# Using the ENCODE regulatory data to interpret non-coding somatic variants in cancer

# [JZ2MG]##### around 2500 word without abstract and removed

## Long Abstract New version

# [JZ2MG]##### long abstract 574 word ###

Though the impacts of somatic mutations within the very limited number of cancer-associated genes are well understood, the overwhelming number of mutations in cancer genomes fall within non-coding regions, rendering them far more difficult to evaluate. Data obtained from the new ENCODE release allows us to bridge these gaps in knowledge.

Here we collected comprehensive data from ENCODE and deeply integrated them to interpret cancer genomes. In particular, we combine computational predictions with EnhancerSeq experiments to generate high quality enhancer lists in several cell lines. To link the enhancers to genes, we performed enhancer target prediction by synthesizing evidences from expression profiles and chromatin status, and further pruned it using Hi-C data for higher accuracy. With this high quality linkage, we are able to define the extended genes by combining coding regions with key regulatory elements in enhancers and promoters for better functional interpretation. In addition, we also explored the full spectrum of the binding profiles from ENCODE and set up high confidence gene regulatory networks for both transcription factor (TF) and RNA binding proteins (RBPs).

We then integrated various signal tracks from comprehensive experiments to carry out rigorous recurrence analysis on the proposed extended gene regions. Specifically, we calibrated a genome-wide background mutation rate (BMR) by regressing out the effects from well-known confounders, such as replication timing and chromatin status. Then we performed a joint burden test on the extended genes to amplify mutation signals that might be weakly distributed in individual elements. Analyses show that our scheme could effectively remove false positives and discover meaningful burdened regions. In the context of leukemia, our analysis identified well-known drivers (such as TP53 and ATM) and key genes (BCL6) that has strong prognostic value but missed by coding region analysis.

We then explored the structure of TF-TF network by organizing it into a stratified hierarchy through comparison of outbound edges to inbound ones. We find that top-level TFs tend to be more associated with tumor-to-normal differential expression, and bottom-level TFs (e.g., EZH2 and NR2C2) tend to be enriched with burdened binding sites. We then rigorously compared TF regulatory networks between loosely matched tumor and normal cell lines and identified significantly rewired (i.e., target-changing) TFs, such as IKZF1 and MYC. By integrating large-scale chromatin features and whole genome sequencing (WGS) data, we demonstrate that such massive tumor-to-normal rewiring events may largely be explained by changes in chromatin structures, rather than direct mutational effects. Patient survival analyses reveal that the regulatory activity of our top rewired TF (IKZF1) is significantly associated with cancer progression.

We further integrated expression data from multiple cohorts into the more generalized TF/RBP network to prioritize key regulators that significantly drives the differential expression between normal and tumor cell lines in multiple cancer types. We identified ZNF687 as a key TF for breast cancer and SUB1 as a key RBP for liver and lung cancer. We further validated the effect of these TF/RBPs through different siRNA knockdown experiments.

Finally, we developed a step-wise scoring workflow to prioritize key variants in a cancer specific way. In particular, we identify several active enhancers in breast cancer pinpointed variants therein that potentially affect their downstream gene expressions. Experiments on both wild and mutant type sequences through luciferase assays confirmed their effects in MCF-7.

Our work demonstrates how careful integration of ENCODE resources offers unprecedented opportunities to accurately characterize oncogenic regulation and serves as a powerful tool to prioritize cell-type specific regulatory elements and variants in cancer.

## Short Abstract

#/\*= requirement from Nature **Articles:** an abstract of approximately 150 words =\*/

#/\*= right now 271 words =\*/

While the majority of somatic mutations occur in non-coding regions, we only understand mutations well in very limited cancer-associated genes. Data obtained from the new ENCODE release allows us to bridge the gaps. Here we collected comprehensive data from ENCODE and deeply integrated them to interpret cancer genomes.

