***Gerstein Lab Experience in Variant Prioritization with Focus on Cancer***

***Experience in non-coding genome annotation.*** Our expertise in non-coding DNA variant annotation stems from our experience analyzing a wide variety of genomic assays. 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}. In addition to identifying, quantifying, and linking non-coding genomic elements, we have 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 non-coding variant prioritization.*** We have extensively analyzed patterns of variation in non-coding regions and their coding targets \cite{21596777,22955619,22950945}. 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.

*****Experience in prioritizing protein-coding variants.*** We have developed a variety of tools that prioritize protein-coding variants. Our Variant Annotation Tool (VAT) characterizes variants according to affected genes and transcript isoforms \cite {22743228}, and our Analysis of Loss of Function Transcripts (ALoFT) software predicts loss-of-function (LOF) mutations and their impact \cite{ALOFT}. Relatedly, our netSNP biological network integration tool \cite{23505346} identifies cancer genes based on connectivity. STRESS \cite{27066750} and Frustration \cite{27915290} are two other tools we built to identify mutations that affect allosteric hotspots in proteins and identify key functional protein regions prone to genetic alterations. Finally, our Intensification tool searches for deleterious mutations within repeat regions of proteins \cite{27939289}.

Figure 1. FunSeq2 workflow

***Experience in variant prioritization based on recurrence, taking into account background mutation rate estimation.*** A major approach to finding driver variants starts with searching for mutation-rich genes or genomic regions. However, high mutation heterogeneity and potential correlations between neighboring sites give rise to substantial overdispersion in mutation counts, which complicates background rate estimation. We developed a computational framework called LARVA, which integrates variants with a set of non-coding functional elements to model mutation counts of the elements and handle overdispersion \cite{26304545}. This framework incorporates regional genomic features such as replication timing to better estimate local mutation rates and finds mutational hotspots. We have identified well-known non-coding drivers and uncovered new potential non-coding driver regions after applying LARVA to hundreds of whole-genome tumor sequences.

***Experience in allelic analysis.*** Our AlleleSeq pipeline quantifies allele-specific expression \cite{21811232}, which can provide a direct readout of the effects of allele-specific variants (ASVs). We also conducted a study of allele-specific activity from RNA-Seq and ChIP-Seq experiments conducted on 1000 Genomes Project \cite{23128226,27089393} individuals. After uniformly reprocessing all datasets, including ones from the gEUVADIS \cite{24037378} and ENCODE \cite{22955616}, we detected ASVs using a beta-binomial test to correct for overdispersion. We then combined the effects of multiple ASVs to assign allelicity scores to genomic elements, indicating that these elements are sensitive to mutations \cite{27089393}.

***Experience in genomics and cancer 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 (i.e., co-lead of the AWG for modENCODE and leadership of the ENCODE & cancer workgroup).\cite{22955616,25164757,22955619,21177976}.[[add in psychencode marker paper ref]] 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.