Prioritization of genetic variants in genes:

We have extensive experience in network studies and functional interpretation of coding mutations. Considering diverse gene functions, variants in genes can have a wide spectrum of global effects, ranging from fatal for essential genes to no obvious damaging effect for loss-of-function tolerant genes. The global effect of a coding mutation is largely governed by the diverse biological networks in which the gene participates. We have integrated multiple biological networks to investigated gene functions. We found that functionally significant and highly conserved genes tend to be more central in various networks \cite{23505346} and position on the top level of regulatory networks \cite{22955619}. Incorporating multiple network and evolutionary properties, we have developed a computation method - NetSNP \cite{23505346} to quantify indispensability of each gene in the genome. The method shows its strong potential for interpretation of variants involved in Mendelian diseases and in complex disorders probed by genome-wide association studies.

While NetSNP identifies and ranks genes, we also plan to develop approaches to quantify variant-specific effects. To this end, we developed Variant Annotation Tool ,VAT, vat.gersteinlab.org, to annotate protein sequence changes of mutations. VAT provides transcript-specific annotations and annotates mutations as synonymous, missense, nonsense or splice-site disrupting changes \cite{22743228}. Loss-of-function mutations, the most severe form of coding changes have attracted lots of interest in clinical studies. We have published a paper where we have systematically surveyed LoF variants in a cohort of 180 healthy people as part of the Pilot Phase of the 1000 Genomes project \cite{22344438}. Using linear discrimination analysis, we developed a method to distinguish LoF-containing recessive genes from benign LOF-containing genes. In this grant, we will substantially expand this analysis in an effort to understand the impact of LoF variants. Specifically, we propose to develop methods that will (1) provide variant-specific functional impact scores (2) Distinguish between recessive, dominant and benign variations. Currently, most methods provide a dichotomous classification consisting of benign versus disease. Given that most rare variants are heterozygous, developing methods to differentiate benign rare variants from disease-causing variants in terms of those that can lead to recessive or dominant disease are much needed.

Homologous regions such as pseudogenes give rise to a multitude of problems in variants calling. Errors due to mismapping of short reads derived from pseudogenes to genic regions leads to false variant calls. On the other hand, real variant calls can be missed due to reads being mapped to pseudogenes rather than the true genes\cite{25157971}. To identify pseudogenes in the human genome, we developed PseudoPipe, the first large-scale pipeline for genome wide human pseudogene annotation\cite{16574694}. We also obtained the “high confidence” pseudogenes by combining computational predictions with extensive manual

curation\cite{22951037,25157146}, and identified parent gene sequence from which the pseudogene arises based on their sequence comparisons\cite{22951037}.

Prioritization of genetic variants in ncRNA:

We also have experience analyzing characteristics of ncRNAs. Our incRNA pipeline combines sequence, structural and expression features to classify newly discovered transcriptionally active regions into RNA biotypes such as miRNA, snRNA, tRNA and rRNA\cite{21177971}. Our ncVar pipeline further analyzes genetic variants across biotypes and subregions of ncRNAs, e.g. showing that miRNAs with more predicted targets show higher sensitivity to mutation in the human population \cite{21596777}.

Prioritization of genetic variants in noncoding DNA:

**** We have considerable experience annotating non-coding regulatory regions of the genome

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{22125477,21177976,20439753,15145574,14724320,17447836,15372033,19164758,16455753,22955619,22950945,18077332,24092746,23505346,21811232,21606091,21253555}

**** **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.