, the probability $p\_{c}^{v\_{c}}$ is first estimated using common variants: $p\_{c}^{v\_{c}}=\frac{\#common variant v\geq v\_{c}}{\#common variant}$. The score of continuous feature is defined as $s\_{c}^{v\_{c}}=1+p\_{c}^{v\_{c}}\*log\_{2}p\_{c}^{v\_{c}}+(1-p\_{c}^{v\_{c}}) \*log\_{2}(1-p\_{c}^{v\_{c}})$.

We will run FunSeq2 on all TopMed samples, generating a score for every variant in every individual, genome-wide.

***Allelic analysis***

Allele-specific variants(ASVs) have the potential to provide a highly direct readout of the functional impact of a variant. These are 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 this we have generated comprehensive lists of allelic variants for ENCODE and 1000 Genomes and found that allelic variants are under differential selection from non-allelic ones\cite{22955619,24092746}\cite{22955620,22955619,24092746}. 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 constructed a personal diploid genome and transcriptome of NA12878 \cite{27089393}

 To create a resource describing allele-specific activity, termed AlleleDB, we conducted a comprehensive allele-specific analysis of RNA-Seq and ChIP-seq experiments from multiple disparate studies, such as gEUVADIS\cite{24037378} and ENCODE\cite{22955616}, conducted on 1000 Genomes individuals\cite{23128226, 27089393}. After reprocessing and harmonizing the heterogeneous data, we use the beta-binomial test to remove the effect of overdispersion distribution of dataset and detect the ASV in a uniform way. However, because ASVs are enriched for rare variants, we prioritize by the ‘allelic genomic element’ with the presence of ASVs. Each element will is assigned an ‘allelicity’ score based on not only its enrichment of allelic variants within the element (in comparison to accessible variants within the elements and having sufficient coverage to make an allelic activity call), but also across the number of individuals having allelic variants in a consistent allelic direction. The scoring system by element is useful in two ways: (1) it allows continuous ranking of genomic elements based on its allelic impact across multiple individuals (as opposed to defining a threshold to make a binary decision of whether an element is ‘allelic’) and (2) it enables incorporation of ASE and ASB into the main prioritization scheme; input variants (even those which are rare, but lie in highly-ranked allelic genomic elements) will be up-weighted according to their scores.

We will use AlleleDB to prioritize variants in allelic elements throughout the genomes of TopMed individuals. We will also use TopMed samples with both whole genome sequencing and RNA-Seq data to expand our AlleleDB resource.

***Rare somatic and germline burden tests***

We have worked on statistical methods for analysis of non-coding regulatory regions. LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichments in noncoding elements, by comparing observed mutation counts with expected counts under a whole genome background mutation model \cite{26304545}. LARVA includes corrections for biases in mutation rate owing to DNA replication timing. LARVA can also be targeted exclusively to coding regions to prioritize genes. We used this tool in a pan-cancer analysis of 760 cancer whole genomes’ variants spanning a number of cancer data portals and some published datasets. Our analyses demonstrated that LARVA can recapitulate previously established coding and noncoding cancer drivers, including the TERT and TP53 promoters \cite{26304545}. Furthermore, we have developed MOAT (Mutations Overburdening Annotations Tool), an alternative, empirical mutation burden approach that evaluates mutation enrichments based upon permutations of the input data (submitted). Both annotation-based and variant-based permutation is supported. We will apply these tools to identify recurrent variants and overburdened elements in TopMed samples.