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

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, 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, Chongzhi Zang, Feng Yue, David M. Gilbert, Michael Snyder, Kevin Yip, Chao Cheng, Robert Klein, X. 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 non-coding regions are important for oncogenesis. Here we endeavor to address this issue through creating a companion resource to the main ENCODE encyclopedia. In particular, we integrate diverse ENCODE data to precisely calibrate background mutation rates. 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), allowing us to achieve better statistical power for mutational 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 the degree of "rewiring" during oncogenesis, classifying changes as either moving toward or away from a stem-like state. Finally, we use our overall resource -- comprising the compact annotations, networks, and burdened regions -- to prioritize non-coding elements and their mutations, and 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. [[MG2JZ thousands?]] [[JZ2MG: yes, if we also count ChIP-seq, eCLIP, ChIP-seq itself is around 900 already]]

Formatted: Highlight

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 many ENCODE cell types are associated with cancer (see Figure 1 and supplementary file), ENCODE data are particularly suited for interpreting somatic variants and gene regulation in cancer. The initial release of the ENCODE annotation was mainly focused on a limited number of cell types using RNA-seq, DNase-seq and ChIP-seq assays

\cite{22955616}. The new release of ENCODE took two new directions. First, it considerably broadened the number of cell types using the original assays. As such, the main ENCODE encyclopedia aims to utilize these to provide a general 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 lines are associated with various types of cancer (Figure 1), including those of the blood (K562), breast (MCF-7), liver (HepG2), and lung (A549). Also, another data-rich, top-tier cell line is for the human embryonic stem cell (H1-hESC). For decades, a 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}.

Here, we integrate ENCODE data to provide deep annotations of cancer genomes, focused on interpreting cancer-related data, such as mutational and transcriptional profiles. In particular, we construct a companion resource to the general encyclopedia, which we call “EN-CODEC” (“companion *ENCODE* encyclopedia resource for *C*ancer”).

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 having more mutations than expected. Hence, we wish to construct an accurate background mutation rate (BMR) model in a wide range of cancer types. However, BMR estimation is challenging: the somatic mutation process can be influenced by numerous confounders (in the form of both external genomic factors and local sequence context factors), and these can result in wrong conclusions if not appropriately corrected \{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 for a BMR prediction, we used a negative binomial regression of the mutation counts against 475 genomic features across 229 cell types, including replication timing, chromatin accessibility, Hi-C, and expression profiles. 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). Consistent with previous results on mutation rates, in breast cancer, we observed an elevated rate in regions with the repressive modification H3K9me3 and a reduced rate in regions with the activating, enhancer-associated mark H3K27ac \cite{25732611, 22820252, 25693567 }. Also, due to the correlated nature of genomic features across cell types, even approximate matching of a specific cancer against an ENCODE cell line can still improve its BMR estimation. Hence, our analyses may easily be extended to other cancer types.

A second aspect to best using the ENCODE data in cancer mutation analysis is maximizing the statistical power of burden tests. In traditional genomic analyses, a comprehensive set of annotations, usually covering as many base pairs as possible, is considered to be beneficial. However, testing every possible nucleotide in the genome in mutation recurrence analysis noticeably reduces statistical power for several reasons (see supplements). [[JZ2MG: logically this sentence here is bad. In last sentence, we talked about why comprehensive annotation is bad, but in the next sentence we changed to why compact is good. But on the other hand, we want to introduce the concept of core and compact]] First, for a single burden test on an individual genomic element (e.g., an enhancer), focusing on a smaller, "core" region, enriched for true functional impact, would significantly improve detectability. Hence, we trimmed the conventional annotations to key "functional territories" by using the TF-binding sites (TFBS) and the shapes of various

Deleted:

Deleted: only

Deleted: Hence, our analyses

Deleted: greatly

Formatted: Highlight

Deleted: motifs

genomic signals (see supplement). [[MG2JZ: note TFBS]][[JZ2MG: not exactly sure what do you mean]] Second, repeated burden tests on a large number of elements, would be subject to large penalty from multiple-testing correction. We, therefore, tried to develop a minimum number of high-confidence annotations in our search for burdened regions by removing low-confidence ones as much as possible. With a particular focus on enhancers, we started by searching for regions supported by multiple lines of evidence in the data-rich top-tier cell types. We developed a machine-learning algorithm to combine shapes of signal tracks from DNase-seq and a battery of up to 10 histone modification marks. We then intersected these predictions with those of putative enhancers called from STARR-seq experiments (using a second algorithm). These experiments provide a direct, albeit noisy, readout of enhancer activity in particular cell types. Such an integrative approach enables us to define a minimal list of enhancers with as few false-positives as possible (see supplement). We also reconciled and cross-referenced our "compact annotation" with the main encyclopedia annotations (see supplement).

