# Large-scale ENCODE data integration to interpret regulatory changes in cancer[[[Shirley: discuss title]]]

## Abstract

Cancer is caused by genomic variants that disrupt the normal physiology of living cells. Though mutations in coding genes have been well characterized, the preponderance of mutations in tumor genomes fall within non-coding regions, rendering them less amenable to characterization and interpretation. Newly released ENCODE data provides an opportunity to bridge these knowledge gaps. For a variety of cancer-derived cell lines, as well as non-cancerous cell lines derived from relevant tissues, ENCODE provides diverse genome-wide assays, such as Repli-seq, ChIP-seq, DNase-seq, STARR-seq, Hi-C, and ChIA-PET. The resulting data and functional maps of the human genome provide a framework to assess the potential for non-coding mutations to dysregulate genes, and ultimately cellular regulation.

We first integrated diverse assays from ENCODE to define high-confidence regulatory elements and their associated genes, thereby defining what we call an extended gene neighborhood around each element. We also developed a regression-based method for background mutation rate calibration, which removes confounding effects from chromatin and replication timing. Using this, we then search for genes with higher than expected mutational frequencies within the extended gene neighborhood. This approach was applied to successfully identify novel highly mutated genes that are associated with patient prognosis, such as such as *BCL6* in leukemia.

In addition, we also integrated extensive binding profiles from ENCODE to build tissue-specific regulatory networks for both transcription factors (TFs) and RNA-binding proteins (RBPs). Intriguingly, when analyzing a hierarchical network, we found that TFs with higher mutation burden tend to be located at the bottom of the hierarchy (e.g., EZH2 and NR2C2), whereas those with dysregulated expression tend to reside at the top. Furthermore, by comparing the tumor-derived network with that from normal cells, we identified highly “rewired” TFs with changed targets and prognostic value, such as IKZF1 and MYC. We then extended the tissue-specific network to build generalized networks across cancers. After combining this with expression profiles from other cohorts, we pinpointed MYC and SUB1 as key regulators that significantly drive tumor-vs-normal differential expression. We then validated their effects through knockdown experiments.

Finally, we propose a prioritization scheme for identifying key mutations in cancer. We identified active enhancers (and seven high impact-mutations therein) in breast cancer, and then validated their functional effects through luciferase assays.

## Introduction

A small fraction of mutations associated with cancer have been well characterized, particularly those coding regions of key oncogenes and tumor suppressors. However, the overwhelming bulk of mutations in cancer genomes – particularly those discovered over the course of recent large-scale cancer genomics initiatives – lie within non-coding regions. Whether these mutations drive cancer development or progression, or simply emerge as byproducts of genomic instability remains an open question. [[[add ref]]]

[[JZ2MG: I am not sure why the “However” sentence is deleted. Usually we should say someone has done some job but there is a lot of holes, so we did something else]]

Several recent studies have begun to address this question by incorporating limited functional genomics data for variant interpretation \{cite 25261935, 27064257, 27807102}. For example, Weinhold *et al*. investigated recurrent non-coding mutations in regulatory regions. Wright *et al*. found cancer risk-associated single-nucleotide variants (SNVs) in enhancer regions that potentially upregulate MYC expression in colorectal cancer \{cite 20065031}. Lawrence *et al*. incorporated large-scale genomics profiles to identify cancer drivers \{cite 23770567}.

