- An integrative scheme for somatic variant prioritization from ENCODE data
- Dissecting mutational landscape in cancer through integrating ENCODE regulatory signals

The overwhelming number of mutations in cancer genomes occurs in the non-coding region of the genome. However, we only really know about the impact of mutations in the fairly small number, approximately 200 cancer genes. The new release of the ENCODE data and annotation allow us to bridge these two disparate facts.

#/********** Figure 1: logic: background rate model -> gene regulatory complex -> novel results predicted by survival analysis **********/
#/***** the following should be one para, just separate from understanding purpose*/

First, we proposed a predictive model by integrating various ENCODE data to precisely calibrate the background mutation rate in multiple cancer types. It could effectively separate the mutational hotspots that are due to well-known genomic confounders (e.g replication timing and chromatin status) from those that are truly associated with tumor genesis.

In addition, to maximize the interpretation of the burdened noncoding regions, we proposed the concept of gene complex by integrating protein coding, proximal, and distal regulatory regions together for our mutational burden analysis. Specifically, for each gene we first defined proximal regulatory regions from the transcription factor binding sites from CHIP-seq experiment. Then we further predicted a list of high confidence distal regulatory regions and their gene targets by integrating CHIP-seq, Enhancer-seq, DNASE-seq, and Hi-C data from the ENCODE project.

Our results show that our integrated scheme could possibly pick up weak mutational signals from various regions and provide more accurate mutation burden analysis. It outperforms simple protein coding gene only analysis by discovering more sensible driver candidates. For example, our scheme does not only predict some well-known drivers such as TP53 and ATM in CLL, but also picked up novel drivers such as BCL6 that is missed by protein coding gene only analysis. Survival analysis showed that the expression level of our novel driver BCL6 could effectively predict CLL progression.

#/********** Figure 2: logic: Rabit -> validation (loregic reguls to be added here!) **********/
Second, integration analysis of eCLIP, CHIP-seq, and RNA-seq data from ENCODE helped to decipher the gene expression regulatory code at transcript resolution and pinpoint the key RNA/DNA binding proteins that is highly associated with tumor specific gene expression. Specifically, we quantified the regulation score for all RNA/DNA binding proteins in multiple cancer types and compared their effects on

gene expression in detail. Our model highly several proteins, like ZNF687 for breast cancer and SUB1 for liver and lung cancer, as key elements that drives the tumor-normal cell differential expression. siRNA RNA-seq experiments were extracted from ENCODE to further validate their effects in corresponding cell lines.

#/********** Figure 3&4: network analysis **********/

Third, we build up a high confidence gene-gene regulation network by integrating both proximal and distal regulation mechanisms. Hierarchical analysis showed that the master transcription factors (TFs) usually demonstrates large correlation with tumor-normal gene expression, and hence influencing gene expression to a larger degree. We also quantified the degree of rewiring for each TF by investigating networks from loosely matched cell lines and prioritized the TFs according to their rewiring score. We discovered a list of highly rewired TFs such as NRF1 and MYC that are highly suspected with tumor genesis. We further divided the target genes into four major groups according to their regulatory status in normal and tumor cell lines. We reasoned that for enhanced and suppressed categories, chromatin status changes are the key driving force for it's regulatory profile; as a contrast, for the static and dynamic targets, other factors such as somatic variants actually plays key roles in their regulatory changes.

#/********** Figure 5: variant prioritization scheme and validation results **********/

Finally, we proposed an integrative scoring workflow to prioritize SNVs in the key elements mentioned above according to their putative deleterious impact. To show the effectiveness of our method, we experimentally validated and characterized a few candidate variants through luciferase assay. Nine out of ten selected variants showed significance negative effect on the downstream gene expression.

.