# Integrating ENCODE data to interpret regulatory changes in cancer

## Abstract

Cancer is caused by mutations in the DNA which disrupt the normal physiology of cells. While mutations on coding genes have been well characterized, those on non-coding regions, representing the majority of somatic mutations in cancer, are still poorly understood. Recent release of ENCODE include diverse profiles of functional genomics assays on cancer and normal cell lines. Integration of ENCODE functional genomics data in cell lines with TCGA molecular profiles on tumor tissues enable us to bridge the knowledge gaps in a number of cancers to better model mutations in non-coding regions.

In cancer, the impact of mutations in a limited number of coding genes is well characterized; in contrast, the preponderance of variants constitute poorly characterized mutations occur in non-coding regions. The new release of the ENCODE data enables us to bridge these knowledge gaps for a number of well-studied cancers of the blood, liver, lung, breast and cervix. For each of these cancer cell lines, ENCODE provides diversity of genome-wide assays (e.g., ChIP-seq, DNase-seq, Enhancer-seq, Hi-C, and ChIA-PET), and these are applied to matched tumor-normal cell lines.

First, the new ENCODE data enables precise calibration of non-coding background mutation rate by removing the confounding effects of chromatin and replication timing from tissue-matched samples. Furthermore, by integrating different ENCODE assay types, we were able to define high-confidence regulatory elements and their target genes. This allows us to define extended gene neighborhoods, which gave more sensitive mutation recurrence analysis than just coding genes. This approach identified novel genes, such as *BCL6* in leukemia, that are recurrently mutated in cancers and associated with patient prognosis.

Second, we integrated the ENCODE data to build a hierarchical regulatory network, including both transcription factors (TFs) and RNA-binding proteins (RBPs). Intriguingly, 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 tumor and normal network, we identified highly “rewired” TFs with changed targets and prognostic value, such as IKZF1 and MYC. Our results indicate that such rewiring events are mainly attributable to epigenetic changes, rather than mutations that disrupt TF motifs.

Third, we proposed a prioritization scheme for key non-coding elements, as well as the mutations within, according to their positions in regulatory networks and potentials to drive expression changes in cancer. In particular, we prioritize CTCF as a key TF for blood cancer and SUB1 as a key RBP for liver and lung cancers, and validated them through siRNA knockdown experiments. Finally, we identify active enhancers and seven high impact mutations therein in breast cancer and validated their functional effects through luciferase assays.

## Introduction

[Question: does the introduction look OK? Are we clear what we have done?]

Mutations associated with cancer have been well characterized in a many key oncogenes and tumor suppressors. However, the overwhelming bulk of mutations in cancer genomes – particularly those discovered from the 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. Newly-released data from the ENCODE Consortium can help address this question by providing comprehensive characterization of non-coding genomic elements, as well as by linking such elements to well-known cancer associated genes.

Here, we endeavor to provide a companion resource to the main ENCODE encyclopedia by building a “cancer encyclopedia”. The main encyclopedia is oriented toward breath of the annotations to describe elements over hundreds of cell lines. In contrast, we focus on cell lines (the “tier 1 & 2" lines) with a wide variety of profiles available. Most of these cell lines are associated with cancers of the blood, liver, lung, cervix, and breast. We show that these cell lines can be used to provide a better understanding of oncogenesis, and we provide a resource for interpreting the wealth of mutational and transcriptional profiles produced by the cancer community.

In particular, we first develop a regression-based method to integrate ENCODE data to calibrate an accurate background mutation rate (BMR). This allows us to accurately find significantly burdened regions in many cancers. We further use ENCODE assays to accurately define non-coding elements, enhancers in particular, and how these elements are linked to known genes. This enables us to delineate regulatory networks involving transcription factors (TFs) and, to a lesser extent, RNA-binding proteins (RBPs). We represent these networks in a variety of ways, including hierarchical models, wherein master regulators occupy the top of the hierarchy. For each regulator in the network, we then calculate a rewiring index that represents the degree to which a regulator differs between normal and cancerous cells. We then generalize our work to make it applicable over many cancer types, both in terms of mutational burden analysis and also to leverage the regulatory network to interpret expression data in a pan-cancer fashion. Finally, we show how our regulatory network and mutational burdening analyses can be combined into a unified workflow to prioritize specific regulators, regions, and variants in throughout the cancer genome. Our inferences from this workflow are then validated using a number of small-scale experimental studies.

## Data for comprehensive functional characterization in ENCODE

[Question: does the data section make sense? Some of tumor/normal matching are very rough, will this cause lots of challenges?]

