# An integrative ENCODE resource for cancer genomics

## Introduction

The 2012 ENCODE release provided RNA-seq, histone and transcription factor (TF) ChIP-seq, and DNase-seq over several cell lines to comprehensively annotate the noncoding regions in the human genome for the first time. The current release broadens the number of cell lines for these assays and considerably expands available tissue data. It also greatly increases the depth of coverage of assays by adding new approaches, such as STARR-seq, Hi-C, and eCLIP. The integration over many assays provides an unparalleled opportunity to develop accurate annotations in a cell-type specific manner, which is particularly useful for interpreting genomic variants associated with disease and cell-state changes that underlie many disease processes. Deep integration over many assays also allows us to connect many regulators and non-coding elements into multi-modal networks, including proximal (TF/RBP-gene) and distal ones (enhancer-gene or TF-enhancer-gene).

Here, focusing on data-rich ENCODE cell types, we performed deep integration over many functional assays to characterize the noncoding genome, which may serve as a valuable resource for disease studies. Cancer is one of the best applications to illustrate the key aspects of this integrative ENCODE resource. Unlike many other diseases, cancer is very much a disease of whole-genome dysregulation. Cancer cells may display aberrant behaviors of key regulators, extensive remodeling of epigenetics, and abnormal transitions between cell states. The wealth of ENCODE functional characterization data allows direct measurement of chromatin status, regulatory changes, and expression perturbations for individual genes. It may also be used to construct comprehensive high-quality networks, to capture tumor-to-normal alterations from a more global perspective. ENCODE data may thus help better elucidate the biological mechanisms of cancer initiation and progression.

Therefore, we present an integrative *ENCODE* companion resource for *C*ancer genomics (ENCODEC). This resource consists of various annotations, networks, and code bundles available online. Its compact noncoding annotations and extended gene definitions can potentially increase the statistical power to interpret variant (both germline and somatic) and expression data in cancer. Its comprehensive experiment-based networks allow us to depict global alterations in network rewiring, hierarchies, and TF/RBP dysregulations. This lead us to the detection of many key cancer genes that are potentially of prognostic value, including the well-known TF MYC and novel RBP SUB1. We also discovered noncoding SVs that activate key oncogenes and SNVs that affect enhancer activity, and successfully validated them through CRISPR and luciferase assays. We further took advantage of the wealth of the ENCODE expression and chromatin state data to cluster various cell types, and highlight a consistent stem-like transition during normal-to-tumor development in many cancer types.

## The breadth and depth of the ENCODE resource

Figure 1 illustrates two key dimensions of the overall ENCODE data set in relation to cancer: its breadth across cell types and depth across assays. Integration over the many different cell types, or many different assays, may contribute novel insights regarding cancer biology.

For example, it is challenging develop a background mutation rate model (BMR) for somatic recurrence in cancer – the somatic mutation process can be influenced by numerous confounders (in the form of both external genomic factors and local sequence context factors), which can result in false conclusions if not appropriately corrected. Researchers have suggested various regression approaches to integrate genomic features for accurate BMR calibrations. ENCODE has dramatically increased available BMR-associated features from less than 200 to over 2,000. Here we show in figure 1 that simple integration of these data sets progressively provides more accurate BMR estimates.

In addition, one can also integrate across assays to investigate how different patterns of epigenetic marks are related to structural variants (SV) in strictly matched cell types. ENCODE has various whole genome assays, which can contribute to accurate SV calls after proper integration with whole genome DNA sequencing data. Interestingly, we found that K562 breakpoints are associated with H4K20me1, which is an activating histone marker only in K562, but not in other cell types.

## An extended gene annotation and its applications in cancer

From the wealth of the ENCODE experiments in the data-rich cell lines, we constructed a deep, integrated annotation with two key characteristics – 1) it is compact to precisely pinpoint the functional sites, and 2) it extends the conventional gene annotation by linking the discontinuous noncoding regulatory regions to genes. Different from conventional gene annotations, which are uniformly defined (although differentially expressed) across cell types, extended gene annotations are highly dynamic and may considerably change from cell to cell. This may benefit the statistical power of many analyses in cancer.