[[JZ2MG: I know that to use “cancer-relevant ENCODE encyclopedia companion resource” is good in some sense, but they are way too long. I prefer the shorter name C-ENCODE]]

One way to approach the functional interpretation noncoding variants is to experimentally evaluate the functional effects of mutating individual bases. This is a major endeavor of the ENCODE Consortium. In the initial release of the ENCODE annotation five years ago, this was predominantly accomplished using RNA-Seq and ChIP-Seq assays on a limited number of cell lines. [[[ref]]]The new release of ENCODE took two new directions. First, it considerably broadened the number of cell types with the main RNA-Seq and ChIP-Seq assays; the main ENCODE encyclopedia aims to utilize this to provide a general, unified annotation resource applicable across many cells. Secondly, ENCODE expanded the number of sophisticated assays such as STARR-seq, Hi-C, ChIA-pet, eCLIP and RAMPAGE on several top-tier cell lines, many of which are cancer-associated. This enables precise definitions of enhancers, direct identification of enhancer-target gene links, and the construction of RNA-binding protein (RBP) networks. Here, we focus on top-tier cell lines and their relevance to cancer by performing large-scale integration of these various assays to construct a in-depth companion resource to the general encyclopedia. We call this the “cancer-relevant ENCODE encyclopedia companion resource” (or the “encyclopedia companion” for short) for interpreting the wealth of mutational and transcriptional profiles produced by the cancer research community. [[[Shirley: need to discuss the name, brainstorm]]]

## Comprehensive functional characterization by ENCODE data integration[[[Shirley: we need a better head]]

The ENCODE top-tier cell lines provide good models not only for studying gene regulation in detail, but also for understanding cancers of the blood (K562), breast (MCF-7), liver (HepG2), lung (A549), and cervix (HeLa-S3). [[[Shirley+kevin: should we mention ES?]] In different contexts, these top-tier cell lines can be "paired" with functional genomics data form normal tissue (often from epigenome roadmap) or another immortalized cell line from corresponding healthy tissue. (figure XXX). Comparisons of these "TN-pairs" normal could help to model the differential gene regulation between tumor and normal tissues. It is worth noting both relating these cell lines to cancers and pairing the tumor-normal matches is approximate in nature, as these matches are not intended to substitute data from real tumor and normal tissues. Nonetheless, they serve as good models for performing a wide variety of functional genomics profiles, perturbation assays, and experimental validations. Furthermore, many of these pairings have been used in previous analyses \{cite 26241649 25144821}(Figure XXX & supp XXX).[[[Need to refer to supp well here. Just say “see supp.”]]]

To build the companion encyclopedia, we started by defining enhancers. Unlike the general encyclopedia, we used genomic signal tracks from a battery of 5 to 10 histone modification marks in combination with DNase-seq. These were used as input into Casper, a machine learning predictor that we developed. Casper integrates the signal shapes of these various signals around enhancers. We then ensembled these predictions with peaks called from STARR-Seq experiments, which directly read out candidate enhancers in the genome. Such an integrative approach gives accurate definitions of enhancers (see supplement). We then used RAMPAGE data to better define promoters, and further linked enhancers to putative promoters using a prediction algorithm based on gene expression. These potential linkages were then filtered through the results of Hi-C and ChIA-pet experiments, which provide a more accurate map of physical interactions to obtain high confidence enhancer target linkages. The enhancer-target linkages, promoters, and RNA binding sites within genes constitute a so-called extended gene neighborhood (Fig1 C).

## We further linked the enhancers and promoters with their associated transcription factors (TF) to construct extended regulatory networks. First, we built tissue-specific distal and proximal TF regulatory networks by linking TF to genes, either by direct TF-promoter interactions or indirectly via TF-to-enhancer-to-gene interactions (Fig1 B). We then pruned these networks to include only the strongest edges using another signal shape algorithm called TIP \{cite 22039215}. In paired "tumor-normal" cell lines, we measured the signed, fractional number of outgoing edges changing, the rewiring index, and ranked TFs by this.[[[did i get right?]]] In addition, we merged our cell-type-specific networks to get a generalized network for pan-cancer analysis. For each network, we then arranged all regulators into a hierarchy. TFs are placed into different levels of the hierarchy to the degree which they directly regulate the expression of other TFs \{cite 25880651} or are in turn regulated by them. We use a simulated annealing procedure to search for the placement that maximizes the number of downward flowing edges. A final hierarchal network structure is shown in Fig1 D. This shows how highly mutated TFs tend to sit at the bottom of the hierarchy and those associated with large expression changes at the top. [[[more... what about merged?]]]