In particular, we provided high quality enhancer list and their gene target linkage to define the extended genes and then set up gene regulatory networks for both transcription factors (TFs) and RNA binding proteins (RBPs). Based on these data, we performed rigorous recurrence analysis on the proposed extended gene regions after integrating comprehensive signals. We identified well-known drivers (such as TP53 and ATM in CLL) and key genes (BCL6) that has strong prognostic value but missed by coding region analysis. We then explored the hierarchy of TF-TF network and find that top-level TFs tend to be more associated with tumor-to-normal differential expression, and bottom-level TFs (e.g., EZH2 and NR2C2) tend to be with more frequent burdened binding sites. We investigated the edge gain and loss events in matched tumor/normal networks and identified significantly rewired TFs (e.g., IKZF1 and MYC) that are highly associated with cancer progression. We further integrated expression data into the more generalized TF/RBP network and identified ZNF687 and SUB1 as key regulators in breast and lung cancer and validated their effects through knockdown experiments. Finally, we developed a step-wise scoring workflow to pinpoint key variants and confirmed their effects through luciferase assays. Our work demonstrates that data from ENCODE after careful integration may serve as a powerful resource for the cancer community to investigate and prioritize key regulators and variants.

**[[dc2all: Starting edits here (I didn’t look at the Abstract – since that has already been edited separately).]]**

**[[dc2all:**

* **yellow highlighting -- designates text that is very unclear and needs to be substantially re-worked**
* **need to try to keep tense consistent -- I think the present tense is prefered by most journals**
* **"cell-line" vs "cell line"?**
* **"Fig. X" vs "Fig X" vs "Figure X"?**
* **"standalone" or "stand-alone"?**
* **"Fig. Sx" vs "Supp. Fig. X"?**
* **"percent" vs "%"**
* **italicize gene names or not -- but best to keep consistent**

**]]**

## Introduction

The advent of whole-genome sequencing (WGS) and personal genomics are providing new opportunities to identify key regulatory elements and deleterious mutations therein, many of which play important roles in carcinogenesis. In turn, these novel findings may enable the development of targeted therapies in clinical studies. Despite the discovery tens of thousands of mutations, only a very small fraction is readily interpretable in terms of their effects on known cancer-associated genes. A lingering uncertainty lies in the degree to which these tens of thousands of variants contribute to cancer. Are they simply neutral passengers created as byproducts of the genomic dysregulation known to be intrinsic to cancer genomes? Alternatively, do key cancer-driving variants remain lurking among this newly discovered pool of mutations? The new data release of ENCODE Consortium may bridge these knowledge gaps by providing accurate non-coding annotations and precisely linking these annotations to well-characterized protein coding genes.

The ENCODE data resource provides a fundamental annotation of the human genome. Annotations are assigned in a cell line-specific manner, and many of the cell lines used are cancerous. Admittedly, this data is imperfect in the context of cancer research: some cell lines are suboptimal for actual tumor samples, and matching data from these cell lines with that from appropriate normal samples is often challenging. However, because of the tremendous richness of assays available in this new ENCODE release, comparisons of tumor-like and normal-like cell lines provide an unprecedented and accurate window into the regulatory and chromatin-related changes associated with cancer.

Here, we endeavor to make the ENCODE resource as useful as possible for cancer research. We optimally match the cell lines with known cancers to better integrate relevant expression profiles and somatic mutations from known cohorts. We then develop methods to integrate comprehensive ENCODE signal tracks to calibrate an accurate background mutation rate (BMR), which is then provided as a resource. This allows us to accurately find burdened regions in many cancers. We use the wealth of ENCODE assays to accurately determine non-coding elements in each cell line (enhancers in particular). It also enables us to delineate regulatory networks involving transcription factors (TFs) and, to a lesser extent, RNA-binding proteins (RBPs). We represent these regulatory 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 score that represents the degree to which a regulator differs between normal and cancerous cells.

For each of the main ENCODE cell lines, our publically-disseminated resource consists of a list of accurately determined enhancers, a list of burdened regions, the regulatory TF network, as well as the most rewired TFs in this regulatory network. Collectively, these resources allow us to prioritize a few key elements as being associated with oncogenesis. Some of these are large-scale elements such as TFs and RBPs. Others are smaller in scale, including specific enhancers and even mutations within these enhancers. We evaluate our findings with small-scale studies, such as luciferase assays or TF knockdowns; most of the findings derived from our integrative analyses are experimentally validated.

