**Annotation of non-coding regions of the genome, including both TF-binding sites and non-coding RNAs**

Our proposed work is based on our past experience in non-coding annotation, as part of our 10-year history with the ENCODE and modENCODE projects. Our TF work includes the development of methods to define the binding peaks of TFs prediction of a TF’s target genes and new machine learning techniques. Furthermore, we developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers, which we have partially validated. We also constructed linear and non-linear models that utilize TF binding and histone modification signals to accurately predict the transcriptional output of a gene in different cell types of several organisms including yeast, worm, fly, and human. We have also constructed regulatory networks for human and model organisms and completed many analyses on them. Furthermore, we conducted large-scale multi-organism regulatory and coexpression network comparisons, along with transcriptome and pseudogene lineage analyses. We also have extensive experience conducting integrated analyses of RNA-Seq datasets generated by the ENCODE, modENCODE, BrainSpan and exRNA consortia. In particular, we developed RSEQtools and IQseq for gene model creation and transcript quantification. We also developed tools that specifically analyze features of ncRNAs, including incRNA and ncVAR for finding and characterizing these elements.

**D-1-a-ii We have experience in allelic analyses**

A specific class of regulatory variants is one that is related to allele-specific events. These are variants that are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and allele-specific expression (ASE). We have previously developed a tool, AlleleSeq, 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. 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. Furthermore, we have constructed a personal diploid genome and transcriptome of NA12878.

**D-1-a-iii Experience in relating annotation to variation: the FunSeq pipeline**

We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region. Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery).

In recent studies, we have integrated and extended these methods to develop a prioritization pipeline called FunSeq. It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). FunSeq links each non-coding mutation to target genes, and prioritizes such variants based on scaled network connectivity. It identifies deleterious variants in many non-coding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. It also detects their disruptiveness in TF binding sites (both loss-of and gain-of function events). Integrating large-scale data from various resources (including ENCODE and The 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations, and it scores somatic recurrent mutations higher than those that are non-recurrent. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples. We have also applied our method to investigate non-coding mutation patterns in subtypes of gastric cancer. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.