# 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 annotate the noncoding regions in the human genome. The current release broadens the number of cell types for these assays and considerably expands available tissue data. It also greatly increased the depth coverage of assays by adding novel assays, such as STARR-seq, Hi-C, and eCLIP. The integration over a massive number of assays provides an unparalleled opportunity to develop compact and accurate annotations in a tissue-specific manner, which is particularly useful for interpreting genomic variants associated with disease. Deep integration over many assays also allows us to connect many of the regulators and non-coding elements into elaborate networks, including proximal (TF/RBP-gene) and distal ones (enhancer-gene or TF-enhancer-gene).

Hence, focusing on several data-rich ENCODE cell types, we performed deep integration over tens of functional assays to deeply characterize the noncoding genome, which may serve as a valuable resource for disease studies. Our resource is particularly well suited to studying cancer for several reasons. First, many of the most data-rich cell types are associated with cancer cells, including cell types from blood, breast, liver, and lung (Fig. 1). Second, the wealth of ENCODE data, such as replication, epigenetic, and transcriptional profiles, may be used to inform our understanding of cancer mutational processes for both single nucleotide variations (SNV) and structural variations (SV). Third, the wealth of ENCODE data can be used to measure epigenetic remodeling and cell-state transitions, which are implicated in oncogenesis. Lastly, high-quality regulatory networks can be reliably constructed from thousands of experiments to provide a systems-level perspective of cancer. One may thus directly measure the perturbations of individual regulators and entire networks to better elucidate the biological mechanisms of cancer initiation and progression.

Therefore, we present an integrative ENCODE companion resource for Cancer genomics (ENCODEC). This resource consists of various annotations, networks, and code bundles available online. It allows us to prioritize key regulators, non-coding elements and variants in relation to cancer. These prioritized elements tend to be burdened with germline and somatic mutations, or to sit at central positions in the regulatory network, or to be associated with large epigenetic, expression, or distal interaction changes. We find that this prioritization uncovers interesting interactions between the well-known oncogene TF MYC and the RNA binding protein (RBP) SUB1 that had not been known before. Finally, it shows how the overall chromatin and epigenetics in a cell changes, moving the cell into a more stem-like state.

## The breadth and depth of the ENCODE resource

Our work takes advantage of the breadth and depth of the ENCODE resource and customizes it for cancer research. First, the somatic mutation process can be influenced by numerous confounders. ENCODE contains more than 2,000 distinct types of epigenetic and replication timing data sets. Aggregating these together in a simple model, one can predict background mutation rates (BMRs) for often highly heterogeneous tumors more accurately than a smaller number of features. As seen in Fig. 1, BMR estimation accuracy even continues to improve after even 15 or 20 features are added. Conversely, one can aggregate across assays within a particular cell type to uncover the mutational mechanisms underlying SVs. For instance, once can aggregate the histone markers across structural variants released by ENCODE, which are called by integrating various types of assays (see. Suppl. sect. xxx). 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.

We also utilized the depth of the ENCODE data to provide compact and accurate annotations with superior properties relative to other annotations (see suppl. sect. xxx). For example, we explored the full catalogue of ENCODE eCLIP experiments to systematically define the post-transcription regulome with noticeably improved resolution and precision over previous efforts (see suppl. sect. xxx). Additionally, in several well-known cancer cell types, we developed a match filter based algorithm to incorporate a large battery of histone marks with chromatin accessibility data for better enhancer predictions. We the further combined STARR-seq data, which directly measures the genome-wide enhancer activities, to accurately define core enhancers. We then incorporated Hi-C and ChIA-PET data to make accurate enhancer-gene linkage predictions.

## An extended gene annotation and its applications

Much current knowledge of disease has been derived by focusing on protein-coding regions. To broaden the scope of elements studied, we also linked our above noncoding annotations to genes in order create a gene-centric annotation (which we call the extended gene). Our extended gene annotation includes both proximal and distal, transcriptional and post-transcriptional level annotations (Fig.