Our annotation is compact because we defined “core” regions of individual regulatory elements that are enriched for functional sites and reduce false positives rates via strict quality control (see suppl. sect. xxx). This can be doneeither by incorporating novel advanced assays such as eCLIP and STARR-seq, or integrating over tens of functional assays, such as DNase-seq and ChIP-seq (see suppl. sect. xxx). We have shown that such an annotation can effectively reduce the noise-to-signal ratio and reduce multiple test correction burden to achieve a better power for various analyses.

On the extended side, a second step of our integrative annotation entails linking the above compact elements to define a high-quality extended gene neighborhood by integrating computational predictions with direct experimental evidence on physical interactions from Hi-C and ChIA-PET experiments (see suppl. sect. xxx). Such a gene-centric approach not only allows us to improve upon existing knowledge of the genetic regions involved in cancer, but also enables a joint evaluation of distributed yet biologically connected genomic regions. This leads to increased power in many analyses (see suppl. sect. xxx).

To illustrate the value of our resource, we first compared the enrichment of cancer GWAS SNPs with respect to various annotations. The enrichment in protein-coding genes significantly increases as we add more relevant annotations for breast cancer and leukemia (Fig. 2C). This trend is more pronounced when the newly added proximal and distal noncoding annotations are from matched cell types. One may further subset the genes according to different subcategories associated with cancer, and identify enrichments. For instance, we observed a significant enrichment in genes from the Cancer Gene Consenus (CGC) in breast cancer based on the extended gene annotation, which was not possible using the conventional coding regions annotation.

We also showed that our extended gene annotation can provide better stratification of gene expression from mutational signals in cancer patients compared to single annotation categories. We combined the mutational and expression profiles from large cohorts, such as TCGA, and found that mutational status in our extended gene definition can explain the expression differences for a larger number of genes than other annotations, such as annotations of coding sequences (CDS). One example of the explanatory potential of the extended gene is seen for the splicing factor SRSF3, which has been shown to affect liver cancer progression \{cite}. In HepG2, aggregating mutational burden within its extended gene annotation exhibits greater significance relative to gene expression, compared to any single annotation category (p=xxx, one sided Wilcoxon test).

By integrating our compact annotation sets, BMR estimates, and accurate extended gene definitions, we obtained larger power for detecting genomic regions (coding and non-coding) that are mutationally burdened. Fig. 2E illustrates genes that are mutationally burdened using our extended gene definitions in several well-known cancer cohorts. For example, in the context of chronic lymphocytic leukemia (CLL), our analyses identified well-known highly mutated genes (such as TP53 and ATM) that have been reported in previous analyses. More importantly, the increased power provided by the extended-gene annotation allowed us to detect genes that would otherwise be missed by an exclusively coding analysis. An example of this is the well-known cancer gene BCL6, which may be associated with patient survival (Fig. 2E).

One can get a physical sense of the importance of the extended gene environment by looking at a situation where genomic variant rearranges the extended gene without affecting coding regions. We found such an example in the breast cancer cell line T47D, where a 130Kb heterozygous deletion changes the chromosome to link a distal enhancer to the promoter and results in the activation of the well-known oncogene ERBB4 (Fig. 2F). This heterozygous deletion is located around 45Kb downstream from the ERBB4 promoter region and potentially merges two Hi-C TADs in an allele-specific way. This change occurs in cancer cells but not in normal cells. We tested this hypothesis through CRISPR editing, by excising an 86bp sequence that contains the CTCF binding sites at the boundary of the two Hi-C TADs from the wild-type allele in T47D cells. This excision confirmed the elevated ERBB4 expression upon CRISPR deletion (as measured by PCR).

## Leveraging ENCODE networks to prioritize key regulators

Building on the extended gene annotation, we constructed detailed regulatory networks. Specifically, we incorporated both distal and proximal networks by linking TFs to genes. This was accomplished either directly by TF-promoter binding or indirectly via TF-enhancer-gene interactions in each cell type (see suppl. sect. xxx). We then pruned the full network to the strongest interactions using a signal shape algorithm which keeps the strongest peaks by weighting their occurrence by the general TF binding profiles (see suppl. sect. xxx). In addition, we reconciled our cell-type specific networks to form a generalized pan-cancer network. Similarly, we also defined an RNA-binding protein (RBP) network from eCLIP experiments. Compared to others, our ENCODE TF and RBP networks can capture more literature-supported regulations and correlate better with knockdown experiments (see suppl. sect. xxx).

