# Integrating ENCODE data to interpret regulatory changes in cancer

## Abstract

In cancer, the majority of catalogued mutations are within coding genes. However, the preponderance of mutations in tumors occur in non-coding regions throughout the genome. While coding mutations yield easily inferred mis-sense, non-sense or frameshift events that lead to altered proteins, non-coding mutations are often exceptionally difficult to characterize. Functional mapping of the non-coding genome in efforts such as the ENCODE Project provides an opportunity to assess non-coding mutations in the light of chromatin state and transcriptional factor binding throughout the human genome. For a variety of cancer-derived cell lines, as well as non-cancerous cell lines derived from the same tissues (allowing in some cases for tissue-matched non-coding background mutation rate), ENCODE provides diversity of genome-wide assays (e.g., ChIP-seq, DNase-seq, Enhancer-seq, Hi-C, and ChIA-PET). The resulting data and functional maps of the human genome provide a framework for assessing the potential for cancer mutations in the non-coding genome to dysregulate genes.

Here we integrate [[get rid of precise]]diverse ENCODE data to define high-confidence regulatory elements and their linkages to genes. This allows us to create definitions of extended gene neighborhoods for mutation recurrence analysis. [[make clear coding & noncoding…. More sensitive is a little cryptic]]

Using this approach, we can identify additional genes associated with patient prognosis, beyond well-known highly mutated oncogenes (e.g., *BCL6* in leukemia). We also integrate the ENCODE data to build hierarchical regulatory networks, including both transcription factors (TFs) and RNA-binding proteins (RBPs). We find that more mutationally burdened TFs tend to be located at the bottom of the hierarchy (e.g., EZH2 and NR2C2), whereas those associated with the largest oncogenic gene-expression changes tend to be at the top. Furthermore, by comparing tumor and normal network, we have identified highly “rewired” (i.e. target changing) TFs, such as IKZF1 and MYC that hold prognostic value. Our results indicate that such rewiring events are mainly attributable to chromatin changes, instead of direct motif loss/gain effects from mutations. [[good to say that we could do this on … talk about the datasets that come to bear]][[unprecedented… large-scale integration]]Finally, we propose a prioritization scheme for key non-coding elements (as well as variants therein) according to their positions in regulatory networks and potentials to drive oncogenic expression changes. We then validate their functional impact in small-scale experimental studies. In particular, we prioritize CTCF as a key TF for blood cancer and SUB1 as a key RNA-binding protein for liver and lung cancers and validated them through siRNA knockdown experiments. Finally, we identify active enhancers and seven high impact mutations therein and validated their functional effects through luciferase assays in breast cancer.

## Introduction

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

Coding mutations associated with cancer have been the focus of extensive study. However, the overwhelming bulk of mutations in cancer genomes – particularly those coming out of the new large-scale cancer genomics initiatives – lie within non-coding regions. The degree to which these variants either drive cancer 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 many well-known cancer associated genes with such elements.

Here, we endeavor to provide a companion resource to the main ENCODE encyclopedia by focusing on cancer and 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 a few cell lines for which we have a very wide variety of assays (the “tier 1 & 2" lines). [[we have this many datasets… lots of integration… we 30 starrseq, …. What encode about… teir1 lines a contrast one histone mark on

Many of these cell lines are associated with various cancers, particularly those of the blood, liver, lung, cervix, and breast. We show that these cell lines can be used to provide a better understanding of the portion of the non-coding genome that is affected during cancer development and progression, and we provide a valuable resource for interpreting the wealth of variant and dysregulated gene expression data 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 identify significantly burdened regions in many cancers. We further use ENCODE assays to 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 networks 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 problems? Is our claim for the matching sound enough?]