## Multi-level data integration better enables recurrent variant analysis in cancer

One of the most powerful ways of identifying key elements and functional mutations in cancer is with recurrence analysis to discover regions with greater mutational burden than expected. However, mutational processes can be influenced by confounding factors (in the form of both external genomic factors and local sequence context effects), which can result in many false positives or negatives without appropriate correction \{cite MutsigCV}. In addition, traditional methods often neglect the natural association of mutations among different types of linked annotations (eg a gene body and its linked enhancer) and evaluate regions separately. Consequently, they sometimes fail to identify mutational signals from distributed yet biologically relevant genomic regions, thereby limiting interpretation.

To address these limitations, we adopt a two-pronged approach for better recurrence analysis. First, we predict an accurate local background mutation rate (BMR) by removing effects of confounding factors in a cancer-specific manner. Specifically, we separated the whole genome into bins (1Mb) and calculated mutation counts under each local context category. For each category, we used a negative binomial regression of the mutation counts against features like replication timing, chromatin accessibility, Hi-C signal, and expression profiles for BMR prediction. In contrast to methods that use unmatched data \{cite MutsigCV}, our approach automatically selects the most relevant features, thereby providing noticeable improvements in BMR estimation, which significantly benefits recurrence analyses (Fig 2A). Almost notably it requires the combination of many different genomic features to get such an accurate estimation (Fig [[[ADD PCA]]])

Second, rather than separately testing standalone annotation categories, we used our extended gene neighborhoods as joint test units that contain both the coding exons and non-coding regulatory elements (Fig 1C). Such a scheme allows for the accumulation of weak mutational signatures distributed across multiple biologically relevant functional elements, which may otherwise be missed if evaluated under individual tests. We demonstrate that our scheme can effectively remove false positives and discover meaningful regions with higher-than-expected mutation counts (Fig 2C). For example, in the context of chronic lymphocytic leukemia (CLL), our analysis identifies well-known highly mutated genes, such as TP53 and ATM, which has been reported from previous coding region analysis. It also discovered genes that are missed by the exclusive analysis of coding regions, such as BCL6. Note that BCL6 has strong prognostic value with respect to patient survival (Fig. 2D), indicating that the extended gene neighborhood could be used as an annotation set for recurrence analysis. In addition, we can easily generalize this BMR calibration approach for other cancer types beyond the five discussed here, as our model will pick an appropriately matched ENCODE signal type.[[[still would like to mention this better]]

## Extensive rewiring events in cell-type specific network in cancer

We then investigated the transcriptional regulatory network in a cell-type specific way. Here, we utilized 4 main tumor-normal cell line pairings described earlier to study how the targets of each common TF changed (i.e., rewired) over the course of oncogenic transformation. We first ranked TFs according the “rewiring index” (Fig. 3 A). In leukemia, well-known oncogenes such as MYC and NRF1 are among the top edge gainers, while IKZF1is the most significant edge loser (Fig 3A). Mutations in this later factor serve as a hallmark of high-risk acute lymphoblastic leukemia. [[[ref]]] In contrast, several ubiquitously distributed TFs, such as YY1, [[[why mention YY1?]]] retain their regulatory linkages (Fig 3A). We observe a similar trend in TFs using a distal, proximal and combined network (see details in supplementary file). We also observe highly rewired TFs such as BHLHE40, JUND, and MYC in lung, liver, and breast cancers (Fig 3).