We analyzed the overall regulatory network by systematically arranging it into a hierarchy (Fig. 3A). Here, regulators are placed at different levels such that those in the middle tend to regulate regulators below them and, in turn, are more regulated by regulators above them3 (suppl. sect. xxx). In this hierarchy, we found that the top-layer TFs are not only enriched in cancer-associated genes (P=xxx, Fisher’s exact test) but also more significantly drive differential expression in model cell types (P=xxx, one sided Wilcoxon Test). These networks also enable investigation of the connections between TFs and RBPs. Interestingly, we found that there are less top-level TF-RBP interactions, as compared to middle and bottom level ones.

Our networks also enable gene-expression analyses in tumor samples. We used a regression-based approach to systematically search for the TFs and RBPs that most strongly drive tumor-normal differential gene expression (suppl. Sect. xxx). For each patient, we tested the degree to which a regulator’s activity correlates with its target’s tumor-to-normal expression changes. We then calculated the percentage of patients with these relationships in each cancer type, and present the overall trends for key TFs and RBPs in Fig. 3A.

As expected, we found that the target genes of MYC are significantly up-regulated in numerous cancer types, consistent with its well-known role as an oncogenic TF4,5. We further validated MYC's regulatory effects using knockdown experiments in breast cancer (Fig. 3). Consistent with our predictions, the expression of MYC targets is significantly reduced after MYC knockdown in MCF-7 (Fig. 3B). We analyzed the RBP network in a manner that was similar to the TF network, and found key regulators associated with cancer (see suppl.). For example, the ENCODE eCLIP profile for the RBP SUB1 has peaks enriched on the 3'UTR regions of genes, and the predicted targets of SUB1 were significantly up-regulated in many cancer types (Fig. 3D). As an RBP, SUB1 has not previously been associated with cancer, so we sought to validate its role. Knocking down SUB1 in HepG2 cells significantly down-regulated its targets (Fig. 3D), and the decay rate of SUB1 targets is lower than those of non-targets (see suppl.). Moreover, we found that up-regulation of SUB1 targets may lead to decreased patient survival in some cancer types (Fig. 3D).

We then used the regulatory network to investigate how these prioritized key regulators interact with other genes. For TFs, we first looked at how MYC's target genes are co-regulated by a second TF. These three-way co-regulatory relationships are shown in Fig. 3C. We found that the most common pattern is the well-understood feed-forward loop (FFL). In this case, MYC regulates both another TF and a common target of both MYC and that TF (Fig. 3C). Since MYC amplification has been discovered in many cancers, understanding which TFs appear to further amplify its effects may yield insights for efforts aimed at MYC inhibition5. Most of the FFLs involve well-known MYC partners such as MAX and MXL1. However, we also discovered many involving NRF1. Upon further examination, we found that that the MYC-NRF1 FFL relationships were mostly coherent, i.e., "amplifying" in nature (Fig. 3C ii). We further studied these FFLs by organizing them into logic gates, in which two TFs act as inputs and the target gene expression represents the output6 (see suppl.). We found that most of these gates follow either an OR or MYC-always-dominant logic gate.

Similarly, with respect to RBPs, we found that the top co-regulatory partner of SUB1 is MYC (SUB1 is a direct target of MYC in many cell types, see suppl. sect.). SUB1 and MYC together form many FFLs in the regulatory network. We hypothesized that MYC can bind to the promoter regions of key oncogenes to initiate their transcription, whereas SUB1 binds to 3UTRs to stabilize oncogenes at the level of RNA transcripts. Such collaboration between MYC and SUB1 results in overexpression of several key oncogenes and leads to proliferation of cancer cells (see suppl. sect. xxx). To validate this hypothesis, we knocked down MYC and SUB1 separately in HepG2 and used qPCR to quantify changes in gene expression. As expected, the expression of oncogenes (such as MCM7, BIRC5, and ATAD3A) is significantly reduced (Fig. 3E).