First, our extended gene definitions include many tissue-specific proximal and distal noncoding regulatory elements that are useful for interpreting cancer-associated GWAS variants. To illustrate this, we calculated the enrichment of cancer GWAS SNPs with respect to various annotations. We observed a positive relationship between increasing GWAS SNP enrichment and the number of included annotations (Fig 2C). We note that, in contrast to unified gene definitions that are identical across different cell types, tissue-specific experiments allow us to build a highly dynamic extended gene definition that is unique to specific cancer types. Indeed, the greatest enrichment of GWAS SNPs is achieved using tissue-matched samples (Fig 2C).

Second, we used the extended gene annotations as a single test unit for recurrence analysis, rather than testing all regions separately. Such a unified scheme enables joint evaluation of the mutational signals from distributed yet biologically connected genomic regions. Fig. 2B illustrates the larger number of known cancer-related genes detected in several cancer cohorts, relative to those derived from the coding regions or promoter sites. For instance, in the context of chronic lymphocytic leukemia (CLL), our joint detection approach identified well-known highly mutated genes (such as TP53 and ATM) \{cite}. More importantly, this joint detection approach allowed us to detect genes that would otherwise be missed by exclusively focusing on coding regions. As an example, we identified the well-known cancer gene BCL6, which may be associated with patient survival (Fig. 2B and refs. 1-3).

Third, our extended gene annotation can provide better stratification of gene expression from mutational signals in cancer patients compared to single annotation categories. For instance, 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).

Finally, we found an example of how an SV introduced extended gene change that may lead to oncogene activation. ERBB4 is a well-known oncogene in many cancer types \{cite}. (Fig. 2D). We identified a 130Kb heterozygous deletion (~ 45Kb downstream from the TSS) that potentially merges two Hi-C TADs and links a distal enhancer to the ERBB4 promoter in T47D cells, but not in normal cells (xxx color track from 4C experiment). We therefore hypothesized that this heterozygous deletion disrupts the insulation of ERBB4 from distal regions, thereby activating its allele-specific expression. We tested this hypothesis through CRISPR editing, by excising an 86bp sequence from the wild-type allele in T47D cells. This excision resulted in elevated ERBB4 expression (as measured by PCR). Our results suggest that ERBB4 activation in T47D may at least in part be due to the 130 kb deletion that disrupts its insulation.

## 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 these networks to include only the strongest edges using a signal shape algorithm1 (see suppl.). 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. eCLIP is an enhanced CLIP protocol that provides single-nucleotide resolution of the RBPs binding signatures2. Compared to imputed networks derived from gene expression or motif analyses, our ENCODE TF and RBP networks provide experimentally based regulatory linkages between functional elements.

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).

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.).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.

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. In all cancer types, we found that the shared target genes' expressions are strongly positively correlated with MYC, while they showed only limited correlation with the second TF (as determined by partial correlation analysis, see suppl.). We further investigated the regulatory control pathways of these triplets. 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. Thus, the ENCODE regulatory network not only helps identify key regulators, but also illustrates how these may work in combination.

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) and also in terms of a simpler co-binding approach (see suppl.).

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 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. On a related note, we organized the cell-type specific networks into hierarchies, as shown in Fig. 4B. Specifically, in blood cancer, the more mutationally burdened TFs sit at the bottom of the hierarchy, whereas the TFs more associated with driving cancer gene expression changes tend to be at the top.

## 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. We projected the xxx RNA-seq data by Residual Component Analysis (RCA) 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). Interestingly, we observed that tumor cells (green in Fig. 5A) are located more proximal to the stem group than its normal counterpart (yellow), which is consistent with recent discoveries \cite{TCGA stemness}. Furthermore, we extended our analysis from transcriptome to both proximal and distal regulatory networks and observed a strong pattern: tumor cells tend to cluster together around stem cells, unlike normal cells.