Our rewiring index only considers direct connections associated with a given TF. However, the targets within the TF regulatory network are characterized by heterogeneous network modules (so called “gene communities”), which are dynamic and usually come from multiple biologically relevant genes. Instead of directly measuring the TF’s target changes for each gene, we determined these gene communities via a mixed-membership model. This enabled us to evaluate each TF’s overall association changes to these gene communities in tumor and normal cells. Similar patterns are observed using this model to using the rewiring index (Fig 3A). [[[but some differences? We need to transition better]]] ; more strikingly, we found top "community changers" were strongly associated with poor prognosis (in particular, IKZF1 in Fig XXX).

What drives the certain factors to highly rewire? We find that the majority of rewiring events are associated with noticeable gene expression and chromatin status changes, but not necessarily with variant-induced motif loss or gain events (Fig. 3A). For example, JUND is a top gainer in CLL. The majority of its gained targets in tumor cell lines demonstrate higher gene expression, stronger active and weaker repressive histone modification mark signals. We found a similar trend for the rewiring events associated with JUND in liver cancer.

## Integrating regulatory networks with tumor expression profiles identifies key regulators in cancer

Next, we extended our network analysis in a pan-cancer fashion by merging the cell type-specific networks for both TFs and RBPs. Then using a machine learning method, we integrated 8,202 tumor expression profiles from TCGA to systematically search for the TFs and RBPs that most strongly drive tumor-specific expression patterns. For each patient, our method tests to the degree a regulators’ tumor-to-normal expression changes are sufficiently correlated with their targets’ expression changes. We then calculated the percentage of patients with these relationships in each cancer type and presented the overall trends for key TFs and RBPs in Fig. 4A.

We find that the target genes of MYC are significantly up-regulated in numerous cancers, which is consistent with its known as an oncogenic TF \cite{22464321}. We further validate MYC’s regulatory effect through CRISPi RNA-seq experiments (Fig 4). Consistent with our calculations, the expression of MYC targets is significantly reduced after MYC knockdown (Fig 4A). After confirming the importance of MYC, we use the regulatory network to understand how MYC works with other TFs. We first looked at all 3,735 triplets involving MYC by requiring that a second TF both interacts and shares a common target with MYC . In all cancer types, we found that MYC expression changes are positively correlated with the expression of most of its targets, while the second TF shows only a limited influence as determined from partial correlations. We then investigated the exact structure of such regulatory relationships. The most common triplet interaction type is a well-understood feed-forward loop (FFL) structure in which MYC regulates both the common target and the second TF. Most of these FFLs involve with well-known MYC partners such as Max and Mxl1. However, we also discovered that many involve another factor called NRF1, which has not been previously associated with MYC. Upon further study, we found that that the MYC-NRF1 FFL relationships were mostly coherent ("amplifying") FFLs. We further studied the MYC-NRF1 FFLs by forming these triplets into a logical gate, in which the two TFs act as inputs and the target gene expression represents the output \{cite 25884877}. We can show that the predominant number of these gates follow either OR or MYC-always-dominant logic. Thus, the ENCODE regulatory network not only helps find key regulators, but also to really demonstrate how they work in combination with other regulators.

We also analyzed the RBP network derived from ENCODE data and found key RBPs associated with cancer. For example, the ENCODE eCLIP experiment has profiled many SUB1 peaks on the 3’UTR regions of genes, and we find that the predicted targets of the RBP SUB1 were significantly up-regulated in many cancer types (Fig. 4C). Sub1 has not been associated with cancer before. [[[true??]]] We thus validated this new association in liver cancer. After knocking down SUB1 in HepG2 cells, its predicted targets are also down-regulated relative to other genes (Fig. 4D). In addition, we found that the decay rate of SUB1 target genes are significantly shorter than non-targets (Fig. 4C). These results indicate that SUB1 may bind to 3’UTR regions to stabilize transcripts. Moreover, we found that the up-regulation of SUB1 target genes is correlated with a poorer patient survival in other cancer types such as lung cancer (Fig. 4).