## Data for comprehensive functional characterization in ENCODE

 From the levels of transcription to chromatin and nuclear organization, the ENCODE Project has provided a surge in functional annotation data across the human genome. Since more than XXX% of the cell lines provided by ENCODE are cancer cell lines, the raw data from ENCODE may serve as an invaluable resource for cancer research (see table S1). Here, we create a comprehensive list of raw datasets (Figure 1A) from ENCODE to interpret cancer genomes. In addition to the raw data being highly relevant to cancer, ENCODE annotations also have great breadth, expanding genomic insight from only the coding region to over xxx% of the annotated noncoding regions of the genome (Table S2). As such, this significant increase in data may greatly benefit the functional interpretation of cancer genomes.

Despite the comprehensive catalog of functional characterization assays in ENCODE, integrating its associated data into cancer research remains challenging for three main reasons. First, cancer is such a heterogeneous disease that it is necessary to use data from optimally-matched cell lines. ENCODE is imperfect for such analysis. We observe that there are only loosely matched tumor-normal pairs for some cancer types, and most cell lines lack data from certain experimental assays (Fig 1A). Therefore, it is necessary to create biologically relevant tumor-normal pairs, as well as to develop appropriate algorithms to learn from this suboptimally matched data. The second challenge arises as a result of the heterogeneous nature of the raw data from various experimental assays. The data must undergo rigorous de-duplication, unified processing, and proper normalization before accurate large-scale integration can be achieved. Lastly, the noncoding annotations in ENCODE, such as TF binding sites and enhancers, are provided as stand-alone regions in the genome, and thus lack clear direct linkages to protein-coding genes. Hence, direct interpretation of the functional impacts of mutations remains elusive.

In this study, we address these challenges in order to most effectively leverage ENCODE data as a resource for cancer research (Figure 1 B-C). To tackle the heterogeneity amongst data types, we constructed a comprehensive data matrix by normalizing raw signals of genomic features that severely confound somatic mutagenic processes (see Supp. File/Section(?) X ). The resultant data matrix can immediately be used for BMR correction. In contrast to previous approaches to annotation (many of which use only histone modification and chromatin accessibility data \{cite chromHMM}), we directly combine large-scale Enhancer-Seq experimental data with computational predictions to generate high-quality enhancer lists in several cell lines (see Supp. File/Section(?) X). To link these enhancers to genes, we perform enhancer target predictions by integrating evidence from expression profiles and chromatin status (see Supp. File/Section(?) X), and we further prune these predictions using Hi-C data for greater accuracy (see Supp. File/Section(?) X). To achieve improved functional interpretation, we use these high-quality linkages to construct what we term “extended genes” – coding regions matched with key regulatory elements in enhancers and promoters (Fig1 B). In addition, we also explore the full spectrum of the binding profiles in ENCODE data, and construct high-confidence gene regulatory networks for both transcription factor (TF) and RNA binding proteins (RBPs; Fig 1C and Fig X in Supp. File/Section(?) X).

## 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 attempts to discern which regions of the genome are more heavily 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, and they thus test 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.

In contrast, we integrate 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). 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). For example, in CLL, the correlation between observed and predicted mutation counts over 1-megabase bins ($ρ$) using replication-timing signals (from K562) increases from XX to XXX relative to that using data from HeLa-S3. Furthermore, despite their possibly high correlations in signal tracks, various functional characterization assays from ENCODE usually represent different biological mechanisms that affect mutagenic progresses (see Supp. File/Section(?) X). Thus, it is important to integrate these features 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 burden analyses (see below).

As oppose to separately testing standalone annotation categories, we employed our extended genes (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). 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 identified well-known drivers (such as TP53 and ATM) as well as other genes (such as BCL6) that were missed by the analysis of coding regions. In addition, 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.

## Extensive rewiring events of several transcription factors in cancer

We organized the TF regulatory network into hierarchies 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. Top-level TFs tend to more strongly influence the differential expression between tumor and normal cells. For instance, 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).

