.

## Introduction

Large-scale functional genomics data are useful for dissecting cancer genomes, particularly for interpreting mutation and expression profiles. The initial ENCODE release in 2012, along with other targeted functional genomic data, have motivated a number of integrative studies. These data allow us to assign functional impact to non-coding mutations, which constitute the bulk of mutations in cancer genomes. For instance, Torchia *et al*. integrated various genomic and epigenetic signals to identify promising therapeutic targets in rhabdoid tumors. In addition, ENCODE data sets (especially those related to replication timing and other signals) are useful for estimating background mutation rates (BMR), which vary greatly over the genome. Lawrence *et al*. incorporated genome-wide features, such as replication timing, methylation, and expression profiles, to identify cancer drivers after BMR correction. ENCODE data are 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.

The new release of ENCODE data has a number of improvements over the previous one. The novel features of the most recent data release allow us to construct EN-CODEC (a customized “companion *ENCODE* encyclopedia resource for *C*ancer”) by integrating 2,656 experiments from 14 experimental assays in 229 cell types. It comprises three parts: a background mutation rate model, compact annotations, and regulatory networks. All annotations, models, analysis results and codes are freely accessibly for users (see supplements). We detail each of these parts 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. [JZ2MG: do we need the real URL?]

Specifically, with a much wider selection of ENCODE cell types, our ENCODE companion resource provides substantially more functional genomics data that can be better matched to particular cancer types of interest, allowing a demonstrably improved BMR model. In addition, for a number of well-known cancer cell types, it incorporates various types of novel assays (such as STARR-Seq, HiC, and ChIA-pet) with a large battery of data on histone marks to accurately define core enhancers and their target genes. Consequently, relative to generic annotations, it constructs more compact annotations in a cell type specific way to maximize statistical power in the search of mutationally burdened regions. Finally, our resource significantly extends TF regulatory networks and constructs additional networks from novel assays such as eCLIP and HiC. In a few prominent cancer types, these provide cell type-specific networks in both tumor and normal cells, which enable us to directly measure regulatory changes during the normal-to-tumor transition. Furthermore, the stem cell data in ENCODE enable us to relate such rewiring events as changes that reflect cellular states which become more or less stem-like in nature. More generally, our network can better explain cancer-specific expression patterns in tumors from resources such as TCGA, and it also helps reveal key regulators that drive large-scale tumor-to-normal expression changes.

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

## Multi-level data integration improves variant recurrence analysis in cancer

One of the most powerful ways of identifying key elements in cancer genomes is through mutation recurrence analysis, the objective of which is to discover regions that harbor more mutations than expected. Hence, we demonstrate how to integrate extensive ENCODE data to construct an accurate background mutation rate model in a wide range of cancer types. Accurate BMR estimation is non-trivial – the somatic mutation process can be influenced by numerous confounders (in the form of both external genomic factors and local sequence context factors), and these can result in false conclusions if not appropriately corrected10.