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

Comment [j1]: Robert suggested to add ENCODEC in title to boost citation

Formatted: Centered, Line spacing: multiple 1.15 li

Formatted: Font:(Asian) +Theme Body Asian (DengXian), 16 pt, Bold, Pattern: Clear

An integrative ENCODE resource for interpreting non-coding mutations and gene regulation in cancer

[JZ2MG: my favorite title so far]
ENCODEC: An integrative ENCODE resource for interpreting non-coding mutations and gene regulation in cancer

Jing Zhang*, Donghoon Lee*, Vineet Dhiman*, Peng Jiang*, William Meyerson, Matthew Ung, Shaoke Lou, Patrick Mcgillivray, Declan Clarke, Lucas Lochovsky, Lijia Ma, Grace Yu, Arif Harmanci, Mengting Gu, Koon-kiu Yan, Anurag Sethi, Qin Cao, Daifeng Wang, Gamze Gursoy, Jason Liu, Xiaotong Li, Michael Rutenberg Schoenberg, Joel Rozowsky, Lilly Reich, Chongzhi Zang, Juan Carlos Rivera-Mulia, Jie Xu, Jayanth Krishnan, Yanlin Feng, Jessica Adrian, James R Broach, Michael Bolt, Vishnu Dileep, Tingting Liu, Shenglin Mei, Takayo Sasaki, Su Wang, Yanli Wang, Hongbo Yang, Feng Yue, David M. Gilbert, Michael Snyder, Kevin Yip, Chao Cheng, Robert Klein, Shirley Liu, Kevin White, Mark Gerstein

Abstract

Most somatic mutations in cancer are non-coding while the characterized drivers are predominantly located in coding regions, creating a conundrum as to whether the non-coding regions are important for oncogenesis. Here we endeavor to create a companion resource to the main ENCODE encyclopedia to address this issue. In particular, we integrate diverse ENCODE data to precisely calibrate background mutation rates and we utilize advanced functional-genomic assays, especially STARR-seq and Hi-C, to develop compact annotations and accurate extended gene models (linking enhancers to coding regions), achieving better statistical power for burden analysis. We also construct detailed regulatory networks to interpret tumor gene expression and mutation profiles, pinpointing effects of key regulators such as the transcription factor MYC and the RNA-binding-protein SUB1 and then validating them. We build cell-type specific networks to directly measure regulatory "rewiring" during oncogenesis, classifying changes as either moving toward or away from a stem-like state. Finally, we integrate the overall ENCODE resource, comprising networks and a compact annotation, to prioritize non-coding elements and mutations and then we validate a subset of them through targeted experiments.

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 genomics 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 assays to interpret cancer genomes.

The rich functional assays and annotation resources developed by the ENCODE Consortium allow us to characterize these non-coding regions in depth \cite{22955616}. Given that around eighty percent of ENCODE cell lines are associated with cancer (see supplement), ENCODE data are particularly suited for interpreting cancer gene regulation[[JZ2MG: this is Shirley's suggestion, but then where is the variant?]]. In the initial release of the ENCODE annotation sets, 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 to conduct RNA-seq, ChIP-seq, and DNase-seq profiles. As such, the main ENCODE encyclopedia aims to utilize these to provide a general and unified annotation resource applicable across many cell types. Second, 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-rich top-tier cell line is the human embryonic stem cell line H1-hESC. For decades, the prevailing paradigm has held that at least a subpopulation of tumor cells have the ability to self-renew, differentiate, and regenerate, in a manner that is similar to stem cells \cite{24333726}. Hence, H1-hESC can serve as a valuable comparison when investigating the degree to which the oncogenic transformation represents stem cell like activities\cite{24333726}. [[JZ2MG: suggest not to mention the differentiated or undifferentiated direction since we do not have the data]]

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 compiled 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 mutation recurrence analysis, the objective of which is to discover regions that undergo more 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 these confounders can result in wrong conclusions if not appropriately

Deleted: genome-wide

Deleted: advanced

Deleted: allows

Deleted: at great

Deleted: cancerous tissues

Deleted:)

Deleted: cancer research.

Formatted: Highlight

Deleted: with

Deleted: assays

Deleted: ,

Deleted: Secondly

Deleted: (

Comment [SL2]: People who are not familiar with ENCODE may not understand "top tier cell lines".

