**An integrative ENCODE resource for cancer genomics: interpreting regulatory changes and non-coding mutations**

# Abstract

Most somatic mutations in cancer are noncoding while the characterized drivers are predominantly located in coding regions, creating a conundrum as to whether the noncoding 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 the diverse ENCODE data to precisely calibrate background mutation rates and synthesize advanced functional-genomic assays, especially STARR-seq and Hi-C, to develop compact annotations and accurate extended gene models (linking enhancers to coding regions) to achieve better statistical power for burden analysis. We also construct detailed regulatory networks to interpret tumor gene expression and mutation profiles, pinpointing and key regulators such as the TF 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 noncoding elements and mutations and then validate a number of them through targeted small-scale studies.