**An integrative ENCODE resource for interpreting non-coding mutations and gene regulation in cancer**

# Abstract

Most somatic mutations in cancer are non-coding while the characterized drivers are predominantly located in coding regions, creating a conundrum as to whether the non-coding regions are important for oncogenesis. Here we endeavor to create a companion resource to the main ENCODE encyclopedia to address this issue. In particular, we integrate diverse ENCODE data to precisely calibrate background mutation rates and we utilize advanced functional-genomic assays, especially STARR-seq and Hi-C, to develop compact annotations and accurate extended gene models (linking enhancers to coding regions), achieving better statistical power for burden analysis. We also construct detailed regulatory networks to interpret tumor gene expression and mutation profiles, pinpointing effects of key regulators such as the transcription factor MYC and the RNA-binding-protein SUB1 and then validating them. We build cell-type specific networks to directly measure regulatory "rewiring" during oncogenesis, classifying changes as either moving toward or away from a stem-like state. Finally, we integrate the overall ENCODE resource, comprising networks and a compact annotation, to prioritize non-coding elements and mutations and then we validate a subset of them through targeted experiments.

MYC section

As expected we find that the target genes of MYC are significantly up-regulated in numerous cancers, consistent with its well-known role as an oncogenic TF and transcriptional activator \cite{22464321}. We further validated MYC regulatory effects through knockdown experiments (Fig 3). Consistent with our predictions, the expression of MYC targets is significantly reduced after MYC knockdown (Fig 3A). We then used the regulatory network to understand how MYC works with other TFs. We first looked at all triplets involving MYC by requiring that a second TF both interacts and shares a common target with MYC. In all cancer types, we found that MYC’s expression levels are positively correlated with the expression levels of most of its targets, while the second TF shows only limited influence as determined by partial correlation analysis.

We further investigated the exact structure of these regulatory relationships. The most common triplet interaction mode is a well-understood feed-forward loop (FFL) whereby, in this case, MYC regulates both another TF and a common target of both MYC and that TF (Figure 3 C). Since MYC amplification is a major determinant of many cancers, understanding which TFs appear to further amplify MYC effects through FFLs may yield insights for efforts aimed at MYC inhibition (PMC4200208). Most of the FFLs we observed involve well-known MYC partners such as MAX and MXL1. However, we also discovered that many involve another factor NRF1. Upon further study, we found that that the MYC-NRF1 FFL relationships were mostly coherent (i.e., "amplifying" in nature). We further studied these FFLs by organizing these triplets into logic gates, in which the two TFs act as inputs and the target gene expression represents the output \{cite 25884877}. We show that most of these gates follow either OR or MYC-always-dominant logic gate. Thus, the ENCODE regulatory network not only helps find key regulators, but also demonstrates how they work in combination with other regulators.