An integrative ENCODE resource for cancer: interpreting non-coding mutations and gene regulation [[correct title???]]

ENCODE comprises thousands of functional genomics data sets, related to numerous cancer types; it is possible to tailor them into a targeted resource for interpreting cancer genomes. In particular, this resource can be used to measure the impact of non-coding mutations, constituting the bulk of the somatic variants. Moreover, by integrating next-generation assays (e.g. STARR-seq) with many epigenetic features, we can significantly refine and make more focused these annotations (beyond a more general genome annotation), increasing the power for recurrent-mutation detection. Second, ENCODE signal data, especially replication timing, allows us to build precise, cancer-matched background models for mutation rates considerably more accurate than previous models. Third, ENCODE data, incorporating new assays, such as Hi-C and RNA-binding protein assays (i.e., eCLIP), in addition to large-scale transcription-factor ChIP-seq, allows the construction of extensive regulatory networks. In some contexts, these networks reveal how connections "rewire" during oncogenesis, as well as how these changes relate to a stem-cell state. More generally, one can use ENCODE networks to prioritize regulators most associated with large-scale expression changes in cancer. Combining the networks with the refined annotations and background mutation models, one can develop a step-wise prioritization scheme for non-coding mutations. Here, we show how this can be instantiated, and we perform a number of small-scale validations (i.e., luciferase assays and shRNA knockdowns) to demonstrate how the resource can prioritize mutations with significant consequences in cancer.