**An integrative ENCODE resource for cancer genomics: interpreting regulatory changes and non-coding mutations**

**Introduction**

A small fraction of mutations associated with cancer have been well characterized, particularly those in coding regions of key oncogenes and tumor suppressors. However, the overwhelming majority of mutations in cancer genomes – especially those discovered over the course of recent whole-genome cancer genomics initiatives – lie within non-coding regions \cite{25261935}. Whether these mutations substantially impact cancer progression remains an open question \cite{26781813}.

Several recent studies have begun to address this question by incorporating limited functional genomics data\{cite 25261935, 27064257, 27807102}. For example, Hoadley *et al.* integrated five genome-wide platforms and one proteomic platform to uniformly classify various tumor types \{cite 25109877}. Torchia *et al.* integrated various genomic and epigenetic signals to identify promising therapeutic targets in rhabdoid tumors \cite{27960086}. Lawrence *et al.* incorporated large-scale genomics profiles to identify cancer drivers \{cite 23770567}. However, there is no systematic integration of thousands of functional genomic data sets from a broad spectrum of advanced assays to interpret cancer genomes.

The rich functional assays and annotation resources developed by the ENCODE Consortium allows us to characterize these non-coding regions at great depth \cite{22955616}. Given that around eighty percent of ENCODE cell lines are associated with cancerous tissues (see supplement) ENCODE data are particularly suited for cancer research. In the initial release of the ENCODE annotation set, this was predominantly accomplished by using RNA-seq and ChIP-seq assays on a limited number of cell types \cite{22955616}. The new release of ENCODE took two new directions. First, it considerably broadened the number of cell types with RNA-seq, ChIP-seq, and DNase-seq assays. As such, the main ENCODE encyclopedia aims to utilize these to provide a general, unified annotation resource applicable across many cell types. Secondly, ENCODE also expanded the number of advanced assays (such as STARR-seq, Hi-C, ChIA-PET, eCLIP and RAMPAGE) on several top-tier cell lines. Many of these top-tier cell lines are associated with various cancer types (Figure 1A), including those of the blood (K562), breast (MCF-7), liver (HepG2), lung (A549), and cervix (HeLa-S3) . In addition, another data-riched top-tier cell line is H1-hESC, a human stem cell line. For decades, the prevailing paradigm has held that at least a subpopulation of a tumor’s cells have the ability to self-renew, differentiate, and regenerate, in a manner that is similar to current thinking for normal stem cells \cite{24333726}. Hence, H1-hESC can serve as a valuable comparison when investigating degree to which their oncogenic transformation represents differentiation or undifferentiated states \cite{24333726}.

Here, we endeavor to collect the data catalog to provide deep annotations of cancer genomes. We performed large-scale integration to construct an in-depth cancer-related companion resource to the general encyclopedia. We complied these resources as the “companion *ENCODE* encyclopedia resource for *C*ancer” (or “EN-CODEC” for short) to interpret cancer-related genomic data, such as mutational and transcriptional profiles.

## **Multi-level data integration improves variant recurrence analysis in cancer**

One of the most powerful ways of identifying key elements in cancer genomes is through recurrence analysis, the objective of which is to discover regions that undergo greater mutation than expected. Hence, we first attempted to construct an accurate background mutation rate (BMR) model in a wide range of cancer types. However, BMR estimation is a challenging problem: the somatic mutation process can be influenced by numerous confounding factors (in the form of both external genomic factors and local sequence context factors), and without appropriate correction, these confounders can result in many false positives or negatives \{cite 23770567}.

We address the issues associated with confounding factors in a cancer-specific manner. Specifically, we separated the whole-genome into bins (1Mb) and calculated mutation counts per bin under each local context category. For each category, we used a negative binomial regression of the mutation counts against 475 features from 300 cell types, including replication timing, chromatin accessibility, Hi-C, and expression profiles for BMR prediction. In contrast to methods that use unmatched data \{cite 23770567}, our approach automatically selects the most relevant features, thereby providing noticeable improvements in BMR estimation (Fig 2A). Notably the combination of many different genomic significantly improves the estimation accuracy in multiple cancer types (Fig 2 B). It is also worth noting that, due to the correlated nature of these genomic features, some cancers without features from apparently matched cell types can still automatically learn from related cell types and achieve a good BMR precision. Hence, our analyses may easily be extended to other cancer types.

 Our next step entailed maximizing the statistical power of burden tests. In traditional analysis, a comprehensive set of annotations is usually thought to be beneficial. However, testing every possible nucleotide in the genome in mutation burden analysis would significantly reduce statistical power (see supplement). First, in terms of an individual test, focusing on shorter core regions with true functional impact would significantly improve computational power. Hence, we trimmed the conventional annotations, such as enhancers, to key regions by looking into shapes of various signal tracks (see supplement). In addition, burden tests would be subject to large penalties from multiple test correction on large numbers of annotations, many of which may include inaccurate or inactive regulatory elements.