The most comprehensive set of assays for ENCODE are available for tier 1 cell lines. These cell lines provide an invaluable resource for studying genomic regulation. Five tier 1 cell lines are derived from cancers, including cancer of the blood (K562), breast (MCF-7), liver (HepG2), lung (A549), and cervix (HeLa-S3). Moreover, for four of these five cell lines, there is another cell line (from relatively healthy tissue) that can be used for comparisons. This provides an approximate ‘normal’ match to these tumor-derived cell lines. Of course in this case both the matching of the cell lines to cancers and then the approximate matching of the tumor-normal pairs is very much approximate in nature – these matchings are not intended to serve as a substitute for data from real cancer tissue. However, approximate matchings can be used to integrate a wide variety of available omics data to determine significant differences between the tumor-derived and immortalized ‘normal’ cells. Importantly, this exercise provides a resource that can then be used to better interpret the more limited volume of data derived from true cancer tissue. [[laying it out up front]]

(Fig 1A).

To build a “cancer-relevant encyclopedia of DNA elements” (C-ENCODE)[[cancer encyclopedia is too pretensions]] , we first constructed 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 marks, we implemented 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 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 low resolution three dimensional physical picture of inter-genomic connections (see Supp. File/Section(?) X). To achieve improved functional interpretation, we used these high-quality linkages to construct what we term “extended gene neighborhoods” – coding regions matched with key regulatory elements, such as enhancers, promoters, and binding sites from regulators (Fig1 B). In addition, we also explored the binding profiles in ENCODE data, and constructed high-confidence gene regulatory networks for both TFs and RBPs (Fig. 1C and Fig. X in Supp. File/Section(?) X). Finally, we merged our integrated analysis with the broader ENCODE data, and we provide consistent identifiers and definitions for the C-ENCODE resource.

In summary, the C-ENCODE resource consists of a list of accurately determined enhancers, a list of burdened regions, 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 experimental assays (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 deleterious mutations in cancer is through recurrence analysis, which identifies regions of the genome that are more frequently mutated than expected. There are two challenges associated with such analysis. First, the mutation process introduces confounding factors (in the form of both external genomic factors and local context effects), which can result in many false positives or negatives (see Supp. File/Section(?) X). Secondly, traditional burden tests often neglect the interplay among annotation categories, thereby testing regions separately. Consequently, these tests are sometimes unable to identify distributed mutation signals from biologically relevant regions, thereby limiting the functional interpretation of the burdened regions.

To address these limitations of traditional recurrence analysis, we integrate the C-ENCODE resource at two levels. 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 deal with internal context effects. In contrast to methods that use unmatched data \{cite MutsigCV}, our regression-based approach demonstrates that matched data usually provides higher BMR prediction precision (Fig 2A, see also Supp. File/Section(?) X). In breast cancer, for example, the correlation between observed and predicted mutation counts over 1-megabase bins () using replication-timing signals (from MCF-7) increases from XX to XXX relative to that using data from HeLa-S3. Thus, it is important to integrate these features to infer BMR (Fig 1B). For example, 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 burden analyses (see 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 lost 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 approach can effectively remove false positives and discover meaningful burdened regions (Fig 2C). For example, in the context of K562 cells derived from a chronic lymphocytic leukemia (CLL), our analysis identifies well-known highly mutated genes (such as TP53 and ATM) [[– NOT written right -- ]]as well as other genes (such as BCL6) that are missed by recurrance analysis of coding regions. BCL6 demonstrates 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 a cancer type that is not related to one of the five on which we focused. This will work with our model and it will pick an appropriately matched C-ENCODE signal type.

