**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 to comprehensively annotate the noncoding regions in the human genome for the first time (ref). The current release broadens the number of cell lines and considerably expands available tissue data. It also greatly increases the depth of of assays by adding new advanced approaches, such as STARR-seq, eCLIP, and Hi-C. ­There is an unparalleled opportunity to integrate over many assays to develop accurate annotations in a cell-type specific manner. This is particularly useful for interpreting genomic variants associated with disease and cell-state changes that underlie many disease processes. Deep integration 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).

[PDM2MG: Don’t feel strongly, but this sentence is a bit funny. Cancer is a disease, so it’s not really cancer itself that’s the application or focus. It’s the investigation/understanding of cancer. Foci (plural of focus) is also a bit funny. It would be funny to say ‘Cancer is a focus to illustrate the key aspects...’, but you could say ‘We focused on cancer...’.]

Here, focusing on data-rich ENCODE cell types, we developed an integrative and network-associated annotation, which may serve as a valuable resource for disease studies. Cancer is one of the best applications to illustrate the key aspects of this ENCODE resource. Unlike many other diseases, cancer is very much a disease of whole-genome alterations and dysregulations \cite{27478938, 26493648}. Cancer cells usually display aberrant behavior of key regulators, extensive epigenetic remodeling, 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 regions on the genome. It may also be used to construct comprehensive high-quality networks, to capture tumor-to-normal alterations from a more global perspective.

[[JZ2MG: we think (3) is not quite relevant, but not strongly. Can we say the full set of ENCDOE data is our resource?]]

Therefore, we present an *ENCODE* companion resource for *C*ancer genomics (ENCODEC). This resource consists of (1) compact noncoding annotations and extended gene definitions that can potentially increase the statistical power to interpret genome variations (both germline and somatic) and gene expressions; (2) comprehensive experiment-based networks that allow us to depict global alterations in network rewiring, hierarchies, and dysregulations; and (3) the full catalogue of ENCODE cell types that allows us to place oncogenic changes relative to normal and stem cells within a global context . As an example, we coupled our resource to cancer expression and mutation data from large cohorts to prioritize key regulators, elements, and single nucleotide variants (SNV). We show the accuracy and utility of our resource using several targeted experimental validations.

## Application of the ENCODE resource highlights its breadth and depth

Figure 1 illustrates two key dimensions of the overall ENCODE data set: its breadth across cell types and depth across assays. Data aggregation and integration over these two dimensions can benefit the interpretation of genome variations in cancer. Here we show two examples of how the integration and aggregation of raw signal level ENCODE data can be used to interpret structural variants (SV) and single nucleotide variants (SNVs).

In Figure 1, we first show how one can do cross-assay comparisons on the assay-rich cell lines to illuminate potential SV impacts. In particular, we surveyed regions around somatic break points in the K562 cell line, which were called using an integrative approach from various assays in ENCODE (see suppl. sect. xxx). We found that the activating histone H4K20me1 occurs preferentially around these breakpoints, which was not observed using GM12878 histone data at these exact same locations. We further examined the GM12878 H4K20me1 levels proximal to benign germline breakpoints from the 1000 Genomes Project and found no obvious trend. One potential implication is that the somatic SVs in tumor cells may be associated with creating active regions of chromatin, which may be important in oncogenesis.

In addition to such comparisons, ENCODE data may also be integrated to provide insight related to SNVs. For instance, cancer variants may be best understood in the context of recurrence analysis by comparison against a background mutation rate model (BMR). However, it is challenging to develop such model with high accuracy because the somatic mutation process can be influenced by numerous confounders, which can result in false conclusions if not appropriately corrected. Researchers have suggested various approaches to integrate genomic features for accurate BMR calibrations (ref). ENCODE has dramatically increased available features, especially histone modifications and replication timing, from less than 200 to over 2,000. This is particularly valuable because cancer cells are usually highly heterogeneous and thus are not necessarily matched by a single cell type in all assays. We show in Figure 1 that simple integration of these data sets by a negative binomial regression provides progressively more accurate BMR estimates, and the estimation performance still improves even after ten or more features are added to the model.

## Linking compact individual annotations to form an extended gene

From the wealth of the ENCODE experiments in data-rich cell types, we constructed a deep, integrated annotation with two key characteristics – 1) the individual noncoding elements are compactly defined to more precisely locate true functional sites, and 2) these discontinuous regulatory regions are precisely linked to genes to form the extended gene definitions. Different from conventional gene annotations, which are uniformly defined across cell types, our extended gene definitions are highly dynamic and may change considerably across cell types. In contrast with conventional efforts to provide annotations as comprehensive as possible, our annotation may improve the statistical power of many analyses in cancer.

