# Integrating ENCODE data to interpret regulatory changes in cancer

## Abstract

#/*=== abstract section 364 words ===*/

Articles have a summary, separate from the main text, of up to 150 words, which does not have references, and does not contain numbers, abbreviations, acronyms or measurements unless essential.

[JZ2MG: 1. Did not mention that our data can be extended to other cancer types, not just 4+1, I think this is important; 2, logically we say richness of data and then directly three points as in the three para, feel something is missing]

In general, one para, changes to 4 paras just for logic checkup.

In cancer, the impact of mutations in a limited number of coding genes is well understood; 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 data enables precise, tissue-matched non-coding background mutation rate calibration by removing the effects of confounders, such as replication timing. Furthermore, by integrating diverse ENCODE data, we are able to define high-confidence regulatory elements and their linkages to genes. This allows us to create definitions of extended gene neighborhoods, which are more sensitive than just coding regions for mutation recurrence analysis. Using this approach, we can identify additional genes associated with patient prognosis, beyond well-known highly mutated oncogenes (e.g., *BCL6* in leukemia).

Second, we integrate the ENCODE data to build a hierarchical regulatory network, 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.

Third, 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.

Articles are typically 3,000 words of text, beginning with up to 500 words of referenced text expanding on the background to the work (some overlap with the summary is acceptable), before proceeding to a concise, focused account of the findings, ending with one or two short paragraphs of discussion.

In total 2673 words,

#/*=== Introduction section 468 words ===*/

#/*=== Data section 426 words ===*/

#/*=== recurrence section 501 words ===*/

#/*=== Rewiring section 589 words ===*/

#/*=== expression section 250 (+111) words ===*/

#/*=== expression section 326 words ===*/

#/*=== expression section 143 words ===*/

## Introduction

Despite the millions of mutations discovered in cancer, only a very small fraction is readily interpretable in terms of their effects on known cancer-associated genes. An uncertainty lies in the degree to which the remaining variants contribute to cancer. Does this newly discovered pool of mutations simply constitute neutral passengers created as byproducts of oncogenic dysregulation? Or are there key variants among this pool that either directly drive tumorigenesis or otherwise affect the regulation of key cancer genes? The new data release of ENCODE Consortium may help to address these questions by providing accurate non-coding annotations and precisely linking these annotations to well-characterized protein coding genes.

The key feature of the ENCODE annotation set is that it relies on a wide variety of diverse experimental assays. For instance, instead of calling enhancers from just one histone modification mark, it offers the potential to integrate many different assays, such as DNase-seq, ChIP-seq, Enhancer-seq, Hi-C and ChIA-PET to more accurately define enhancers and link them to coding genes. Despite the comprehensive catalog of functional characterization assays in ENCODE, it is still challenging to directly integrate the ENCODE data to interpret cancer genome. Optimally link these cell line specific data, especially those from first tier cell lines with supreme data richness, to relevant cancer types is necessary before accurate large-scale integration can be achieved. Admittedly, some “matchings” are imperfect because these cell lines might not be as accurate as tissues from patients. However, it is not currently possible to conduct such a

wide variety of assays on actual tissue. Thus, such matchings provide a valuable opportunity to obtain accurate non-coding annotation in cancer.

Here, we endeavor to make the ENCODE resource as useful as possible for cancer research. We first match several cancers to top-tier ENCODE cell lines to better integrate relevant expression profiles and somatic mutations from known cohorts. We then develop a regression-based method to integrate comprehensive ENCODE data to calibrate an accurate background mutation rate (BMR), which can be released as a resource. This allows us to accurately find mutationally burdened regions in many cancers. We further use the wealth of ENCODE assays to accurately determine non-coding elements (enhancers in particular) and accurately link them 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 score that represents the degree to which a regulator differs between normal and cancerous cells. The top rewired regulators are found to be associated with patient prognoses. We finally propose a step-wise workflow to prioritize key regulators, functional elements, and mutations therein and validate their impacts through small-scale experiments.

## Data for comprehensive functional characterization in ENCODE

We first surveyed diverse sets of genome-wide assays spanning across 367 cell types released for ENCODE (details see supplementary file). To accurately characterize noncoding elements, we mainly focused blood, breast, liver, lung, and cervical cancers because their relevant tier 1 cell lines in ENCODE demonstrate best data breadth and depth (Fig 1A). We list the optimally matched tumor-normal pairs for each cancer type, and summarize the available experimental assays after de-duplication and unified processing. We observe that every cancer type contains a lot of highly heterogeneous data, but they also lack data from certain experimental assays (Fig 1A). Therefore, it is necessary to provide sensible normalization before large-scale integration and develop appropriate algorithms to learn from other sub-optimally matched data.

To tackle the heterogeneity amongst data types, 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 relying on single histone modification marks \{cite chromHMM}, we propose an ensemble-based method called ESCAPE, which performs large-scale data integration to identify active enhancers accurately. This integration involves prediction using a diverse collection of histone ChIP-seq, DNase-seq and Enhancer-seq. We further link these to genes not by simple correlation but by optimally investigating how the histone modification marks on the enhancer help predict the gene expression of the potential target gene. This group of potential linkages is then filtered through the results of the Hi-C experiments, which provide lower resolution but also a more accurate physical picture of the connections. (see Supp. File/Section(?) X). To achieve improved functional interpretation, we use these high-quality linkages to construct what we term “extended gene neighborhood” – coding regions matched with key regulatory elements, such as enhancers, promoters, and binding sites from regulators (Fig1 B). In addition, we also

explore the binding profiles in ENCODE data, and construct high-confidence gene regulatory networks for both TFs RBPs (Fig 1C and Fig X in Supp. File/Section(?) X).

Finally, 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 (see supplementary materials). Collectively, these resources allow us to prioritize a few key elements as being associated with oncogenesis, some of which have been validated using small-scale experimental assays (see table S1).

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

#/*=== recurrence section 501 words ===*/

One of the most powerful ways of identifying key elements and deleterious mutations in cancer is through recurrence analysis, which attempts to identify those regions of the genome that 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, 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.

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). 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 ($\rho$) using replication-timing signals (from MCF-7) increases from XX to XXX relative to that using data from HeLa-S3. Furthermore, despite the possibly high correlations in signal tracks, various functional characterization assays from ENCODE usually represent different biological mechanisms that affect mutagenic processes (see Supp. File/Section(?) X). Thus, it is important to integrate these features to infer BMR (Fig 1B). For example, $\rho$ 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). 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. 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

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 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 leukemia, 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 (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). 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 (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.

Alternatively, we also 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).

A remaining uncertainty lies in the underlying causes of this rewiring. It is necessary to investigate whether it is 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).

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

To optimally leverage ENCODE data for studying various cancer types, we extended our network analysis from strictly matched tumor-normal cell lines to more generalized networks. We can use both TF networks and those derived from RBPs – these are newly available data types in ENCODE, but for which we have no matched tumor normal pairs (see suppl. for description of merged networks).

Using a regression-based 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 select 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).

[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 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.

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

#/*=== conclusion section 326 words ===*/

The above description of the regulatory network and the optimum determination of mutation recurrence provide a way to prioritize key genomic features. 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 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). Note that both our HisShape enhancer prediction method and the ESPC algorithm (on EnhancerSeq experiment) predict this to be an enhancer (Fig. 5D). Hi-C and ChIA-PET data indicate that this region is regulating an 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

#/*=== conclusion section 143 words ===*/

In the context of oncogenesis, 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 is that we demonstrate how our intricate analysis can be improved, either by adding new data types or simply scaling an individual 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.