**Cell-type specific regulatory networks highlight extensive rewiring events during oncogenesis**

For data-rich cell types with numerous TF ChIP-seq experiments, we built cell-type specific regulatory networks. These networks enable direct comparisons of network rewiring during oncogenesis. To achieve the best pairing given the existing data, we constructed a "composite normal" by reconciling multiple related normal cell types (see suppl.). Although the pairings are only approximate, many of them have been widely used in prior studies (see suppl.). Furthermore, they leverage the extensive functional characterization assays in ENCODE to provide us with a unique opportunity to study regulatory alterations in cancer on a large scale for the first time.

In particular, we measured the signed fractional number of edges changes for "tumor-normal pairs” (which we call the "rewiring index"), to study how TF targets change in the oncogenic transformation. In Fig. 4A, we ranked TFs according to this index. In leukemia, well-known oncogenes (such as MYC and NRF1) were among the top edge gainers, while the well-known tumor suppressor IKZF1 is the most significant edge loser (Fig. 5A). Mutations in IKZF1 serve as a hallmark of various forms of high-risk leukemia7,8. We observed a similar rewiring trend using distal, proximal, and combined networks (details in suppl.). This trend was also consistent across a number of cancers: highly rewired TFs such as BHLHE40, JUND, and MYC behaved similarly in lung, liver, and breast cancers (Fig. 5).

In addition to direct TF-to-gene connections, we also measured rewiring using a more complex gene-community model. Here, the targets within the regulatory network were characterized in terms of heterogeneous modules from multiple genes (so called "gene communities" ). Instead of directly measuring the changes in a TF's targets between tumor and normal cells, we determined the changes in its gene communities via a mixed-membership model (see suppl.). Similar patterns to direct rewiring were observed using this model (Fig. 5A).

We organized the cell-type specific networks into hierarchies, as shown in Fig. 5B. Specifically, in blood cancer, the more mutationally burdened TFs sit at the bottom of the hierarchies, whereas the TFs more associated with driving cancer gene expression changes tend to be at the top. In addtion, we found that the strongest edge gainers and edge losers in rewiring events often sit at the top level of the network hierarchies.

We found that the majority of rewiring events were associated with noticeable gene-expression and chromatin-status changes, but not necessarily with mutation-induced motif loss or gain events (Fig. 5A). For example, JUND is a top gainer in K562. The majority of its gained targets in tumor cells demonstrate higher levels of gene expression, stronger active and weaker repressive histone modification mark signals, yet few of its binding sites are mutated. This is consistent with previous work that indicates most non-coding risk variants are not well-explained by a mutational [LU4] model9. With a few notable exceptions (see suppl.), we found a similar trend for the rewiring events associated with JUND in liver cancer and, largely, for other factors in a variety of cancers.

## Stemness measurement during oncogenic transformation through regulatory networks

A prevailing decades-old paradigm has held that at least a subpopulation of tumor cells has the ability to self-renew, differentiate, and regenerate in a manner similar to that in stem cells. One of the strengths of ENCODE is its many stem cell lines, including the H1 stem cell, which is one of the most data-rich cell types. We leveraged the large number of stem cells in ENCODE and the additional data available with this ENCODE release, to place tumor-associated cell types relative to normal cells and stem cells in cell space. We projected the RNA-seq data by Reference Component Analysis (RCA, \cite{nat rca paper}) to cluster various cell types according to the similarity of their transcriptomes. We found that various types of stem cells, including data-rich H1 cells, form a tight cluster (blue in Fig. 5A). As is observed from Fig. 5, there is potentially a trend where the transition from normal to tumor cells is moving toward a stem cell. This is true for a variety of different cancers. This observation is consistent with previous efforts using expression and methylation analysis. Interestingly, we observed a consistent (or even stronger) pattern from the proximal and distal chromatin data, which can be viewed as the underlying cause of the observed gene expression changes.

It is also well-known that dysregulation of key oncogene TFs are hallmarks of tumor progression. Key genes, such as MYC, initiate overexpression of other oncogenes in tumor cells. To test the hypothesis that oncogenic TFs contribute to the state of cell differentiation, we measured the perturbations introduced by oncogenic TFs through expression comparisons before and after TF knockdowns. Interestingly, the overall expression profiles reverted slightly back towards normal state upon oncogene knockdowns.