## Extensive rewiring events of transcription factors in cancer

We next investigated transcription regulatory networks in a tissue-specific manner. In each cell type, we organized TF regulatory networks 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 the blood cell comparison (K562 vs. GM12878), top-level TFs tend to more strongly influence the differential expression between the CLL-derived K562 cells (tumor) and the immortalized GM12878 lymphoblastoid cells (normal). 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 networks, rewiring (i.e., target changing) analysis may help to identify cancer-associated deregulation. Hence, we investigate rewiring events in TF networks using multiple formulations (see Supp. File/Section(?) X). Specifically for the CLL example, out of the 69 common TFs in K562 and GM12878 from ENCODE, we remove the general TFs and restricted our rewiring analysis to the remaining 61 (see Supp. File/Section(?) X). We first rank TFs according to their respective number of lost and gained edges, a “rewiring index” (Fig. 3 A, see also Supp. File/Section(?) X). Several oncogenes (such as MYC and NRF1) are among the top edge gainers. In contrast, IKZF1 (somatic mutations in which serve as a hallmark of high-risk acute lymphoblastic leukemia, or ALL) is the most significant edge loser, with up to xxx% of lost edges in K562 (Fig 3A). However, several ubiquitously distributed TFs (such as YY1) retain their regulatory linkages (as shown in Fig 3A). We observe a similar trend in TFs using a distal, proximal and combined network (see details in supplementary file). Similarly, we also observe highly rewired TFs in lung and liver cancer (see fig XX) though 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 more elaborate changes that include not only direct connections, but also the whole neighborhood of connections with which a TF associates through a variety of 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 gainer (Fig 3A). To study the consequences of network rewiring under this model, we performed the survival analysis on xxx AML patients, in which we find 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 all 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 observe 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. Its number of binding sites increases from xxx to xxx in K562. In addition, up to xxx% of its binding sites co-bind with other TFs.

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 a lot of the gain/loss events are associated with substantial expression changes (of at least 2-fold) and changes to chromatin states. However, only xxx percent of them could be potentially due to direct motif loss/gain effects. (Fig. 3D).

[[call out examples that are not blood cancer]]

dotted line around B parallel to D\\

separate B & D

in last figure

move A over and put a thin line

put an arrow to the right

put the cell klines that myc ^& sub1 are validated in – show they are not pancancer

highlight the gride

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

Next we generalized the cell type specific network analysis across multiple cancer types. We produced a generalized pan-cancer regulatory network by merging the cell type specific networks to make a merged regulatory network for both TFs and RBPs. This generalized network can then be used for interpreting the many different gene expression data sets that are now available for different cancers. Using a regression-based learning approach, we normalizes these data in a consistent fashion to the uncut expression data, thereby creating a consistent gene expression resource.

Using a machine learning method (see Supp. File/Section(?) X), we integrated thousands of patient expression profiles from multiple cohorts to systematically search for TFs and RBPs that drive tumor-specific expression patterns (Table Sx). In particular, for each regulator-cancer type pair, we selected the best explanatory binding profile and estimate the fraction of patients with differentially regulated target genes (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 (Figure 3C). As another example, the predicted targets of the TF CTCF were found to be significantly up-regulated in multiple tumors (star in Supp. Fig. 2).

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

[Questions: do you think this prioritization make sense? Is the description of the exp clear enough? What else need to add in the experiment part?]

The above description of the regulatory network and the optimum determination of mutation recurrence provide an approach 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: a) are most frequently rewired; b) sit within network hubs or on top of the network hierarchy; or c) significantly drive oncogenic expression changes. We then prioritize functional elements that are associated with highly prioritized regulators, undergo large regulatory and chromatin changes, or (most importantly) are highly mutated in tumor cohorts. 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 CTCF and SUB1) to knock down experiments to validate their regulation effects (Fig 4D). We then identified several active enhancers in noncoding regions, and validated their ability to initiate transcription using luciferase assays (see Supp. File/Section(?) X). In addition, we further selected key SNVs within these enhancers that are important for gene expression control (table Sx). Of the 8 motif-disrupting SNVs that we tested, we observed 6 variants with consistent up- or down-regulated activity 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 indicate that this is an enhancer (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 regulatory traffic, and motif analysis predicts this C to G mutation can significantly disrupts the FOLS2 binding affinity (see Supp. File/Section(?) X). A luciferase assay demonstrates 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 values of ENCODE as a resource for cancer research, and leverages ENCODE to provide a step-wise 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 non-coding annotations (through progressively more accurate Enhancer-seq experiments and deeper Hi-C experiments) will enable better linkages. Likewise, the recurrence analysis can be further improved by collecting better-matched data sets 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.