Integrating ENCODE data to interpret regulatory changes in cancer

Abstract

Cancer is caused by mutations in the DNA which disrupt the normal physiology of cells. While mutations on coding genes have been well characterized, the preponderance of mutations in tumors occur in non-coding regions and are still poorly understood. The new release of ENCODE data provides an opportunity to bridge these knowledge gaps. For a variety of cancer-derived cell lines, as well as non-cancerous cell lines derived from relevant tissues, ENCODE provides diverse genome-wide assays, such as Repli-seq, ChIP-seq, DNase-seq, STARR-seq, Hi-C, and ChIA-PET. The resulting data and functional maps of the human genome provide a framework to assess the potential for non-coding mutations to dysregulate genes.

In this paper, we first integrated diverse assays from ENCODE to define high-confidence regulatory elements and their gene linkage to define the extended gene neighborhood. We also developed a regression based method for background mutation rate calibration. It removes confounding effects from chromatin and replication timing and search for genes with higher than expected mutation frequency in the extended gene neighborhood. This approach successfully identified novel highly mutated genes, such as *BCL6* in leukemia, that are associated with patient prognosis.

Besides, we also integrated extensive binding profiles from ENCODE to build up tissue specific regulatory networks for both transcription factors (TFs) and RNA-binding proteins (RBPs). Intriguingly, through networks hierarchy analysis we found that TFs with higher mutation burden tend to be located at the bottom of the hierarchy (e.g., EZH2 and NR2C2), whereas those with dysregulated expression tend to reside at the top. Furthermore, by comparing tumor and normal network, we identified highly "rewired" TFs with changed targets and prognostic value, such as IKZF1 and MYC. We then extended tissue specific network to build up generalized networks across cancers. After combining with expression profiles from other cohorts, we pinpointed MYC and SUB1 as key regulators that significantly drive tumor to normal differential expression and then validated their effects through knockdown experiments.

Finally, we proposed a prioritization scheme for key mutations in cancer. We identified active enhancers and seven high impact mutations therein in breast cancer and validated their functional effects through luciferase assays.

Introduction

Mutations associated with cancer have been well characterized in key oncogenes and tumor suppressors. However, the overwhelming bulk of mutations in cancer genomes – particularly those discovered from the recent large-scale cancer genomics initiatives – lie within non-coding regions. Whether these mutations drive cancer development or progression, or simply emerge as byproducts of genomic instability remains an open question.

[JZ2MG: cited potential reviewer Matthieu Lupien]

Several recent studies begin to address this question by either directly employing a small group of non-coding annotations or incorporating limited functional genomics features for

mutation effect interpretation \{cite 25261935, 27064257, 27807102 }. For example, Weinhold et al investigated recurrent non-coding mutations in regulatory regions like promoters and discovered mutations in promoter that reduce gene expression and suggests poor prognosis. Wright et al found cancer risk-associated single-nucleotide variation (SNV) in enhancer regions that potentials upregulate MYC expression through long range interactions in colorectal cancer \{cite 20065031}. Lawrence et al incorporated several expression, chromatin, and replication timing profiles to quantify somatic mutation burden and identify cancer drivers \{cite 23770567}. However there is no systematical integration of thousands of functional genomic data sets from tens of experimental assays to interpret the cancer genome.

The newly-released data from the ENCODE Consortium can benefit such integrative analysis by providing comprehensive characterization of non-coding regulatory elements and linking them to cancer associated genes. The second phase of ENCODE was focusing on using RNA-seq and ChIP-seq data from multiple cell lines to define non-coding regulatory elements. Phase three ENCODE went into two directions. On one hand, it expanded the cell lines and tissue for these RNA-seq and ChIP-seq data to get a general catalog of regulatory element, which has been covered in the main ENCODE encyclopedia paper; on the other hand, focusing on the top tier cell line it expanded the number of sophisticated assays such as STARR-seq, Hi-C, ChIA-pet, and RAMPAGE. These improvements enable us to accurately identify distal regulatory elements such as enhancers and link them to genes, whereas the broader catalogue gives more general set of regulatory elements in many tissues. Here, we endeavor to provide a companion resource to the main ENCODE encyclopedia by focusing on cancer and building a "cancer encyclopedia". The main encyclopedia is oriented toward breath of the annotations to describe elements over hundreds of cell lines. In contrast, we focus on top tier cell lines with a wide variety of assays available. Most of these cell lines are associated with cancers of the blood, liver, lung, cervix, and breast. We show that these cell lines can be used to provide a better understanding of oncogenesis, and we provide a resource for interpreting the wealth of mutational and transcriptional profiles produced by the cancer community.