The human regulatory network specifies the combinatorial control of gene expression states from various regulators. Edge gain and loss analysis in matched tumor-normal TF networks may help to identify cancer-associated dysregulation. Hence, we investigated such rewiring events in TF networks using multiple formulations (see Supp. File/Section(?) X). Specifically, out of the 69 common TFs in K562 and GM12878 from ENCODE, we removed the general TFs and restricted our rewiring analysis to the remaining 61 (see Supp. File/Section(?) X). We first ranked TFs according to their respective number of lost and gained edges (Fig 3 A, see also Supp. File/Section(?) X). For example, 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). On the other hand, 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.

In addition, we also used a more complicated mixed-membership model to look more abstractly at local gene communities to re-rank the TFs (see Supp. File/Section(?) X). Similar patterns were observed using this model. We also observed that MYC (a well-known oncogene) became 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 found IKZF1 to be significantly associated with tumor progression (see Supp. File/Section(?) X).

Considerable TF rewiring is observed between tumor and normal samples. We investigate the potential underlying causes of this rewiring, and find that the majority of rewiring events result from changes in chromatin status, rather than from variant-induced motif loss or gain events (Fig. 3A). For example, JUND is a top gainer in K562 (with xxx gains and xx losses). We find that up to 30.5% (58.1%) of the gain (loss) events are associated with substantial expression changes (of at least 2-fold), and that xxx% have huge chromatin changes. Among those edges, only xxx variants were found in 100 CLL samples, and among these, up to xxx motif gain/loss variants could potentially affect rewiring events. Together, these analyses implicate a limited role for the effects of motif-changing events during the transition from normal to cancer cells.

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

To optimally leverage ENCODE data for studying various types of cancers, we extended our network analysis from strictly matched tumor-normal cell lines to more generalized networks by a regression-based learning method called RABIT (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 estimated the fraction of patients with differentially regulated target genes (see Supp. File/Section(?) X). The overall trends for the key TFs and RBPs discovered are given in Fig. 4A. The predicted impacts of regulators on tumor gene expression (as inferred from RABIT) are highly consistent with previous findings. For example, RABIT predicts the target genes of *MYC* to be 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 predicted 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 of our novel predictions, the predicted targets of the TF *ZNF687* were significantly up-regulated in breast and prostate tumors (star in Supp. Fig. 2). Thus, this integration between ENCODE and expression data has revealed many previously unidentified regulators with possible roles in driving cancer-specific expression patterns.

[JZ2MG: loregic to be here!]

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 result 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 network 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. Such unique patterns of co-association in cancer cell lines suggest a differential combinatorial code.

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

Here, we proposed a multi-resolution prioritization scheme to pinpoint key regulatory elements and SNVs that are important for carcinogenesis (workflow in Fig.5 A). We start by searching for key regulators (such as TF or RBPs) that are either massively rewired or drive tumor-normal differential expression. We then prioritize the functional elements (such as enhancers and TF-binding sites) governed by the key regulators through recurrence analysis. Lastly, we scrutinize each SNV therein by synthesizing features from annotation, conservation, and motif gain/loss events to pinpoint the impactful variants for small-scale functional characterization.

Using this framework, we 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 region is in chromosome 6, 13.5xxx (Fig. 5C). This enhancer is located in the noncoding region. Both histone modification and DHS signals implicate its regulatory role as being active (Fig. 5C). Our xxx based enhancer prediction method identified this as an enhancer, and we note that it was further validated using an EnhancerSeq experiment (Fig. 5D). Hi-C data indicate that this region is regulating an upstream gene XXX (DL to fill in). xx out of the XXX Chip-Seq experiments demonstrate significant binding events here, and the C to G mutation strongly 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

In the context of oncogenesis, this study comprehensively demonstrates the effectiveness of using ENCODE data to prioritize key regulatory elements and SNVs at different scales. Our scheme can immediately be applied to interpret the large number of noncoding variants from massive cohorts to pinpoint key elements for detailed functional characterization.