Deleted:)

Formatted: Highlight

Deleted:) .

Deleted: H1-hESC, a

Deleted: a tumor's

Deleted: current thinking for normal

Formatted: Highlight

Deleted: their oncogenic

Deleted: differentiation or undifferentiated states

Deleted: greater mutation

Formatted: Highlight

Deleted: without appropriate correction,

Deleted: many false positives or negatives

corrected \{cite 23770567}. [[JZ2MG: basically no difference of her version and our version, but hers might be easier to experimentalists to understand?]]

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 across 229 cell types, including replication timing, chromatin accessibility, Hi-C, and expression profiles for BMR prediction. In contrast to methods that use data from unmatched cell types \{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 features significantly improves the estimation accuracy in multiple cancer types (Fig 2 B). In addition, due to the correlated nature of these genomic features across cell types, imperfect matching of some cancer with an ENCODE cell line can still improve BMR precision. Hence, our analyses may easily be extended to other cancer types.

A second step to utilize the ENCODE data in the mutation recurrence analysis in cancer is to maximize 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 recurrence analysis will significantly reduce statistical power (see supplements). First, in terms of an individual test, focusing on shorter core regions with true functional impact would significantly improve detectability. Hence, we trimmed the conventional annotations, such as enhancers, to the key regions by looking into shapes of various signal tracks (see supplements). Second, burden tests would be subject to large penalty from multiple testing correction on a large number of annotations, which might 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 types of 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 by our computational method ESCAPE 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 the novel regions.

A final aspect to increase the statistical power is to link the compact noncoding regulatory elements to 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. 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), which usually 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 identified well-known highly mutated genes, such as TP53 and ATM, which have been reported from previous coding region analysis. It also discovered genes that were 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.

Deleted: over

Deleted: data

Deleted: It is also worth noting that

Deleted: burden

Deleted: Besides

Deleted: from STARR-seq experiments

Deleted:

Deleted: ESCAPE

Deleted: those which were

Deleted: detection

Deleted: the

Formatted: Highlight

Deleted: .). Given their association with well-known oncogenic genes, such a joint test scheme also

Deleted: identifies

Deleted: has

Deleted: are

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 transcription factor (TF) network, we first built distal and proximal TF regulatory networks by linking TFs to genes, either directly by TF-gene promoter interactions or indirectly via TF-enhancer-gene interactions in each cell type. We then pruned these networks to include only the strongest edges using another signal shape algorithm \{cite 22039215}. In addition, we merged the cell-type-specific networks to get a generalized pan-cancer network. Similarly, we also defined an analogous RNA binding protein (RBP) network (in a simpler format). Compared to imputed networks from gene expression or 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 regulators' activity is correlated with their targets' tumor-normal expression changes. We then calculated the percentage of patients with these relationships in each cancer type, and presented 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 cancers, which is consistent with its well-known role as an oncogenic TF and transcriptional activator \cite{22464321}. We further validated MYC's regulatory effect through knockdown experiments (Fig 3). Consistent with our predictions, the expression of MYC targets is significantly reduced after MYC knockdown (Fig 3A). We then used 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 further investigated the exact structure of regulatory relationships of MYC with other TFs. The most common triplet interaction mode is a well-understood feed-forward loop (FFL) whereby, in this case, MYC regulates both another TF and a common target of both MYC and that TF (Figure 3 C). Since MYC amplification is a major determinant of many cancers, understanding which TFs appear to further amplify MYC effects through FFLs may yield insights for efforts aimed at MYC inhibition /cite{PMC4200208}. Most of these FFLs we observed involve well-known MYC partners such as MAX and MXL1. However, we also discovered that many involve another factor NRF1. Upon further examination, 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 gate. 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 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. 3C). As an RBP, SUB1 has not previously been associated with cancer, so we sought to validate its role. Knocking down of SUB1 in HepG2 cells significantly down-regulated its target genes relative to other genes (Fig. 3D), and the decay rate of SUB1 target genes is significantly lower than non-