Data for comprehensive functional characterization in ENCODE

The most comprehensive set of assays for ENCODE are available for top tier cell lines. They provide good models not only for studying gene regulation in details, but also for understanding cancers of the blood (K562), breast (MCF-7), liver (HepG2), lung (A549), and cervix (HeLa-S3). For four of these five top tier cell lines, there is another immortalized cell line from corresponding healthy tissue. Therefore, comparisons of the data from cancer and normal cell lines could help model gene regulation in tumor versus normal tissues. It is worth noting that both relating these cell lines to cancers and pairing the tumor-normal matches are very approximate in nature, as these matches are not intended to substitute data from real tumor and normal tissues. Nonetheless, they are good models to perform a wide variety of functional genomics profiles, perturbation assays, and experimental validations. In addition, the wide variety of available omics data generated on these cell lines in ENCODE can be used to better interpret molecular profiles from tumor tissues and understand gene dysregulation in cancers.

(Fig 1A).

[JZ2MG: logic: raw data -> gene level -> network level]

To build a cancer encyclopedia with these cell lines, we first collected comprehensive functional genomics data to characterize factors that potentially affect somatic mutagenic processes. Then at the gene level, we tried to accurately identify both distal (enhancers) and proximal (promoters and regulator binding sites) regulatory elements. Specifically, we first focused on identifying enhancers and linking them to genes through an ensemble method (Fig.

S2). In contrast to methods relying on a single assay, we first used a pattern recognition based algorithm called CASPER on ChIP-seq and DNase-seq signals to search for enhancer candidates and then pruned them using peaks from our STARR-seq pipeline ESCAPE. We further applied our enhancer linkage prediction method JEME based on ChIP-seq, DNase-seq, ChIA-pet, and RNA-seq data to link these enhancers to genes. These potential linkages were then filtered through the results of Hi-C experiments, which provide a more accurate yet lower resolution map of chromatin interactions. For each gene, we combined these enhancers with proximal regulatory elements to construct what we termed "extended gene neighborhoods" – coding regions matched with key regulatory elements – to better interpret gene regulation (Fig1 B). Furthermore, at the network level we also explored the binding profiles in ENCODE and constructed high-confidence gene regulatory networks for both TFs and RBPs (Fig1 C). Finally, we merged our efforts with the broader ENCODE encyclopedia and provided consistent identifiers and definitions for the cancer encyclopedia.

In summary, our cancer encyclopedia consists a list of regions with higher than expected mutation frequency in cancer, accurately determined enhancers and gene linkages, the extended gene neighborhoods, the regulatory network of TFs (and for some lines RBPs), as well as the characteristics of TF/RBPs within the network, such as positions in network hierarchy, rewiring status, tumor/normal differential expression driving potential and prognostic value in various cancer types. Collectively, these resources allow us to prioritize a few key elements as being associated with oncogenesis, some of which are then validated using small-scale experiments (see table S1).

Multi-level data integration better enables recurrent variant analysis in cancer

One of the most powerful ways of identifying key elements and functional mutations in cancer is through recurrence analysis, which finds regions of the genome that are mutated more than expected. However, mutation process could be influenced by or associated with confounding factors (in the form of both external genomic factors and local context effects), which can result in many false positives or negatives in recurrence analysis. In addition, traditional methods often neglect the association among annotation categories and evaluate regions separately. Consequently, sometimes they fail to identify mutation signals from dispersed yet biologically relevant genomic regions, thereby limiting the interpretation power.

To address these limitations of traditional recurrence analysis, we integrated the cancer encyclopedia resources at two levels for better recurrence analysis. First, we predict an accurate local BMR by regressing out the confounding effects of features in a cancer-specific manner. Specifically, we prepare a covariate matrix by integrating 475 features at 1mb bins to remove those effects that may confound the BMR. We then separated the whole genome into 64 categories according to the local 3-mers and run separate regression models to further remove confounders from intrinsic sequence contexts. In contrast to methods that use unmatched data \{cite MutsigCV}, our regression-based approach with matched data usually yields higher BMR prediction precision (Fig 2A). In breast cancer, for example, the spearman's correlation (ρ) between observed and predicted mutation counts over 1-megabase bins increases from XX to XXX when using replication-timing signals from MCF-7 instead of HeLa-S3. This underlies the importance of integrating chromatin features from matched tissues to infer BMR (Fig 1B). For example, ρ only ranges from xxx-xxx using matched replication timing, but its range increases to xxx-xxx by adding 1 PC from the remaining covariates. It progressively increases to the xxx-xxx regime by adding PCs to the full model through forward selection (Fig 1B, see Supp. File/Section(?) X). Such noticeable improvements in BMR estimation significantly improve the recurrence analyses below.

