Annotating regulatory sites, enhancers, and the regulatory network

We have developed a set of tools to identify regulatory elements and analyze regulatory networks. PeakSeq \cite{19122651} and MUSIC \cite{25292436} are efficient ChIP-Seq data processing algorithms that can be used to locate transcription factor (TF) binding sites. In order to establish regulatory relationships between TFs and target genes, we introduced a probabilistic model-based method, TIP (Target Identification from Profiles) \cite{22039215}. We have applied machine-learning methods that integrate multiple genomics features to classify regulatory regions from ENCODE data of >100 TF binding sites. In particular, we were able to identify potential enhancers from regions classified as gene-distal regulatory modules \cite{22950945}. Making use of these potential enhancers, we developed the Function-based Prioritization of Sequence Variants (FunSeq) tool \cite{24092746} for identification of candidate drivers in tumor genomes, and more recently, a more elaborate and flexible framework for this tool. \cite{25273974}

We have developed many approaches for studying biological networks. We developed methods to construct and analyze the regulatory networks of human and model organisms \cite{20439753,21177976,22125477,21430782,22955619} based on ENCODE and modENCODE datasets. We constructed and analyzed a hierarchical regulatory network \cite{22955619,17003135, 21045205, 20523742, 20351254}. Overall, we found that the hierarchy rather than centrality ("hubiness") better reflects the importance of regulators and that in many organisms, including humans, the highest degree of collaboration is between regulators from the middle level. We integrated regulatory networks with gene expression to uncover different types of functional modules \cite{15372033,19723326,12902159, 14555624}. We also introduced several software tools for network analysis including Topnet, \cite{14724320} tYNA \cite{17021160} and PubNet \cite{16168087}.

Relating genomics and chromatin features with gene expression

We have extensive experience on constructing statistical models to predict gene expression levels based on TF binding and HM signals proximal to transcription start site (TSS). \cite{22060676,21177976,21324173,21926158,22955978} For example, we constructed linear and nonlinear models that utilize TF binding signals as input to predict the transcriptional output of a gene \cite{22955978}. We applied these methods on a diverse set of model organisms from yeast to humans \cite{22060676,21177976,21324173,21926158} and achieved high predictive expression levels based on binding signals of 40 TFs in K652 \cite{22955978}. These predictive models revealed several important trends, (i) TF binding and HM signals have comparable predictive accuracy when applied to gene expression; (ii) they are highly redundant in predicting expression \cite{21926158}, (iii) TF binding signals achieve the highest prediction accuracy at the TSS (iv) given the high correlation of different TF signals, only a small number of TFs are required to achieve a good prediction of gene expression. Finally, statistical models trained on protein-coding gene expression can be also utilized to predict non-coding gene expression \cite{21926158} suggesting that non-coding genes share the same regulatory mechanism with protein coding ones.

Comparative Genomics, comparing model organisms to human

Capitalizing on the uniformly processed and matched experimental data obtained by mod/ENCODE consortia, we have performed a series of comparative studies across distant metazoan phyla. A comparative analysis of human, worm, and fly revealed remarkable conservation of general properties of regulatory networks. \cite{25164757} We discovered co-expression modules shared in animals and enriched in their developmental genes. To examine the degree of conservation on how chromatin features affect gene expression, we constructed a 'universal model' for quantitative prediction of coding and non-coding gene expression levels from chromatin features at the promoter. The model is based on a single set of organism-independent parameters and in the three model organisms, achieved accuracy comparable to the organism-specific models. \cite{25164755} We performed a multi-organism comparison of pseudogenes and found that they are much more lineage specific than protein-coding genes, reflecting the different genome remodeling processes in each organism \cite{25157146}. We also 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.\cite{25249401} This integrates co-association networks of individual species utilizing the orthology relationships of genes between species.

Determining SVs and relating variants with annotation

We have a lot of experience in large-scale structural variant calling through being active members of the 1000 Genomes Consortium \cite{21787423,21293372,20981092,23128226}. We have developed a number of SV calling algorithms, including BreakSeq, which compares raw reads with a breakpoint library (junction mapping) \cite{20037582.}, CNVnator, which measures read depth and estimates copy number variation \cite{21324876}, AGE, which refines local alignment \cite{21233167}, PEMer, which uses paired ends \cite{19236709}. We have also developed approaches for quantifying retroduplicaton variation \cite{24026178}, array based approaches to structural variation \cite{19037015} and a sequencing-based bayesian model \cite{21034510}. Applying some of these methods to skin we were able to detect somatic mosaicism \cite{23160490}.

We have also investigated the relationships between networks and variants. We have found that highly conserved genes tend to be more central in interaction and regulatory networks (i.e. more connectivity is associated with more constraint) \cite{22955619, 24092746, 22955620}. Moreover, we examined the impact of adaptive evolution to protein interaction networks, finding proteins under positive selections tend to locate at network periphery \cite{18077332}. We also have demonstrated that networks can be used practically to prioritize the most deleterious variants in cancers \cite{24092746}.

We have previous studied how SNVs and polymorphisms can create cis-regulatory variants that are associated with allele-specific (AS) binding (ASB), particularly of transcription factors or DNA-binding proteins, and AS 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 based on the construction a personal diploid genome sequence (and corresponding personalized gene annotation) using genomic sequence variants (SNPs, indels, and structural variants). Using AlleleSeq, by constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between AS binding and expression \cite{22955619, 22955619, 24092746, 22955620}.