## Step-wise prioritization scheme and result validations

Collectively, we propose a step-wise prioritization scheme that allows us to pinpoint key regulator genes, noncoding elements, and single nucleotides associated with oncogenesis, as schematized in Fig. 6A. Specifically, we first highlighted regulators that are either frequently rewired, or located in network hubs, or sit at the top of the hierarchy, or significantly drive expression changes in cancer. We then prioritized functional elements associated with these regulators that are either highly mutated in tumors, or undergo large changes in expression or regulation. Finally, on a nucleotide level, by estimating their ability to disrupt or introduce specific binding sites, we pinpoint impactful genomic variants at a fine scale.

##

To demonstrate the utility of our ENCODE resource, we instantiated our prioritization workflow in a few select cancers and experimentally validated the results. In particular, as described above, we subjected some key regulators, such as MYC and SUB1, to knockdown experiments to validate their regulatory effects (Fig. 3B and 3D). We highlighted large scale structural variations that potentially disrupt oncogene insulation and validated their effects through CRISPR engineered deletions (Fig. 2E). Finally, we selected key SNVs based on their disruption of enhancers with strong influence on gene expression. These SNVs were prioritized based on mutation recurrence in breast-cancer cohorts, as well as enhancer motif disruption scores. [LU12] Of the eight motif-disrupting SNVs that we tested, six exhibited consistent up- or down-regulation relative to the wild-type in multiple biological replicates.

One particularly interesting example, illustrating the value of ENCODE data integration, is in an intronic region of CDH26 in chromosome 20 (Fig. 6C). The signal shapes for both histone modification and chromatin accessibility (DNase-seq) data indicate its active regulatory role as an enhancer in MCF-7. This was further confirmed by STARR-seq (Fig. 6C). Hi-C and ChIA-PET linkages indicated that the region is within a topologically associated domain (i.e., a “TAD”) and validated a regulatory connection to the breast-cancer-associated gene SYCP2. We further observed strong binding of many TFs in this region in MCF-7. Motif analysis predicts that a common mutation in breast cancer affects this region, and significantly disrupts the local binding affinity of several TFs, such as FOSL2 (Fig. 6C). Luciferase assays demonstrated that this mutation introduces a 3.6-fold reduction in expression relative to the wild-type, indicating a strong repressive effect on enhancer functionality.

## Conclusion

This resource underlines the importance of deep data integration over many novel assays to understanding cancer genomes. A compact annotation with extended gene definitions can benefit various analyses in cancer. The value of our resource can go far beyond noncoding variant interpretation, to allow exploration ofgenome-wide gene dysregulations, epigenetic remodeling, network perturbations, and cell state transitions. These are also keys aspects of genome interpretation that are tightly associated with oncogenesis. The successful validation of our prioritized regulator genes, noncoding elements, and SNVs demonstrates the value of our resource.

A key caveat related to our resource concerns network rewiring in cell-type specific networks. The utility of these networks in cancer is based on associating them to various cancer types and then pairing a specific cancer network with a composite normal. Both pairings are approximate. Moreover, cancer is well-knowns for its heterogeneity. Tumor cells from a given patient usually show distinct molecular, morphological, and genetic profiles. Linking cancer to one specific cell-type may not fully capture the heterogeneity seen in cancer. To place this limitation in context, it can even be challenging to obtain a representative match between tumor and normal tissues taken from a single patient. Further technological advances, such as single cell sequencing, may help to provide more biological insights at a higher resolution. Nevertheless, we feel that our networks currently provide the best available view of the regulatory changes in oncogenesis since no other system currently has data at this scale.

In general, our study underscores the value of large-scale data integration, and we note that expanding the scale of our approach in a number of dimensions is straightforward. For example, we successfully formed compact annotations and regulatory networks for model systems already replete with advanced functional assays like eCLIP and STARR-seq; our methods can be readily extended to further model systems when they are similarly assayed in the future. Given the rewiring formalism presented here, it should be straightforward to expand the analysis to greater numbers of TFs. (In fact, the re-wiring formalism provides a way of selecting candidate key TFs and cell types.) We anticipate that this will provide a clearer and more accurate picture of the spectrum of regulators that are affected by extensive chromatin changes, and thus help prioritize research efforts in cancer.