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 toward normal state upon oncogene knockdowns

## Step-wise prioritization scheme pinpoints deleterious features associated with oncogenesis

Summarizing the above, our companion resource consists of annotations of (1) overall somatic and germline mutational burden scores; (2) accurate and compactly defined regulatory elements by integrating various novel functional assays, including eCLIP and STARR-seq; (3) enhancer-target-gene linkages and extended gene neighborhoods that are obtained by integrating Hi-C and multi-histone-mark experiments; (4) tumor-normal differential expression, chromatin, and 3D structural changes; (5) TF regulatory networks, both merged and cell-type specific, based on both distal and proximal regulation; (6) an analogous but less-developed network for RBPs; (7) attributes of TF/RBPs derived from network analysis, such as position in the network hierarchy, regulatory potential, and rewiring status. All the resources mentioned above are available online through the ENCODE website as simple flat files and computer codes (see suppl.).

Collectively, these resources allowed us to prioritize key genomic features associated with oncogenesis at regulator, element, and nucleotide levels. Our prioritization workflow is schematized in Fig. 6A. We first searched for key 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 gene expression, or TF binding, or chromatin status. Finally, on a nucleotide level, by estimating their ability to disrupt or introduce specific binding sites, we pinpoint impactful genome variants at a fine scale level.

## Small-scale validation experiments on prioritized regulators and elements

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. 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 SYCP210. 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 highlights the value of deep data integration over many novel assays to annotate noncoding elements of the genome. We provided accurate tissue-specific extended gene annotations and extensive regulatory networks through integration of thousands of experiments. We believe that one of the best applications of our resource is to cancer research.

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 particular cancer types and then pairing a specific cancer network with a composite normal. Both correspondences are approximate. Nevertheless, we feel that our networks currently provide the best available view of the regulatory changes in oncogenesis. No other system has this scale of TF-ChIP data. Moreover, the heterogeneous nature of cancer means that tumor cells from a given patient usually show distinct molecular, morphological, and genetic profiles11. Cell-type specific or tissue-type specific analyses may not fully capture the heterogeneity seen in cancer. However, 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.

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 actually provides a way of selecting candidate 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.

1 Cheng, C., Min, R. & Gerstein, M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* **27**, 3221-3227, doi:10.1093/bioinformatics/btr552 (2011).

2 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514, doi:10.1038/nmeth.3810 (2016).

3 Cheng, C. *et al.* An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. *Genome Biol* **16**, 63, doi:10.1186/s13059-015-0624-2 (2015).

4 Dang, C. V. MYC on the path to cancer. *Cell* **149**, 22-35, doi:10.1016/j.cell.2012.03.003 (2012).

5 McKeown, M. R. & Bradner, J. E. Therapeutic strategies to inhibit MYC. *Cold Spring Harb Perspect Med* **4**, doi:10.1101/cshperspect.a014266 (2014).

6 Wang, D. *et al.* Loregic: a method to characterize the cooperative logic of regulatory factors. *PLoS Comput Biol* **11**, e1004132, doi:10.1371/journal.pcbi.1004132 (2015).

7 de Rooij, J. D. *et al.* Recurrent deletions of IKZF1 in pediatric acute myeloid leukemia. *Haematologica* **100**, 1151-1159, doi:10.3324/haematol.2015.124321 (2015).

8 Boer, J. M. *et al.* Prognostic value of rare IKZF1 deletion in childhood B-cell precursor acute lymphoblastic leukemia: an international collaborative study. *Leukemia* **30**, 32-38, doi:10.1038/leu.2015.199 (2016).

9 Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343, doi:10.1038/nature13835 (2015).

10 Masterson, L. *et al.* Deregulation of SYCP2 predicts early stage human papillomavirus-positive oropharyngeal carcinoma: A prospective whole transcriptome analysis. *Cancer Sci* **106**, 1568-1575, doi:10.1111/cas.12809 (2015).

11 Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328-337, doi:10.1038/nature12624 (2013).