Rather than separately testing standalone annotation categories, we employ our extended gene neighborhoods which contains both the coding exons and non-coding regulatory elements as joint test units (Fig 1C). Such a scheme allows for the accumulation of weak mutation signals distributed across multiple biologically relevant functional elements, which may otherwise be missed if evaluated under individual tests. We demonstrate that our scheme can effectively remove false positives and discover meaningful regions with more than expected mutations (Fig 2C). For example, in the context of K562 cells derived from a chronic lymphocytic leukemia (CLL), our analysis identifies well-known highly mutated genes, such as TP53 and ATM, that has been reported from previous coding region analysis. It also discovered new genes such as BCL6 that are missed by the analysis of coding regions. BCL6 has strong prognostic value with respect to patient survival (Fig. 2D), indicating that the extended gene neighborhood could be used as an annotation set for recurrence analysis. In addition, we can easily generalize this BMR calibration approach for other cancer types not in the five we are focusing on, as our model will work to pick an appropriately matched ENCODE signal type.

Extensive rewiring events in tissue specific network in cancer

We then investigated the transcription regulation network in a tissue specific way. In each cell type, we organized the TF regulatory network into a hierarchy by comparing the inbound and outbound edges of each factor, thereby enabling us to investigate the global topology of TF regulation (Fig. 1E, see also Supp. File/Section(?) X). TFs in different levels of the hierarchy reflect the extent to which they directly regulate the expression of other TFs \{cite 25880651}. For example, TFs in the top layer have more outbound than inbound edges in the network, and thus play larger roles in regulating other TFs (Supp. Fig. xx). In this representation, two patterns readily emerge. In leukemia, top-level TFs tend to more strongly influence the differential expression between tumor and normal cells. The average Pearson correlation between TF binding events and tumor-normal expression changes increases from 0.125 in the bottom layer to 0.270 in the top layer (Table Sx). TFs in the bottom layer are more frequently associated with burdened binding sites in general, perhaps reflecting their increased resilience to mutation (see Supp. Section X, Table Sx).

When comparing the common regulators in approximately matched tumor and normal regulatory networks, rewiring (i.e., target changing) analysis may help to identify cancer-associated deregulation. Hence, we investigated rewiring events in TF networks using multiple formulations (see Supp. File/Section(?) X). Specifically, for leukemia we removed the general TFs and restricted our rewiring analysis to 61 common TFs in K562 and GM12878 from ENCODE. We first ranked TFs according a "rewiring index" (Fig. 3 A), which calculates their respective number of lost and gained edges. Oncogenes such as MYC and NRF1 are among the top edge gainers. In contrast, IKZF1, whose somatic mutations serve as a hallmark of high-risk acute lymphoblastic leukemia, is the most significant edge loser, with up to xxx% of lost edges in K562 (Fig 3A). In contrast, several ubiquitously distributed TFs, such as YY1, retain their regulatory linkages (Fig 3A). We observe a similar trend in TFs using a distal, proximal and combined network (see details in supplementary file). We also observe highly rewired TFs in lung and liver cancers (see fig XX) although we do not have as many common TFs between tumor and normal cell lines for these tissues.

Our rewiring index only considers direct connections associated with a given TF. One may also consider rewiring that include not only direct connections, but also the whole neighborhood of connections with which a TF associates through membership and topic models. In particular, we used a mixed-membership model to look more abstractly at local gene neighborhoods to re-rank the TFs (see Supp. File/Section(?) X). Similar patterns are observed using this model. We also observed that MYC (a well-known oncogene) becomes a top edge gainer (Fig 3A). To study

the consequences of network rewiring under this model, we performed survival analysis on xxx AML patients and found IKZF1 to be significantly associated with prognosis.

A remaining uncertainty lies in the associated factors of such rewiring. We find that the majority of rewiring events are associated with noticeable changes in chromatin status, but not necessarily with variant-induced loss or gain events (Fig. 3A). For example, JUND is a top gainer in K562 and majority of its gained or lost targets experienced substantial expression and chromatin status changes (at least 2-fold). It is interesting to further investigate the causal relationship between such changes and the rewiring events.[JZ2MG: since we are not able to make a conclusion here, I just said something very vague, but maybe it is too vague…]

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

Next, we merged the tissue specific networks and performed a generalized pan-cancer generalized network analysis across multiple cancer types for both TFs and RBPs. Using a machine learning method, we integrated 8,202 tumor expression profiles from TCGA to systematically search for TFs and RBPs that drive tumor-specific expression patterns. Our method tests whether the regulatory targets are sufficiently correlated with the regulator's molecular status across tumors. The final output is the estimated fraction of patients with target genes differentially regulated between each pair of cancer type and regulator. The overall trends for the key TFs and RBPs detected are given in Fig. 4A. The predicted impacts of regulators on tumor gene expression are highly consistent with previous findings.