We therefore focused on a minimum number of high-confidence annotations in our search for burdened regions. With a particular focus on enhancers, we started by searching for regions supported by multiple evidence. We first proposed a machine learning algorithm (CASPER) to combine shapes of signal tracks from DNase-seq and a battery of 5 to 10 histone modification marks. We then assembled the CASPER predictions with peaks called from STARR-seq experiments, which directly read out candidate enhancers in the genome. Such an integrative approach enables accurate enhancer definitions (see supplement). We also reconciled these enhancers with the main encyclopedia annotations by reporting the overlapping regions and providing new IDs to those which were novel.

A final aspect to increase the power is to link the compact noncoding regulatory elements to the protein-coding genes to form an extended gene region as a whole test unit. As with the exon regions within genes, a natural consequence of this is a set of discrete regions that potentially affect gene expression. Such a unified annotation enables a joint evaluation of the mutational signals from distributed yet biologically relevant genomic regions. As a result of sparse data, traditional methods solely rely on computational correlation, resulting in problematic extended gene definitions. Here we use direct experimental evidence and physical interactions from Hi-C and ChIA-PET experiments, combined with a machine learning algorithm that takes into consideration the wide variety of histone modification marks and gene expression to achieve accurate enhancer-target gene linkages. Finally, the conserved enhancer-target linkages, refined promoters, and RNA-binding sites from eCLIP experiments within genes constitute a so-called extended gene neighborhood (Fig1C). Given their association with well-known oncogenic genes, such a joint test scheme also results in much more interpretable burdened regions.

We demonstrate that our multi-level integration scheme can effectively remove false positives and discover meaningful regions with higher-than-expected mutation counts (Fig 2C). For example, in the context of chronic lymphocytic leukemia (CLL), our analysis identifies well-known highly mutated genes, such as TP53 and ATM, which has been reported from previous coding region analysis. It also discovered genes that are missed by the exclusive analysis of coding regions, such as BCL6. Note that BCL6 has strong prognostic value with respect to patient survival (Fig. 2D), indicating that the extended gene neighborhood may be used as an annotation set for recurrence analysis.

## **Integrating regulatory networks and tumor expression profiles identifies key regulators in cancer**

The ENCODE annotation set also provides detailed regulatory networks instantiated from experimental assays suitable for cancer research. Specifically, for the TF network, we first built distal and proximal TF regulatory networks by linking TFs to genes, either directly by TF-gene interactions through promoters or indirectly via TF-enhancer-gene interactions in each cell type (Fig1 B). We then pruned these networks to include only the strongest edges using another signal shape algorithm \{cite 22039215}. In addition, we merged our cell-type-specific networks to get a generalized network for pan-cancer analysis. Similarly, we also defined an analogous RBP network (in a simpler format). Compared to imputed networks from motif analysis, our ENCODE TF and RBP regulatory networks were built using actual ChIP-seq and eCLIP experiments, which provide much more accurate regulatory interactions between functional elements.

The integrated networks are useful for interpreting the oncogenic changes evident in cancer gene expression data from tumor samples. In particular, using a machine learning method, we integrated 8,202 tumor expression profiles from TCGA to systematically search for the TFs and RBPs that most strongly drive tumor-specific expression patterns. For each patient, our method tests the degree to which a regulators’ regulation is sufficiently correlated with their targets’ 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.

We find that the target genes of MYC are significantly up-regulated in numerous cancers, which is consistent with its well-known role as an oncogenic TF and a transcription activator \cite{22464321}. We further validate MYC’s regulatory effect through knock-down experiments (Fig 3). Consistent with our predictions, the expression of MYC targets is significantly reduced after MYC knockdown (Fig 3A). After confirming the importance of MYC, we use the regulatory network to understand how MYC works with other TFs. We first looked at all triplets involving MYC by requiring that a second TF both interacts and shares a common target with MYC. In all cancer types, we found that MYC’s expression levels are positively correlated with the expression levels of most of its targets, while the second TF shows only limited influence (as determined by partial correlation analysis).

We then investigated the exact structure of such regulatory relationships. The most common triplet interaction mode is a well-understood feed-forward loop (FFL) in which MYC regulates both the common target and the second TF. Most of these FFLs involve well-known MYC partners such as Max and Mxl1. However, we also discovered that many involve another factor called NRF1. Upon further study, we found that that the MYC-NRF1 FFL relationships were mostly coherent (i.e., "amplifying" in nature). We further studied these FFLs by organizing these triplets into logic gates, in which the two TFs act as inputs and the target gene expression represents the output \{cite 25884877}. We show that most of these gates follow either OR or MYC-always-dominant logic. Thus, the ENCODE regulatory network not only helps find key regulators, but also demonstrates how they work in combination with other regulators.

