## Introduction

Large-scale functional genomics data are useful for dissecting cancer genomes, particularly for interpreting mutation and expression profiles. Much interpretive work has already been done based on the initial ENCODE release in 2012 and other targeted functional genomic data . In particular, functional genomics data such as from ENCODE data allows us to assign functional impact to non-coding mutations, which form the overall bulk of mutations in the cancer genome. Torchia et al. integrated various genomic and epigenetic signals to identify promising therapeutic targets in rhabdoid tumors. [[fit in better]] Secondly, the ENCODE data, particularly, the replication time data and other signal data is useful for developing background mutation rates, which vary greatly over the genome. For instance, Lawrence et al. incorporated large-scale genomics profiles to identify cancer drivers.[[ref]][[write better]]

ENCODE data is also useful for connecting non-coding elements (such as enhancers or promoters) into regulatory networks, which are pivotal for understanding cancer from a systems-biology perspective. For example, Leiserson et al discovered significantly mutated subnetworks that contain well-known cancer signaling pathways in various cancer types.[[ref]]

The new release of ENCODE data has a number of improvements over the last release, which allows us to construct ENCODEC (a customized “companion ENCODE encyclopedia resource for Cancer”). ENCODEC comprises three parts: a background mutation model, compact annotations, and regulatory networks. We detail each below and provide illustrations of how they may be used to dissect cancer genomes after combining mutation and expression profiles from large cancer cohorts such as TCGA.

In particular, with a much wider selection of ENCODE cell types, the resource provides substantially more precise matching of signal data[[need better word]] to specific cancer types of interest, allowing a demonstrably better background model to be constructed. In addition, for a number of well-known cancer cell types, it incorporates novel assays (such as STARR-Seq) together with a large battery of histone markers to accurately define core enhancers and their targets. It thus constructs a more accurate and compact annotation than generic genome annotation, maximizing statistical power in search of regions with more mutations. Finally, the resource significantly extends TF regulatory networks with considerably more extensive ChIP-Seq coverage and constructs novel networks from RNA-binding profiles and Hi-C data, linking enhancers and promoters. This enables the construction of a more extensive regulatory network. In a few model contexts, this provides cell type-specific networks in both tumor and normal cells, which enables us to directly measure regulatory changes during the normal-to-tumor transition. Furthermore, the stem cell data in ENCODE allows not only a comparison with oncogenesis but also an understanding of the degree to which it relates to a stem-like state. More generally, the ENCODEC network can better explain cancer-specific expression patterns derived from cancer patients, from such resources as TCGA and help pinpoint key regulators that drive large-scale tumor-to-normal expression changes.

We combined the network analysis with the compact annotation and mutational burdening (from the enhanced background model) to propose a step-wise prioritizing scheme. We validated the functional impact of prioritized mutations and elements using small-scale experiments such as shRNA-seq and luciferase assays. We emphasize this as an illustration of how new annotation sets can immediately be used to analyze existing cancer mutation data and cancer expression.