## Step-wise prioritization schemes pinpoint deleterious SNVs in cancer

Summarizing the analysis described above, our companion encyclopedia consists of number annotation resources: (1) a BMR model and matching procedure and a list of regions with higher-than-expected mutational frequencies in various cancers, (2) accurately determined enhancers, promotors and enhancer-target-gene linkages; (3) extended gene neighborhoods, (4) tumor-normal differential expression and chromatin changes, (5) a regulatory network of TFs; (6) based on the network, for each TF position in the network hierarchy and rewiring status; (7) an analogous but less annotated network for RBPs. Collectively, these resources allow us to prioritize key features as being associated with oncogenesis. The workflow in Fig. 5A describes this prioritization scheme in a systematic fashion. We first search for key regulators that are frequently rewired, located in network hubs or at top of the network hierarchy, or significantly driving expression changes in cancer. We then prioritize functional elements that are associated with top regulators, undergo large regulatory changes in terms of TF binding and chromatin status, or are highly mutated in tumors. Finally, on a nucleotide level, we can pinpoint impactful SNVs for small-scale functional characterization by their ability to disrupt or create specific binding sites, or which occur in positions of particularly high conservation.

[[JZ2MG: the first sentence is very confusing, seem that these enhancers are associated with these TFs. I suggest we just remove the 1st sentence. Also we mentioned this in detail already]]

Using this framework, as we described above, we subject a number of key regulators, such as MYC and SUB1, to knockdown experiments to validate their regulatory effects in particular cancer contexts (Fig 4D). Next here, we also identified several active enhancers in noncoding regions, associated with breast cancer, and validated their ability to influence transcription using luciferase assays in MCF7. We further selected key SNVs, based on significantly recurrent mutations in breast cancer cohorts, within these enhancers that are important for controlling gene expression. Of the eight motif-disrupting SNVs that we tested, six showed consistent up- or down-regulation relative to the wild type. One particularly interesting example, illustrating the unique value of ENCODE data integration, is in the intronic region of CDH26 in chromosome 20 (Fig. 5C). Both histone modification and chromatin accessibility (DNase-seq) signals indicated an active regulatory role in MCF7, which was further confirmed as an enhancer by both CASPER and ESCAPE (STARR-seq) (Fig. 5D). Hi-C and ChIA-PET data indicated that the region is within a topologically associated domain (TAD) and validated a regulatory linkage to the downstream breast-cancer-associated gene SYCP2 \cite{26334652, 24662924}. We found that 21 out of the 52 ChIP-Seq experiments in MCF-7 from ENCODE demonstrate a high frequency of chromatin interactions in this region.[[[what does this mean?]]] Motif analysis predicts that the particular mutations found in the cohorts can significantly disrupt the binding affinity of several TFs, such as FOLS2, in this region.[[[correct???]]] Luciferase assays demonstrate that this mutation introduces an xx-fold reduction in expression relative to wild type expression levels, indicating a strong repressive effect on this enhancer’s functionality.

## Conclusion

This study highlights the value of our companion to the encyclopedia as a resource for cancer research. First, we show that, by integrating many different types of assays on a large scale, we can achieve a very accurate annotation of ENCODE tier-1 cell lines, related to cancer, building up an extensive regulatory network. Then we show how comparisons within this resource itself can illuminate potential regulatory changes in cancer (ie key rewiring TFs). Next, we show how the resource can be generalized into a pan-cancer regulatory network and BMR framework to help interpret patient data from cancer cohorts, both gene expression and mutation data. Finally, we show how we can leverage the companion resource to provide a prioritization scheme for pinpointing key regulatory elements and SNVs for small-scale follow-up. This study underscores the value of large-scale data integration, and we note that expanding this approach (either by integrating additional data types and/or using tumor mutation and expression data on a larger scale) is straightforward. We also anticipate that an additional step would be to carry out many of the ENCODE assays on specific tissues and tumor samples. Though volume of material needed for such analyses may present challenges, we show that such a framework is technically feasible and provides further opportunities for the future.