We also analyzed the RBP-network derived from ENCODE eCLIP data, and found key regulators associated with cancer. For example, the ENCODE eCLIP experiment has profiled many SUB1 peaks on the 3’UTR regions of genes, and we find that the predicted targets of the RBP SUB1 were significantly up-regulated in many cancer types (Fig. 3C). As a RBP, SUB1 has not previously been associated with cancer. We thus validated this new association in liver cancer. After knocking down SUB1 in HepG2 cells, its predicted targets are also down-regulated relative to other genes (Fig. 3D). In addition, we found that the decay rate of SUB1 target genes are significantly shorter than non-targets (Fig. 3C). These results indicate that SUB1 may bind to 3’UTR regions to stabilize transcripts. Moreover, we found that the up-regulation of SUB1 target genes is correlated with a poorer patient survival in other cancer types, such as lung cancer (Fig. 4).

We further present the overall regulatory network by systematically arranging the network into a hierarchy. TFs are placed into different levels such that those on the top tend to regulate the expression of other TFs and those at the bottom are in turn more regulated by higher-level TFs \{cite 25880651}. A final hierarchical network structure is shown in Fig 4. We find that the top-layer TFs are not only enriched in cancer associated-genes, but also more significantly drive tumor-to-normal gene differential expressions.

## **Extensive rewiring events in the regulatory network**

For the top-tier cell types with numerous TF ChIP-seq experiments, we constructed cell-type-specific regulatory networks relating to specific cancers and compared them with networks built from their paired normal cell types. We proposed the concept of a “composite normal” by reconciling multiple related normal cell types, as shown in Fig. 5. The pairings -- relating cancerous cell lines to specific tumors and then matching them to normal cell types -- are approximate in nature. However, many of these pairings have been widely used in the literature before (see supplement). Furthermore, with the enrichment of functional characterization assays in ENCODE, they provide us with a novel opportunity to directly understand the regulatory alterations in cancer by looking at specific network changes that are "rewired" in the process of oncogenesis.

In "Tumor-normal pairs", we measured the signed, fractional number of edges changing (i.e., what we call the “rewiring index”), to study how the targets of each common TF changed (i.e., become rewired) over the course of oncogenic transformation. We first ranked TFs according to this index (Fig. 5 A). 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 this latter factor serve as a hallmark of various forms of high-risk leukemia \cite{26202931, 26713593, 26069293}. Interestingly, IKZF1 loss has been found to be associated with the well-known BCR-ABL fusion transcript, which is present in K562, and usually confers poor clinical outcome \cite{26069293}. In contrast, several ubiquitously distributed TFs retain their regulatory linkages (Fig 5A). We observed a similar trend in TFs using a distal, proximal, and combined network (see details in supplement). The trend was consistent across highly rewired TFs such as BHLHE40, JUND, and MYC in lung, liver, and breast cancers (Fig 5).

In addition to the simple direct TF to gene connection-based model, we also measured rewiring using a more complex gene community model. The targets within the TF regulatory network were characterized by heterogeneous network modules (so called “gene communities”), which usually come from multiple biologically relevant genes. Instead of directly measuring the TF’s target changes for each gene, we determined the change in gene communities via a mixed-membership model. This enabled us to evaluate each TF’s overall association changes to these gene communities in tumor and normal cells. Similar rewiring patterns were observed using this model (Fig 5A).

We then tested whether the gain or loss events from the normal-to-tumor transition result in a network that is more similar to or different from those in stem cells like H1-hESC. Interestingly, we find that the gainer group tends to rewire away from the stem cell’s regulatory network, while the loser groups are more likely to rewire toward the stem cell.

We also find that the majority of rewiring events were associated with noticeable gene expression and chromatin status changes, but not necessarily with variant-induced motif loss or gain events (Fig. 5A). This is consistent with previous discoveries that most non-coding risk variants are not well-explained by the current model \cite{25363779}. For example, JUND is a top gainer in CLL. The majority of its gained targets in tumor cell lines demonstrate higher gene expression, stronger active and weaker repressive histone modification mark signals, yet few of its binding sites are mutated. We found a similar trend for the rewiring events associated with JUND in liver cancer. On a related thread, we organized the cell-type-specific networks to cell-type-specific hierarchies, as shown in Figure 3. Specifically, in blood cancer the more mutationally burdened TFs actually sit at the bottom of the hierarchy, whereas the TFs that are more associated with driving cancer gene expression changes tend to be at the top.