Deleted: through promoters
Deleted: (Fig1 B)
Deleted: our
Deleted: network for
Deleted: analysis
Deleted: regulation
Deleted: present
Deleted: We found
Deleted: find
Deleted: a
Deleted: After confirming the importance of MYC, w
Deleted: use
Formatted: Highlight
Deleted: then
Deleted: such
Deleted: .
Formatted: Not Highlight
Deleted: the
Deleted: and the second
Deleted: in which MYC regulates both the common target and the second TF
Deleted: called
Deleted: study
Deleted: (
Deleted:).
Deleted: experiment
Deleted: profiled many SUB1
Deleted: we find that
Deleted: the RBP
Deleted: a
Deleted: . We thus validated
Deleted: in liver cancer. After knocking
Deleted: , its predicted targets are also
Deleted:). In addition, we found that
Deleted: are
Deleted: shorter

targets (see supplements). Moreover, we found that the up-regulation of SUB1 target genes is correlated with a poorer patient survival in some cancer types, such as lung cancer (Fig. 3D). These results suggest that SUB1 may have oncogenic roles by binding to the 3'UTR regions to stabilize its target transcripts.

We further presented 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 found that the top-layer TFs not only enriched in cancer associated-genes, but also more significantly drive differential gene expressions in tumors.[[JZ2MG: Shirley's comment is reasonable, For such a big figure, the textual explanation is too brief and vague]]

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 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 IKZF1 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 changes to these gene communities in tumor and normal cells. Similar rewiring patterns were observed using this model (Fig 5A).

We next 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, 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.

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

Deleted: Fig. 3C). These results indicate that SUB1 may bind to 3'UTR regions to stabilize transcripts.

Deleted: other

Deleted: 4).

Deleted: present

Comment [SL3]: For such a big figure, the textual explanation is too brief and vague.

Deleted: find

Deleted: are

Deleted: are

Deleted: tumor-to-normal gene

Deleted: .

Deleted: relating to specific cancers

Deleted: this latter factor

Deleted: ,

Deleted: then

Deleted: we find that

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 annotations 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 mutation burdens in a diverse selection of different cancer types, (2) accurate and refined enhancers and promotors by integrating tens of different functional assays, including STARR-seq, and their comparison with those in ENCODE; (3) enhancer-target-gene linkages and extended gene neighborhoods, obtained by 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 merged 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 at the 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 subjected 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 MCF-7. We selected key SNVs, based on mutation recurrence in breast cancer cohorts, within these enhancers that are important for controlling gene expression. 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 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 MCF-7, 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 strong binding of many TFs in this region in MCF-7. Our motif-based analysis predicts that the particular mutation from a breast cancer patient can significantly disrupt the binding affinity of several TFs, such as FOSL2, in this region (Fig. 6D). 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 this enhancer's functionality.

Conclusion

This study highlights the value of our EN-CODEC 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

Deleted: MCF7

Deleted: exhibit

Deleted: MCF7

Comment [SL4]: The grammar in this sentence is wrong: Histone mark and chromatin accessibility can't be confirmed as an enhancer.

Deleted: massive

Deleted: events from

Deleted: s

Deleted: found in the cohorts

Deleted: demonstrate

Comment [SL5]: Should we use past tense in this paragraph, since all of the work has been done?

of annotations to look for regions with higher-than-expected mutation burdens. We also build extensive regulatory networks of various forms from thousands of ChIP-seq and eCLIP experiments to directly study the regulatory changes that accompany transformation to cancer, as well as pinpoint key regulators. Finally, we leverage the companion resource to provide a prioritization scheme to pinpoint key features for small-scale experimental follow-up studies.

EN-CODEC comprises two resources: 1) generalized annotations, such as the BMR model and merged networks and hierarchies for pan-cancer studies; and 2) cancer-specific annotations from pairing the top-tier cell lines to particular cancer types. We did realize 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 other literatures. In addition, cancer is a heterogeneous disease that even the tumor cells from one patient usually show 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 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 addition, larger cohorts of expression and mutation profiles from many cancer types may be used to discover novel key features in cancer genomes. We demonstrate that such a framework is technically feasible and provides further opportunities for the future.

Deleted: that are involved in cancer progression

Deleted: are

Deleted: such

Deleted: s

Comment [SL6]: This doesn't seem to be an example of the previous sentence.