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, some based on the initial ENCODE release in 2012 and other on more targeted functional genomic data sets. In particular, functional genomics data allows us to assign functional impact to non-coding mutations, which constitute the bulk of mutations in cancer genomes. For instance, [[more refs]]Torchia et al. integrated various genomic and epigenetic signals to identify promising therapeutic targets in rhabdoid tumors. Secondly, the ENCODE data sets (especially those related to replication timing and other across-genome "signals") are useful for estimating background mutation rates [[do we really need the BMR label?]], which vary greatly over the genome. [[more ref]]Lawrence et al. incorporated genome-wide features, such as [[repeats]]replication timing, methylation, and expression profiles, to identify cancer drivers after BMR correction. Third, ENCODE data are also useful for connecting non-coding elements (such as enhancers and promoters) into regulatory networks, which are important for understanding cancer from a systems perspective. [[better example]]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 last release, which allows us to construct a customized ENCODE companion resource for Cancer genomics (ENCODEC). This consists of a set of freely distributed annotation files and codes available online (see supplement). It comprises three main parts: a background mutation rate model, compact annotations, and regulatory networks. 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 from the TCGA.

In particular, with a much wider selection of cell types than the previous ENCODE release, ENCODEC provides substantially more functional genomics data that can be better matched to specific cancer types of interest, allowing a demonstrably improved background mutation rate estimation. In addition, for a number of well-known cancer cell types, it incorporates a large battery of data on histone marks with various types of more specialized assays (such as STARR-Seq, HiC, and ChIA-pet) to define core enhancers and their target genes accurately. Consequently, relative to generic annotations, it constructs more compact annotations to maximize statistical power in the determination of mutationally burdened regions. Finally, our resource significantly extends TF regulatory networks with considerably more extensive ChIP-Seq coverage and constructs additional networks from more recent assays such as eCLIP and HiC. For a few prominent cancer types, these provide cell type-specific networks in model tumor and normal cells, enabling direct measurement of potential regulatory changes in oncogenesis. Furthermore, the large amount of stem-cell data in ENCODE enables us to relate such rewiring events as changes that reflect cellular states which become more or less stem-like in nature.[[old sentence w more background — see bottom]] 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 compact annotation sets and mutational burdening (from the enhanced background model) to propose a step-wise prioritizing scheme to highlight key mutations associated with cancer progression. We validated the functional impact of prioritized mutations and elements using focused experiments such as shRNA-seq and luciferase assays. Such prioritization serves as an illustration of how the new ENCODEC resource can immediately be used to help analyze existing cancer mutation data and cancer-associated gene expression.

[[

Also, another data-rich, top- tier cell line is for the human embryonic stem cell (H1-hESC). For decades, a prevailing paradigm has held that at least a subpopulation of tumor cells have the ability to self-renew, differentiate, and regenerate, in a manner that is similar to stem cells9. Hence, H1-hESC can serve as a valuable comparison when investigating the degree to which the oncogenic transformation represents stem-cell-like activities.

]]