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

Summarizing the analysis above, the EN-CODEC resource consists of numerous annotation summarized in Fig. 6 : (1) a BMR model with matching procedure for relevant functional genomics data and a list of regions with higher-than-expected mutations in a diverse selection of different cancers, (2) accurate and refined enhancers and promotors by integrating tens of different functional assays, including STAR-seq, and their comparison with those in ENCODE encyclopedia; (3) enhancer-target-gene linkages and extended gene neighborhoods, based integrating experimentally determined linkages from Hi-C and detailed histone mark and expression correlation, (4) tumor-normal differential expression, chromatin, and regulatory changes, (5) TF regulatory networks, both overall and cell type specific; (6) TFs’ position in the network hierarchy and their rewiring status; (7) an analogous but less annotated network for RBPs.

Collectively, these resources allow us to prioritize key features as being associated with oncogenesis. The workflow in Fig. 6A describes this prioritization scheme in a systematic fashion. We first search for key regulators that are frequently rewired, located in network hubs or at top of the network hierarchy, or significantly drive expression changes in cancer. We then prioritize functional elements that are associated with top regulators, undergo large regulatory changes with respect to gene expression, TF binding, and chromatin status, or are highly mutated in tumors. Finally, on a nucleotide level, we can pinpoint impactful SNVs for small-scale functional characterization by their ability to disrupt or introduce specific binding sites, or which otherwise occur in positions under strong purifying selection.

Using this framework, we subject a number of key regulators, such as MYC and SUB1, to knockdown experiments in order to validate their regulatory effects in particular cancer contexts (Fig 3D). We also identified several candidate enhancers in noncoding regions associated with breast cancer, and validated their ability to influence transcription using luciferase assays in MCF7. We selected key SNVs, based on significantly recurrent mutations in breast cancer cohorts, within these enhancers that are important for controlling gene expression. Of the eight motif-disrupting SNVs that we tested, six exhibit consistent up- or down-regulation relative to the wild type in multiple biological replicates.

One particularly interesting example, illustrating the unique value of ENCODE data integration, is in an intronic region of CDH26 in chromosome 20 (Fig. 6C). Both histone modification and chromatin accessibility (DNase-seq) signals indicate an active regulatory role in MCF7, which was further confirmed as an enhancer by both CASPER and ESCAPE (STARR-seq; Fig. 5D). Hi-C and ChIA-PET data indicated that the region is within a topologically associated domain (TAD) and validated a regulatory linkage to the downstream breast-cancer-associated gene SYCP2 \cite{26334652, 24662924}. We observed massive binding events from TFs in this region in MCF-7. Our motif-based analysis predicts that the particular mutations found in the cohorts can significantly disrupt the binding affinity of several TFs, such as FOSL2, in this region (Fig. 6D). Luciferase assays demonstrate that this mutation introduces a 3.6-fold reduction in expression relative to the wild type, indicating a strong repressive effect on this enhancer’s functionality.

## Conclusion

This study highlights the value of our companion to the main ENCODE encyclopedia as a resource for cancer research. By integrating many different types of assays, we first demonstrate that we can build an accurate BMR model for a wide range of cancer types, and improve the quality and quantity of annotations to look for regions with higher-than-expected mutations. We can also build extensive regulatory networks of various forms from thousands of ChIP-seq and eCLIP experiments to direct study the regulatory changes that accompany transformation to cancer, as well as pinpoint key regulators that are involved in cancer progression. Finally, we leverage the companion resource to provide a prioritization scheme to pinpoint key features for small-scale follow-up studies.

EN-CODEC comprises two resources – generalized annotations, such as the BMR model and merged networks and hierarchies for pan-cancer studies and cancer-specific annotations drives from pairing the top-tier cell lines to particular cancer types. We note that the representative tumor and normal cell types and their pairings used here are rough in nature. However, some pairings have already been widely used in the literature. In addition, the heterogeneous nature of cancer results in complex biology within a given sample: the tumor cells from one given patient exhibit distinct molecular, morphological, and genetic profiles \cite{24048065}. It is difficult to obtain a "perfect" match even from data of real tumor and normal tissues.

This study underscores the value of large-scale data integration, and we note that expanding the scale of these approaches (either by integrating additional data types and/or using tumor mutation and expression data on a larger scale) is straightforward. We also anticipate that an additional step may entail carrying out many of the ENCODE assays on specific tissues and tumor samples. For example, a larger number of genomic features from matched cell types could result in better BMR estimation; more advanced functional characterization assays may generate compact and accurate annotation sets with larger statistical power in burden analyses; and more ChIP-seq/eCLIP experiments would provide more detailed regulatory networks to understand regulatory alterations during cancer progression. In additional, larger cohorts of expression and mutation profiles from many cancer types may be used to discover novel key features in cancer genomes. Although the volume of material needed for such analyses may present challenges, we demonstrate that such a framework is technically feasible and provides further opportunities for the future.