First, it is natural to increase power by reducing the level of noise in an annotation in two mechanisms: decreasing multiple test burden by removing false positive elements, or defining core regions by removing nonfunctional nucleotides per element. According to the former approach, we performed strict quality control of our annotation by deep integration (see supplements). For example, to define distal regulatory elements, we integrated the shape of up to ten histone modification ChIP-seq experiments per cell type using a support vector machine approach (see suppl. sect. xxx). Then we intersected these elements with positive scores from STARR-seq experiments. This resulted in a shorter list of enhancers with high accuracy (see suppl. sect. xxx). Related to the latter approach to increasing power, we restricted individual annotated elements down to a core set of annotations enriched for functional sites by incorporating novel advanced assays such as eCLIP. As a result, our annotations are shorter in length but higher in degree of conservation (see supplement).

A second step to increase power entails linking the above compact annotations to define an extended gene neighborhood (Fig. 2A). For example, to define accurate enhancer-target linkages, we first correlated activity between the chromatin marks and gene expression using an advanced machine learning approach. Then we intersected the most correlated regions with physically-based but lower resolution connections from Hi-C experiments to generate a high-confidence subset (see suppl. sect. xxx). Such a gene-centric approach not only allows us to improve upon existing knowledge of genetic regions, but also enables a joint evaluation of distributed yet biologically connected genomic regions.

To illustrate the value of this extended gene annotation, we first show an example of cancer GWAS SNV enrichment analysis. The enrichment in protein-coding genes significantly increases as we add more relevant proximal and distal annotations for both breast cancer and leukemia (Fig. 2B). This trend is more pronounced when the newly added 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 Consensus (CGC) in breast cancer based on the extended gene annotation, which was not possible using the conventional non-coding regions annotations.

We also show a second example of how the extended gene may be applied to cancer somatic variant recurrence analysis. It is possible to jointly test the somatic mutation burden on an extended gene with greater statistical power compared with separate tests on each element composing the extended gene. We performed such an analysis in well-known cancer cohorts (Fig. 2E). 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 \cite{xxx}. 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.

Our extended gene definitions can also be used for differential expression analysis based on mutation status . For example, we combined the mutation and expression profiles from large cohorts, such as TCGA, and found that mutation status in our extended gene definition can better 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 mutations within its extended gene annotation exhibits greater significance relative to gene expression, compared to any single annotation category (p=xxx, one sided Wilcoxon test).

One can get a physical sense of the importance of the extended gene environment by looking at a situation where a genomic variant rearranges the extended gene structure 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). (The enhancer not being connected ERBB4 in normal breast tissue.) We found that 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. 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 linking genomic elements. Specifically, we built both distal and proximal networks 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 \cite{TIP}.sect. xxx). In addition, we merged each cell-type specific networks to form a generalized 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 TF and RBP 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 (see 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). We found that RBPs affect gene expression to a lesser extent than TFs (P=xxx, one sided Wilcoxon Test), which may due to the fact that most RBPs carry out post-transcriptional level regulation, such splicing, RNA localization, and transportation, rather than direct transcription initiation. These networks also enable investigation of the connections between TFs and RBPs. Interestingly, we found that there are fewer bottom-level direct TF-RBP interactions, as compared to top and middle level ones (P=xxx, one sided Wilcoxon Test). The well-known oncogene MYC is one of the master TFs (top-level of the hierarchy), which not only directly regulates the expression of other TFs, but also targets many RBPs.

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 in different cancers (see 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. 3B.

As expected, we found that the target genes of MYC are significantly up-regulated in numerous cancer types -- in fact, the most up-regulated of any TF -- 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. 3C). Consistent with our predictions, the expression of MYC targets is significantly reduced after MYC knockdown in MCF-7 (Fig. 3C). We analyzed the RBP network in a manner 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 investigate its role. Knocking down SUB1 in HepG2 cells significantly down-regulated its targets, and the decay rate of SUB1 targets is lower than those of non-targets (Fig. 3D). Moreover, we found that up-regulation of SUB1 targets may lead to decreased patient survival in some cancer types.

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. 3E. 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. Many 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 (see suppl. sect. xxx). We further studied the FFLs by organizing them into logic gates, in which two TFs act as inputs and the target gene expression represents the output. 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. In fact, SUB1 is a direct target of MYC in many cell types (see suppl. sect. xxx) and they also 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 3’ UTRs to stabilize oncogenes at the level of RNA transcripts. Such synergistic 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. Comparison of these networks between matched tumor and normal cell types enable measurement of the change in connections (ie 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. sect. xxx). Although the pairings are only approximate, many of them have been widely used in prior studies (see suppl. sect. xxx). 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.

We first organized cell-type specific networks into hierarchies, as shown in Fig. 4A. These hierarchies may expose meaningful regulatory relationships. For example, in blood cancer, we found that the strongest edge gainers and losers in rewiring events often sit at the top level of the network hierarchy. In addition, we found 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 particular, we measured the fractional number of edges changes for "tumor-normal pairs”, to study how TF targets change in the oncogenic transformation. We call this the "rewiring index" and In Fig. 4A and ranked TFs according to it. 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. 4A). 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. 4C).

In addition to direct TF-to-gene connections, we also measured rewiring using a gene-community model. Here, the targets within the regulatory network were characterized in terms of the modules of multiple genes to which they belonged (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 the gene communities it regulates (via a mixed-membership model) (see suppl. sect. xxx). Similar patterns to direct rewiring were observed using this model (Fig. 4C).

