**Non-coding analysis and tools**

We have extensively analyzed patterns of genetic variation in non-coding regions, along with their coding targets. Our work employs metrics such as diversity and fraction of rare variants to characterize selection on various classes and subclasses of functional annotations. In recent studies, we developed a collection of tools that can identify sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations) [1-3]. FunSeq links each noncoding mutation to target genes and prioritizes such variants based on scaled network connectivity [1-2]. This tool 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 hypersensitive sites. FunSeq also detects the disruptiveness of variants in TF binding sites (both LoF and gain-of function events). By 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 score somatic recurrent mutations higher than those that are non-recurrent. To better annotate noncoding regions, we have developed matchedFilter, a signal processing-based enhancer predictor that uses shape-matching filters of histone modification signals inputs and enhancer reporter assays for model training [4]. We also recently developed RADAR, a tool that prioritizes mutations in noncoding RNA regions, primarily using information about RNA-protein interaction [5].

We have developed statistical methods for the analysis of non-coding regulatory regions. LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichment in non-coding elements by comparing observed mutation counts with expected counts under a whole-genome background mutation model [6]. LARVA includes corrections for biases in mutation rate owing to DNA replication timing. LARVA can be targeted to coding regions to prioritize genes. We used this tool in a pan-cancer analysis of variants in 760 cancer whole genomes, spanning a number of cancer data portals and published datasets. Our analyses demonstrated that LARVA can recapitulate previously established coding and non-coding cancer drivers, including the TERT and TP53 promoters. We also developed MOAT (Mutations Overburdening Annotations Tool), an alternative empirical mutation burden approach that evaluates mutation enrichment based upon permutations of the input data. This tool supports both annotation-based and variant-based permutation [7].

We also have extensive experience analyzing cancer genomes through our participation in the ENCODE and Pan-cancer Analysis of Whole Genomes (PCAWG) consortium. We developed a custom annotation by leveraging advanced assays, such as eCLIP, Hi-C, and whole-genome STARR-seq on a number of data-rich ENCODE cell types, a key aspect of which is comprehensive and experimentally derived networks of both transcription factors and RNA-binding proteins (TFs and RBPs) [8]. In addition, we leveraged the comprehensive variant dataset from the PCAWG project to demonstrate that-in addition to the dichotomy of high- and low-impact variants-there is a third group of medium-impact putative passengers [9]. Furthermore, we adapted an additive-effects model from complex-trait studies to show that the aggregated effect of putative passengers, including undetected weak drivers, provides significant additional power (∼12% additive variance)

for predicting cancerous phenotypes, beyond PCAWG-identified driver mutations. Moreover, we performed the first whole-genome analysis of papillary renal-cell carcinoma (pRCC), where we found a germline SNP (rs11762213) in *MET* gene predicting prognosis. We also scrutinized noncoding mutations, discovering potentially impactful ones associated with *MET* [10].

We have developed several tools for analyzing genomic variants from multiple aspects. We built an agnostic machine-learning-based workflow, called SVFX, to assign a “pathogenicity score” to somatic and germline SVs in various cancer types [11]. In particular, we generated somatic and germline training models, which included genomic, epigenomic, and conservation-based features for SV call sets in diseased and healthy individuals. Besides, we developed a method that identifies drivers and quantifies tumor growth that was based solely on the VAF spectrum for an individual sample [12]. Drivers introduce perturbations into the spectrum, and our method uses the frequency of hitchhiking mutations preceding a driver to measure this. Additionally, we reported sigLASSO, a software tool to determine the active mutation signatures from a full repertoire of potential ones, which helped elucidate mechanisms of cancer development [13].