The ENCODE tier 1 cell lines are especially important because the most comprehensive set of profiles are conducted for these cell lines. They provide good models not only for studying gene regulation in details, but also for understanding cancers of the blood, breast, liver, lung, and cervix. For four of the five cell lines within tier 1, there is another immortalized cell line from corresponding healthy tissue. Therefore, comparisons of the data from cancer and normal cell lines could help model gene regulation in tumor versus normal tissues. It is worth noting that both matching the cell lines to cancers and matching the tumor-normal pairs are very approximate in nature, as these matches are not intended to substitute data from real tumor and normal tissues. Nonetheless, they are good models to perform a wide variety of functional genomics profiles, perturbation assays, and experimental validations. In addition, the data generated on these cell lines can be used to better interpret molecular profiles from tumor tissues and understand gene deregulation in cancers.

 (Fig 1A).

To build a cancer encyclopedia with these cell lines, we first construct a comprehensive data matrix by normalizing raw signals of genomic features that severely confound somatic mutagenic processes (see Supp. File/Section(?) X ). In contrast to previous approaches that rely on single histone modification mark, we implement an ensemble-based method called ESCAPE, which performs large-scale data integration to accurately identify active enhancers. This integration involves predictions using a diverse collection of histone mark ChIP-seq, DNase-seq and Enhancer-seq datasets. We further link these to genes by optimally investigating how the histone modification marks on enhancers help predict the gene expression of the potential target gene. This group of potential linkages is then filtered through the results of Hi-C experiments, which provide a more accurate yet lower resolution map of chromatin interactions (see Supp. File/Section(?) X). To improve functional interpretation, we used these high-quality linkages to construct what we termed “extended gene neighborhoods” – coding regions matched with key regulatory elements, such as enhancers, promoters, and binding sites from regulators (Fig1 B). In addition, we explored the binding profiles in ENCODE data, and construct high-confidence gene regulatory networks for both TFs and RBPs (Fig. 1C and Fig. X in Supp. File/Section(?) X). Finally, we merge our Cancer encyclopedia with the broader ENCODE encyclopedia, and we provide consistent identifiers and definitions.

In all, our resource consists of a list of accurately determined enhancers, a list of regions with high mutation burdens in cancer, the regulatory network of TFs (and for some lines RBPs), as well as the most rewired TFs in this regulatory network (see supplementary materials). Collectively, these resources allow us to prioritize a few key elements as being associated with oncogenesis, some of which are then validated using small-scale experiments (see table S1).

## 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 through recurrence analysis, which finds regions of the genome that are more mutated than expected. However, mutation process could be influenced by or associated with confounding factors (in the form of both external genomic factors and local context effects), which can result in many false positives or negatives in recurrence analysis (see Supp. File/Section(?) X). In addition, traditional burden tests often neglect the association among annotation categories and test regions separately. Consequently, these tests sometimes fail to identify mutation signals from dispersed yet biologically relevant genomic regions, thereby limiting the interpretation power of burden tests.

In contrast, we integrated the ENCODE resources at two levels for better recurrence analysis. First, we predict an accurate local BMR by regressing out the confounding effects of features in a cancer-specific manner (see Supp. File/Section(?) X). Specifically, we prepare a covariate matrix by normalizing 475 features from ENCODE to remove those effects that may confound the BMR. We then separated the whole genome into 64 categories according to the local 3-mers and run separate regression models to further remove confounders from intrinsic sequence contexts. In contrast to methods that use unmatched data \{cite MutsigCV}, our regression-based approach with matched data usually yields higher BMR prediction precision (Fig 2A, see also Supp. File/Section(?) X). In breast cancer, for example, the spearman’s correlation ($ρ$) between observed and predicted mutation counts over 1-megabase bins increases from XX to XXX when using replication-timing signals from MCF-7 instead of HeLa-S3. This underlies the importance of integrating chromatin features from matched tissues to infer BMR (Fig 1B). For example, $ρ$ only ranges from xxx-xxx using matched replication timing, but its range increases to xxx-xxx by adding 1 PC from the remaining covariates. It progressively increases to the xxx-xxx regime by adding PCs to the full model through forward selection (Fig 1B, see Supp. File/Section(?) X). Such noticeable improvements in BMR estimation significantly improve the burden analyses below.