We found that the majority of rewiring events were associated with noticeable gene-expression and chromatin-status changes, but not necessarily with direct variant-induced motif loss or gain events (Fig. 4A). For example, JUND is a top gainer in K562. Most 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, either by SNVs or SVs. This is consistent with previous work that indicates most non-coding risk variants are not well-explained by a mutational model. With a few notable exceptions, 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 (see suppl. sect. xxx).

We found that the strongest edge gainers and losers in rewiring events often sit at the top level of the network hierarchy. In addition, we found 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.

## 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 stem cells (ref). One of the strengths of ENCODE is its many stem cell lines, including H1, 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. First, we projected all the RNA-seq data into a low dimension space by Reference Component Analysis (RCA, \cite{nat rca paper}). We found that various types of stem cells, including data-rich H1 cells, form a tight cluster (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, along a single “stem-like principal component.” This is true for a variety of different cancers. This observation is consistent with previous efforts using expression and methylation analysis (ref). Notably, we observed a consistent (or even stronger) pattern from proximal and distal chromatin data, which can be viewed as the underlying cause of the observed gene expression changes.

It is 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. We can use the cell space diagram, 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, along the stem-like component. One can see this difference more precisely and test it statistically if one restricts just to the single transition from K562 (Fig 5).

## Step-wise variant prioritization with targeted validations

Collectively, as schematized in Fig 6, ENCODEC enables a step-wise prioritization scheme that allows us to pinpoint key regulator genes, noncoding elements, and SNVs associated with oncogenesis. Specifically, we first highlighted regulators that are either frequently rewired, located in network hubs, 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 burdened by mutations, undergo large chromatin changes, or change in extended gene linkages. Finally, on a nucleotide level, by estimating their ability to disrupt or introduce specific binding sites, we can 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 (Fig. 3B and 3D) and we measured the effect of SVs on enhancer linkages via 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 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

In this paper, we have described an ENCODE companion resource that provides a deep integrative annotation. This annotation includes compact individual noncoding annotations, extended gene structures, and organizes these together into various networks at a large scale. We show how our resource can help describe oncogenic transformations in cell space. The rewiring of our constructed regulatory networks can be used to help explain what happened in select oncogenic transformations. Altogether, we can use the annotation to prioritize key regulators and variants, especially in cancer.

There remain several caveats associated with our resource. First, our resource associates cancer types with ENCODE cell lines and then secondarily pairs a specific cancer with a composite normal. Both types of pairings are, by nature, approximate. Tumor cells from a given patient show distinct molecular, morphological, and genetic profiles. Linking cancer to one specific cell-type may not fully capture this 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 allow cell-type or tissue-type comparisons at a higher resolution. Nevertheless, we feel that our networks currently provide the best available view of the regulatory changes in oncogenesis. Second, cancer driver discovery is usually a multiple-step process that requires coordinated, large-scale community effort, such as in the TCGA and PCAWG projects. In addition to the careful calibration required for statistical identification of drivers (recurrence), extensive follow up validations are required as obtained by experiment. We hope that the wealth of ENCODE epigenetics data, accurate annotations, precise linkage to genes, and comprehensive networks can provide researchers with useful perspective on cancer biology. [all2MG: suggest to revise as a limitation but still useful. We discussed without a conclusion]

PDM2MG: I tried to rework this a bit, but stopped (do not want to change too much at this stage). However, the relationship of our resource to driver detection is not made clear. What impression do we want to give? We are implying something like ‘Our ENCODE resource is not sufficient for driver detection on its own, but can be applied to driver detection in the context of a larger community effort.’ Do we want to say this explicitly?

Somewhat counter-intuitively, in the context of disease studies, we suggest skepticism regarding a comprehensive annotation that attempts to assign functional impact to every base in the genome. Rather, often the most useful annotation is as accurate and compact as possible, both in terms of extent of individual annotation blocks and in number of elements, thereby increasing power. We demonstrated how such an annotation can be beneficial to the interpretation of expression and mutation data from cancer cohorts. Furthermore, an optimum annotation often has a network-connected aspect rather than just comprising a simple linear genomic block. This network aspect has the important feature of connecting non-coding elements to genes. In both cancer and rare disease studies, researchers have found that high impact variants tend to be tightly associated with genes (\cite {PCAWG driver paper and CMG}). Hence, connecting non-coding elements to genes may be valuable in terms of variant interpretation. Through large scale data integration, we have tried to prototype analysis in such a compact, network annotation.

All together, we successfully formed compact annotations and regulatory networks for model cell types with advanced functional assays like eCLIP and STARR-seq. The example applications of our resource underline the importance of deep data integration over many novel assays to understand cancer genomes. We want to further point out that the value of our resource can go beyond noncoding variant interpretation. Oncogenesis often involves genome-wide gene dysregulations, epigenetic remodeling, network perturbations, and cell state transitions. We have shown how our resource can be used to characterize such events. The successful validation of our prioritized regulator genes, noncoding elements, and SNVs demonstrates the value of our resource. [all2MG: suggest a better closing sentence]