A final aspect of our compact annotation to increase statistical power is linking the noncoding regulatory elements to protein-coding exons to form an extended gene region as a single test unit. Such a unified annotation enables joint evaluation of the mutational signals from distributed yet biologically connected genomic regions. Traditional methods for linking rely solely on the correlation of individual signals (e.g., between the activity of one histone mark at enhancer and gene expression of neighboring genes) and potentially result in inaccurate 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.

Putting together our compact annotation, BMR estimation, and accurate extended gene definitions allows us to get maximal power in detecting genomic regions, coding and non-coding, burdened by mutations (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 that have been reported from previous analyses. It also discovered genes missed by the exclusive analysis of coding regions, such as BCL6. This gene has strong prognostic value for patient survival (Fig. 2D).

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

The ENCODE annotation provides detailed regulatory networks directly based on experimental assays suitable for cancer research. Specifically, for the transcription factor (TF) network, we built distal and proximal TF regulatory networks by linking TFs to genes, either directly by TF-promoter binding or indirectly via TF-enhancer-gene interactions in each cell type. We then pruned these networks to include only the strongest edges using a signal shape algorithm \{cite 22039215}. In addition, we merged all our cell-type-specific networks to form a generalized pan-cancer network. Similarly, we also defined an RNA-binding protein (RBP) network. Compared to imputed networks from gene expression or motif analysis, our ENCODE TF and RBP networks were built using actual ChIP-seq and eCLIP experiments, which provide much more accurate regulatory linkages between functional elements.

The networks are useful for interpreting the 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 most strongly driving tumor-specific expression. For each patient, our method tests the degree to which regulators' activity correlates with their targets' tumor-to-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.

Formatted: Highlight

Deleted: ,

Deleted: called from

Deleted: is that helps

Deleted: re

Deleted: gene

Deleted: an RNA

Deleted:)

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 \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 3B). We then used the regulatory network to understand how MYC works with other TFs. We first looked at all triplets involving MYC, requiring that a second TF both interacts with 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 these regulatory triplets. The most common one 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 (Figure 3 C). Since MYC amplification is a major determinant of many cancers, understanding which TFs appear to further amplify its effects 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 many involve 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 them 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 an OR or MYC-always-dominant logic gate. Thus, the ENCODE regulatory network not only identifies key cancer regulators, but also demonstrates how they work in combination with other regulators.

We also analyzed the RBP-network derived from 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 targets (Fig. 3D), and the decay rate of SUB1 targets is significantly lower than non-targets (see supplement). Moreover, we found that the up-regulation of SUB1 targets is correlated with a poorer patient survival in some cancer types, such as lung cancer (Fig. 3D). These results suggest that SUB1 may have an oncogenic role.

We further analyzed the overall regulatory network by systematically arranging it into a hierarchy (Fig 4). Here, TFs are placed on different levels such that those in the middle tend to regulate TFs at the bottom and, in turn, are more regulated by top-level \{cite 25880651}. In the hierarchy, we found that the top-layer TFs not only enriched in cancer associated-genes but also more significantly drive differential gene expressions in tumors.

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 (see supplement). 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, they leverage the extensive functional characterization assays in ENCODE to provide us with a novel opportunity to directly understand the regulatory alterations in cancer.

In particular, in "tumor-normal pairs", we measured the signed, fractional number of edges changing -- what we call the "rewiring index" -- to study how the targets of each common TF changed over the course of oncogenic transformation. In Fig. 5A, we ranked TFs according to this index. In leukemia, well-known

Deleted: 3A

Deleted: level \

Deleted: . ... [1]

Deleted: , as shown in Fig. 5

Formatted: No Spacing,TextBody

Deleted: "