Rather than separately testing standalone annotation categories, we employ our extended gene (detailed above) as joint test units (see Supp. File/Section(?) X). Such a scheme allows for the accumulation of weak mutation signals distributed across multiple biologically relevant functional elements, which may otherwise be missed if evaluated under individual tests (Fig. Sx in Supp. section X). Furthermore, it enables to collectively assess the overall burdening associated with a well-known cancer associated gene, potentially the burdening in the protein coding regions may be matched by apparently un-connected mutations in the regulatory regions. We demonstrate that our scheme can effectively remove false positives and discover meaningful burdened regions (Fig 2C). For example, in the context of leukemia, our analysis identifies well-known highly mutated genes (such as TP53 and ATM) as well as other genes (such as BCL6) that are missed by the analysis of coding regions. BCL6 has strong prognostic value with respect to patient survival (Fig. 2D), indicating that the extended gene should be used as an annotation set for recurrence analysis. In addition, we can easily generalize this BMR calibration approach for cancer types not in the five we are focusing on, as our model will work to pick an appropriately matched ENCODE signal type.

## Extensive rewiring events of several transcription factors in cancer

We then investigated the transcription regulation network in a tissue specific way. In each cell type, we organized the TF regulatory network into a hierarchy by comparing the inbound and outbound edges of each factor, thereby enabling us to investigate the global topology of TF regulation (Fig. 1E, see also Supp. File/Section(?) X). TFs in different levels of the hierarchy reflect the extent to which they directly regulate the expression of other TFs \{cite 25880651}. For example, TFs in the top layer have more outbound than inbound edges in the network, and thus play larger roles in regulating other TFs (Supp. Fig. xx). In this representation, two patterns readily emerge. In leukemia, top-level TFs tend to more strongly influence the differential expression between tumor and normal cells. The average Pearson correlation between TF binding events and tumor-normal expression changes increases from 0.125 in the bottom layer to 0.270 in the top layer (Table Sx). TFs in the bottom layer are more frequently associated with burdened binding sites in general, perhaps reflecting their increased resilience to mutation (see Supp. Section X, Table Sx).

When comparing the common regulators in approximately matched tumor and normal regulatory networks, rewiring (i.e., target changing) analysis may help to identify cancer-associated deregulation. Hence, we investigated rewiring events in TF networks using multiple formulations (see Supp. File/Section(?) X). Specifically for leukemia, we removed the general TFs and restricted our rewiring analysis to 61 common TF ChIP-seq in K562 and GM12878 from ENCODE (see Supp. File/Section(?) X). We first ranked TFs according , which calculates their respective number of lost and gained edges. Oncogenes such as MYC and NRF1 are among the top edge gainers. In contrast, IKZF1, whose somatic mutations serve as a hallmark of high-risk acute lymphoblastic leukemia, is the most significant edge loser, with up to xxx% of lost edges in K562 (Fig 3A). In contrast, several ubiquitously distributed TFs, such as 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 in lung and liver cancers (see fig XX) although we do not have as many common TFs between tumor and normal cell lines for these tissues.

Our rewiring index only considers direct connections associated with a given TF. One may also consider rewiring that include not only direct connections, but also the whole neighborhood of connections with which a TF associates through membership and topic models. In particular, we used a mixed-membership model to look more abstractly at local gene neighborhoods to re-rank the TFs (see Supp. File/Section(?) X). Similar patterns are observed using this model. We also observed that MYC (a well-known oncogene) becomes a top edge gainer (Fig 3A). To study the consequences of network rewiring under this model, we performed survival analysis on xxx AML patients and found IKZF1 to be significantly associated with tumor progression (see Supp. File/Section(?) X).

The combinatorial regulation of many TFs jointly determines the “ON” and “OFF” states of genes as part of maintaining homeostasis in healthy cells. The disruption of co-regulatory relationships for key elements in cancer cell lines ultimately results in erroneous gene expression patterns. We quantified the co-association status of each TF, and observed major co-association changes in some of the key TFs when comparing the regulatory networks of K562 to GM12878. For example, ZNFXXX is a suppressor TF that shows only marginal co-binding events in GM12878, with binding sites increase from xxx in GM12878 to xxx in K562. In addition, up to xxx% of its binding sites co-bind with other TFs (in which cell?).

A remaining uncertainty lies in the underlying causes of this rewiring. Is it a direct effect of mutations, which could knock out a binding site? Or it is due to indirect effects of chromatin changes, which could cover and uncover binding sites? We find that the majority of rewiring events result from changes in chromatin status, rather than from variant-induced loss or gain events (Fig. 3A). For example, JUND is a top gainer in K562 (with xxx gains and xx losses). We find that majority of the gain/loss events are associated with substantial expression changes (of at least 2-fold) and changes to chromatin states. Only xxx percent of them could be potentially due to direct motif loss/gain effects. (Fig. 3D).