We find that the target genes of MYC are significantly up-regulated in numerous cancers, which is consistent with the known role of MYC as an oncogenic TF. We further validate MYC's regulation effect through CRISPi RNA-seq experiments. Consistent with our prediction, expression of MYC targets are significantly reduced after MYC knockdown (Fig 4A). In addition, we further investigated how MYC interacted with other TFs to jointly control their target gene expressions. We found that genes shared with MYC and other TFs only showed marginal partial correlation with MYC's co-regulatory TFs, indicating a major role of MYC for gene expression control. We then the particular network motif feed-forward loops (FFLs) with MYC as the master regulator. Among the 38 other TFs that form FFLs with MYC, we selected NRF1 since it has the most common targets except two well-known MYC partner TFs. We then checked the logic gate usage among MYC and NRF1 and found that either MYC itself or the or gates dominates across multiple cancer types, confirming the major role of MYC in gene regulation. We also discovered that these MYC-NRF1 FFLs are mostly coherent ones, where NRF1 serves as an amplifier to the MYC effects on the target genes. This is consistent with the previous discovery that NRF1 is intervening to MYC's apoptotic function \{cite 12533512 }.

In addition to recapitulating existing knowledge from previous studies, our analysis also predicts previously unidentified functions for regulators in cancer. For example, the predicted targets of the RBP SUB1 were significantly up-regulated in many cancer types (Fig. 4C). Moreover, the up-regulation of SUB1 target genes is correlated with a worse patient survival in cancer types such as lung cancer (Fig. 4). Previously, SUB1 was considered as a TF. However, the ENCODE eCLIP experiment has profiled many SUB1 peaks on gene'3UTR regions. In HepG2 cell where SUB1 eCLIP experiment was done, the decay rate of SUB1 target genes are significantly shorter than non-targets (Fig. 4C). After knocking down SUB1, its predicted targets are also down regulated comparing to other genes (Fig. 4D). These results indicate that SUB1 may bind gene 3'UTR regions to stabilize transcript level. From our integrated analysis, the

higher SUB1 activity through regulatory binding on 3'UTR regions is likely to drive tumor specific expression patterns in many cancer types.

Step-wise prioritization schemes pinpoint deleterious SNVs in cancer

The above description of the regulatory network and mutation recurrence analysis provided an approach to prioritize key genomic features associated with cancer. The workflow in Fig.5 A describes this prioritization scheme in a systematic fashion. First, we start by searching for key regulators that frequently rewired, locate in network hubs or on 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 and chromatin changes, or (most importantly) 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 create specific binding sites, or which occur in positions of particularly high conservation.

Using this framework, we subject a number of key regulators, such as MYC and SUB1, to knockdown experiments to validate their regulatory effects (Fig 4D). We then identified several active enhancers in noncoding regions, and validated their ability to influence transcription using luciferase assays. We further selected key SNVs within these enhancers that are important for gene expression control. Of the eight motif-disrupting SNVs that we tested, six showed consistent up- or down-regulation effect in expression relative to the wild type. One particularly interesting example, illustrating ENCODE data integration, is on chromosome 6, 13.5xxx (Fig. 5C). This enhancer is located in a noncoding region. Both histone modification and DHS signals implicate its regulatory role as being active (Fig. 5C), and both our CASPER enhancer prediction method and the STARR-seq experiment support its enhancer function (Fig. 5D). Hi-C and ChIA-PET data link this region to a downstream gene SYCP2. We found that 21 out of the 52 ChIP-Seq experiments in MCF-7 from ENCODE demonstrate high frequency of chromatin interactions in this region. Motif analysis predicts the C to G mutation in cancer can significantly disrupts the FOLS2 binding affinity. Luciferase assays demonstrate that this mutation introduces an xx-fold reduction in expression relative to wild type expression levels, indicating a strong repressive effect on this enhancer's functionality.

Conclusion

This study highlights the values of our cancer encyclopedia as a resource for cancer research and leverages it to provide a prioritization scheme to pinpoint key regulatory elements and SNVs for small-scale validations. A key inspiring aspect of our analysis is that by sheer data integration from various assays we demonstrate importance of more accurate non-coding elements and their linkages definitions to interpret both mutation and expression data in cancer. We can straightforwardly envision a path forward through expending this framework by integrating additional types of assays and applying to other cancer relevant cell lines and eventually tissues. We also anticipate that by employing more mutation and expression profiles from larger cohorts will definitely benefit our interpretation. However, there are some limitations in this work. One is to appropriately match cancer and normal cell line/tissues for different cancer types. The major leap would be taken is to do this type of work on real cancer tissue. Nevertheless, we are still very encouraged to see that the resource we collected and some of the frameworks we developed are useful to understand some bits of the expression and mutation drive real tissues.