Deleted: ,

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}. [[MG2JZ: do we want to mention so much about IKZF1]] We observed a similar trend in TFs using distal, proximal, and combined network (see details in supplement). The trend was consistent across 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. The targets within the TF regulatory network were characterized by heterogeneous network modules (so called "gene communities"), which come from multiple biologically relevant genes. Instead of directly measuring the changes in a TF's targets between tumor and normal, we determined the changes in its gene communities via a mixed-membership model. [[MG2JZ: "its communities," ok?]] Similar patterns to the direct rewiring 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 or less similar to that in stem cells like H1-hESC. Interestingly, the gainer TF group tends to rewire away from the stem cell's regulatory network, while the loser group is 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 mutation-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}. [[MG2JZ: we need to reword the last sentence.]] For example, JUND is a top gainer in K562. The majority of its gained targets in the tumor 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 and, largely, for other factors in a variety of cancers (see supplement), with some notable exceptions [[see supplements]]. Related to this, we organized the cell-type-specific networks into cell-type-specific hierarchies, as shown in Figure 4. 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.

Step-wise prioritization schemes pinpoint deleterious SNVs in cancer

Summarizing the analysis above, the EN-CODEC resource consists of 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 mutational burden in a diverse selection of different cancers; (2) highly accurate, minimal and compactly defined enhancers and promotors found from integrating many functional assays, including STARR-seq; (3) enhancer-target-gene linkages and extended gene neighborhoods, obtained by integrating linkages from Hi-C and multi-histone mark and expression correlation; (4) tumor-normal differential expression, chromatin, and regulatory changes; (5) TF regulatory networks, both merged and cell-type specific, based on both distal and proximal regulation; (6) for each TF, its position in the network hierarchy and rewiring status; and (7) an analogous but less-developed network for RBPs.

Collectively, these resources allow us to prioritize key features as being associated with oncogenesis. Our prioritization scheme is shown in as a workflow in Fig. 6A. We first search for key regulators that are frequently rewired, located at the network hubs, sit at the top of the network hierarchy, or significantly drive expression changes in cancer. We then prioritize functional elements that are associated with these regulators, are highly mutated in tumors, or undergo large changes in gene expression, TF binding, or chromatin status. Finally, on a nucleotide level, we pinpoint impactful SNVs for small-scale functional

Deleted: . 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}

Formatted: Highlight

Deleted:

Deleted: its

Formatted: Highlight

Formatted: Highlight

Deleted: CML

Deleted: [[MG2JZ: CLL or CML?]]

Deleted: [[MG2JZ: see supplement]].

... [2]

Deleted: 3

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 some 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). The shape of both histone modification and chromatin accessibility (DNase-seq) signals indicate its active regulatory role as an enhancer in MCF-7. This was further confirmed by STARR-seq (Fig. 5D). Hi-C and ChIA-PET data indicated that the region is within a topologically associated domain 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 significantly disrupts 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

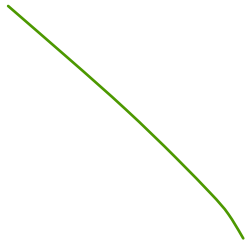
This study highlights the value of our EN-CODEC companion to the main ENCODE encyclopedia as a resource for cancer research. By integrating many different types of assays, we first demonstrate that we can construct an accurate BMR model for a wide range of cancers and customize non-coding annotations to maximize their power in mutational burdening calculations. We also built extensive regulatory networks from thousands of ChIP-seq and eCLIP experiments to directly study the regulatory changes associated with cancer, as well as highlight key regulators. Finally, we leveraged the resource to provide a prioritization scheme to pinpoint key elements for follow-up experiments.

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 realize that the representative tumor and normal cell types and their pairings used here are rough in nature. However, cancer is a heterogeneous disease such 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 real tumor and normal tissues taken from a single patient.

Deleted: Note, h

Deleted: In addition,

Our study underscores the value of large-scale data integration, and we note that expanding the scale of our approach is straightforward. 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 further compact annotation sets, and more ChIP-seq/eCLIP experiments on additional factors would provide more detailed regulatory networks. Larger patient cohorts of expression and mutation profiles from many cancer types may be used to discover novel key features in cancer genomes. We also anticipate that an additional step may entail carrying out many assays on specific tissues and tumor samples. We hope that we demonstrate here that such large-scale integration is technically feasible and provides further opportunities for the future.



Page 4: [1] Deleted	jingzhang.wti.bupt@gmail.com	6/1/17 11:38:00 AM

Page 5: [2] Deleted	jingzhang.wti.bupt@gmail.com	6/1/17 11:49:00 AM

[[MG2JZ: see supplement]].