## Integrating ENCODE data with tumor expression profiles identifies key regulators in cancer

[Question: is our description about RABIT part accurate? Is there something else we should add?]

Using a regression-based learning method (see Supp. File/Section(?) X), we integrated the molecular profiles of 8,202 TCGA tumors with 921 ENCODE regulatory binding profiles to systematically search for TFs and RBPs that drive tumor-specific expression patterns (Table Sx). For each tumor sample, our integration framework selects a set of regulator binding profiles to best explain the expression difference between the tumor sample and normal controls. For each cancer type, our framework further tests whether the regulatory targets identified in a ENCODE experiment are sufficiently correlated with the regulator’s molecular status across tumors. The final output of our framework is the estimated fraction of patients with target genes differentially regulated between each pair of cancer type and regulator (see Supp. File/Section(?) X).

The overall trends for the key TFs and RBPs detected are given in Fig. 4A. The predicted impacts of regulators on tumor gene expression are highly consistent with previous findings. For example, we find that the target genes of MYC are significantly up-regulated in numerous cancers (star in Fig Sx), which is consistent with the known role of MYC as an oncogenic TF. In addition to recapitulating existing knowledge from previous studies, our analysis also predicts previously unidentified functions for regulators in cancer. For example, the predicted targets of the RBP SUB1 were significantly up-regulated in many cancer types (Fig. 4C). Moreover, the up-regulation of SUB1 target genes is correlated with a worse patient survival in cancer types such as lung cancer (Fig. 4). Previously, SUB1 was considered as a TF. However, the ENCODE eCLIP experiment has profiled many SUB1 peaks on gene’3UTR regions (Supp Fig. X). In HepG2 cell where SUB1 eCLIP experiment was done, the decay rate of SUB1 target genes are significantly shorter than non-targets (Fig. 4C). After knocking down SUB1, its predicted targets are also down regulated comparing to other genes (Fig. 4D). These results indicate that SUB1 may bind gene 3’UTR regions to stabilize transcript level. From our integrated analysis, the higher SUB1 activity through regulatory binding on 3’UTR regions is likely to drive tumor specific expression patterns in many cancer types.

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

The above description of the regulatory network and mutation recurrence analysis provide a way to prioritize key genomic features associated with cancer. The workflow in Fig.5 A describes this prioritization scheme in a systematic fashion. First, we start by searching for key regulators that frequently rewired, locate in network hubs or on top of the network hierarchy, or significantly drive expression changes in cancer. We then prioritize functional elements that are associated with top regulators, undergo large regulatory and chromatin changes, or (most importantly) 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 or chromatin changes.

Using this framework, we subject a number of key regulators (such as MYC and SUB1) to siRNA knockdown to validate their regulatory effects (Fig 4D). We then identified several active enhancers in noncoding regions, and validated their ability to influence transcription using luciferase assays (see Supp. File/Section(?) X). We further selected key SNVs within these enhancers that are important for gene expression control (table Sx). Of the eight motif-disrupting SNVs that we tested, six showed consistent up- or down-regulation effect in expression relative to the wild type (Fig. 5B and Supp. File/Section(?) X). One particularly interesting example, illustrating ENCODE data integration, is on chromosome 6, 13.5xxx (Fig. 5C). This enhancer is located in a noncoding region. Both histone modification and DHS signals implicate its regulatory role as being active (Fig. 5C), and both our HisShape enhancer prediction method and the EnhancerSeq experiment support its enhancer function (Fig. 5D). Hi-C and ChIA-PET data link this region to a downstream gene SYCP2. 21 out of the 52 ChIP-Seq experiments demonstrate that the region has high frequency of chromatin interactions, and motif analysis predicts the C to G mutation in cancer can significantly disrupts the FOLS2 binding affinity (see Supp. File/Section(?) X). 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

[Question: do you want to make further claims?]

This study highlights the values of ENCODE as a resource for cancer research, and leverages ENCODE to provide a prioritization scheme to pinpoint key regulatory elements and SNVs for small-scale validations. One of the key aspects of our analysis clearly scales with larger quantities and more diverse data types. In particular, we anticipate that higher quality annotations of non-coding elements through progressively more accurate Enhancer-seq experiments and deeper Hi-C experiments will link them better with target genes. Likewise, the recurrence analysis can be further improved by collecting better-matched datasets and expanding the size of tumor cohorts. In that the analyses presented here improve upon data integration, it provides future investigations with a blueprint for similar studies going forward. By amassing ever-larger data sets, we may obtain a more accurate picture of the cancer genome through large-scale data integration.