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

## Abstract Old version

We understand the impact of somatic mutations well in a very limited number of cancer genes; in contrast, the overwhelming number of mutations in cancer genomes occur in non-coding regions. The new release of the ENCODE data allows us to bridge these two facts. First, the new ENCODE data enables precise tissue-matched genome-wide background mutation rate calibration in a variety of tumors by separating the effect of well-known confounders, such as replication timing and chromatin status. Furthermore, by integrating large scale ChIP-seq, DNase-seq, Enhancer-seq, Hi-C, and ChIA-PET data from ENCODE, we are able to define with high confidence distal and proximal regulatory elements and their linkages to annotated genes. This enables us to create extended gene definitions, and we are able to show these are more sensitive than coding regions in terms of burdening analysis. In particular in leukemia, in addition to well-known drivers such as TP53 and ATM, it allows us to pick up other key genes such as BCL6, which can then be associated with patient prognosis. Second, we integrated the ENCODE data to build up a high confidence TF-gene regulatory network. This enabled us to identify highly rewired (i.e. target changing) TFs, such as NRF1 and MYC by comparing tumor and normal samples. By integrating large-scale chromatin features, we demonstrated that such massive rewiring events between tumor and normal cell lines are mainly attributable to the chromatin structure changes instead of direct mutational effect. Furthermore, we also found that TFs with more mutationally burdened binding sites (e.g., EZH2 and NR2C2) tend to be located at the bottom hierarchy of the TF regulation network. Third, using the ENCODE regulatory network, we developed integrative scoring workflow to prioritize key elements (and mutations in them) according to their role in cancer and then validated these in small-scale studies. In particular, we prioritized ZNF687 as a key TF for breast cancer and SUB1 as a key RNA binding protein for liver and lung cancer and validated them through siRNA knockdown experiments. Finally, we identified key enhancers and mutations in them in breast cancer and then validated their functional effect through luciferase assays.

## 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.

Jing Zhang 1/26/2017 4:33 PM

**Comment [1]:** This sentence reads a little bit weird since the "though" part is pretty long, but I am not a native speaker, so please suggest

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

Jing Zhang 1/26/2017 3:08 PM
**Comment [2]:** Need to confirm with Kevin Yip

Jing Zhang 1/26/2017 4:36 PM
**Comment [3]:** It is more coherent to say enhancers and promoters, since we just mentioned how to set up enhancers. But this will limit our extended gene definition fur further extension. Somehow, we could potentially say that the extended gene can be easily incorporated with other regulatory annotations

Jing Zhang 1/26/2017 4:40 PM
**Comment [4]:** Strong is not a good word here? Please suggest alternative.

Jing Zhang 1/26/2017 4:45 PM
**Comment [5]:** Is this correct in grammar?

Jing Zhang 1/26/2017 4:50 PM
**Comment [6]:** We should differentiate regulatory elements from regulators. RE means for example TFBS, but regulators means TF/RBP

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.