*Reference*

1. Khurana, E., et al., Integrative annotation of variants from 1092 humans: application to cancer genomics. Science, 2013. 342(6154): p. 1235587.
2. Fu, Y., et al., FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol, 2014. 15(10): p. 480.
3. Lou, S. *et al.* GRAM: A GeneRAlized Model to predict the molecular effect of a non-coding variant in a cell-type specific manner. *PLoS Genet* **15**, e1007860, doi:10.1371/journal.pgen.1007860 (2019).
4. Sethi A, et al. A cross-organism framework for supervised enhancer prediction with epigenetic pattern recognition and targeted validation. bioRxiv. 2018.
5. Zhang J, et al. RADAR: annotation and prioritization of variants in the post-transcriptional regulome of RNA-binding proteins. bioRxiv. 2018.
6. Lochovsky, L., et al., LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res, 2015. 43(17): p. 8123-34.
7. Lochovsky, L., et al., MOAT: efficient detection of highly mutated regions with the Mutations Overburdening Annotations Tool. *Bioinformatics* 34.6 (2018): 1031-1033.
8. Zhang, Jing, et al. "An integrative ENCODE resource for cancer genomics." *Nature communications* 11.1 (2020): 1-11.
9. Kumar, Sushant, et al. "Passenger mutations in more than 2,500 cancer genomes: Overall molecular functional impact and consequences." Cell 180.5 (2020): 915-927.
10. Li, Shantao, Brian M. Shuch, and Mark B. Gerstein. "Whole-genome analysis of papillary kidney cancer finds significant noncoding alterations." PLoS genetics 13.3 (2017): e1006685.
11. Kumar, Sushant, et al. "SVFX: a machine-learning framework to quantify the pathogenicity of structural variants." bioRxiv (2019): 739474.
12. Salichos, Leonidas, et al. "Estimating growth patterns and driver effects in tumor evolution from individual samples." Nature communications 11.1 (2020): 1-14.
13. Shantao, Forrest W. Crawford, and Mark B. Gerstein. "Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood." Nature communications 11.1 (2020): 1-12.

**Single-cell analysis and tools**

Single-cell RNA-seq

In collaboration with Prof. Nenad Sestan’s and Flora Vaccarino’s group at Yale, together with groups at USC, the Allen Brain Institute and elsewhere, we analyzed large amounts of RNA-seq data to characterize the transcriptome of the human brain during development. The aim of this project was to create a comprehensive map of gene expression and to understand how the human brain changes throughout life. We have already developed RSEQtools, a suite of tools that performs common tasks on RNA-seq data such as calculating gene expression values, generating signal tracks of mapped reads, and segmenting that signal into actively transcribed regions. This project provides a reference atlas of gene expression in different regions of the brain that will provide valuable information to help interpret neurological function and dysfunction. More recently, this collaborative consortium has started work on PsychENCODE, a project aimed at understanding regulatory variants in the context of their functional connections to psychiatric disease. The project’s approach involves a comprehensive examination of the genome, transcriptome, epigenome, and proteome in relation to brain function.

We have extensive experience with using single-cell RNA-seq data to deconvolve expression data from bulk tissue. In particular, we have successfully deconvolved the bulk RNA-seq data across 1,866 individuals in PsychENCODE and GTEx using single-cell RNA-seq data via non-negative least squares [1].

RNA sequencing (RNA-seq) of sputum samples can be challenging to interpret due to the complex and heterogeneous mixtures of human cells and exogenous (microbial) material. Therefore, we developed a pipeline that integrates dimensionality reduction and statistical modeling to grapple with the heterogeneity [2]. LDA(Latent Dirichlet allocation)-link connects microbes to genes using reduced-dimensionality LDA topics. We validated our method with single-cell RNA-seq and microscopy and then applied it to the sputum of asthmatic patients to find known and novel relationships between microbes and genes.

Single-cell ATAC-seq

To simulate single-cell ATAC-seq (scATAC-seq) experiments with known cell type labels, we have developed an efficient and scalable scATAC-seq simulation method (SCAN-ATAC-Sim) that down-samples bulk-tissue ATAC-seq (e.g. from representative cell lines) in an organized fashion [3]. Our simulation protocol recreates the homogeneous signal-to-noise ratio in a single scATAC-seq experiment by integrating the different amounts of background noise in separate bulk-tissue experiments and independently samples twice without replacement to account for the diploid genome.

*Reference*

1. Wang, Daifeng, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* 362.6420 (2018).
2. Spakowicz, Daniel, et al. Approaches for integrating heterogeneous RNA-seq data reveal cross-talk between microbes and genes in asthmatic patients. *Genome biology* 21.1 (2020): 1-22.
3. Chen, Zhanlin, et al. SCAN-ATAC Sim: a scalable and efficient method to simulate single-cell ATAC-seq from bulk-tissue experiments. *bioRxiv* (2020).