In total 2970 words

Introduction: 368

Data section: 540

BMR section: 438

Rewiring section: 324

Generalized network section: 568

Validation: 476

Conclusion: 217

Large-scale ENCODE data integration to interpret regulatory changes in cancer

EN-codec: A large scale integrative resource from ENCODE for cancer research

EN-codec: A large scale ENCODE integrative scheme to interpret cancer genomes

Introduction

A small fraction of mutations associated with cancer have been well characterized, particularly those coding regions of key oncogenes and tumor suppressors. However, the overwhelming bulk of mutations in cancer genomes – particularly those discovered over the course of 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 \cite{26781813}.

Several recent studies have begun to address this question by incorporating limited functional genomics data for variant interpretation \{cite 25261935, 27064257, 27807102}. For example, Hoadley et al integrated five genome-wide platforms and one proteomic platform to uniformly classify various tumor types \{cite 25109877}. Torchia et al integrated various genomic and epigetnic 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 systematical integration of thousands of functional genomic data sets from tens of experimental assays of various types to interpret the cancer genome.

One way to approach noncoding variant functional interpretation problem is to experimentally evaluate the functional effects of mutating individual bases. This is a major endeavor of the ENCODE Consortium. In the initial release of the ENCODE annotation years ago, this was predominantly accomplished using RNA-Seq and ChIP-Seq assays on a limited number of cell types \cite{22955616}. The new release of ENCODE took two new directions. First, it considerably broadened the number of cell types with the main RNA-Seq, ChIP-Seq, and DNase-seq assays; the main ENCODE encyclopedia aims to utilize this to provide a general, unified annotation resource applicable across many cells. Secondly, ENCODE expanded the number of sophisticated assays such as STARR-seq, Hi-C, ChIA-pet, eCLIP and RAMPAGE on several top-tier cell lines, many of which are cancer-associated. This enables precise definitions of enhancers, direct identification of enhancer-target gene links, and the construction of RNA-binding protein (RBP) networks. Here, we focus on top-tier cell lines by performing large-scale integration of these various assays to construct an in-depth cancer related companion resource to the general encyclopedia. We call this the "companion *ENCODE* encyclopedia resource for cancer" (or "EN-codec" for short) for interpreting the wealth of mutational and transcriptional profiles produced by the cancer research community.

2

Style Definition: Heading 1

Style Definition: Heading 2

Style Definition: No Spacing,TextBody

Comment [j1]: SL comments: model instead of "regulatory changes"
go resource: large scale integration for cancer resources, no cancer encyclopedia
integration models/resources for cancer

Comment [j2]: Key question: do we call SNV variant or mutation in this paper?

Comment [j3]: I prefer to use resource

Comment [j4]: I prefer to use resource

Deleted:

Comment [j5]: This work is targeting Josh Stuart as a reviewer. He is the Last PI on this paper except TCGA consormtium

Deleted: Weinhold et al. investigated recurrent non-coding mutations in regulatory regions

Formatted: Font:Not Italic

Comment [j6]: Matthieu Lupien is in this group but not a predominant author

Formatted: Font:(Default) Times New Roman, Font color: Auto, Pattern: Clear

Deleted: Wright *et al*. found cancer risk-associated single-nucleotide variants (SNVs) in enhancer regions that potentially upregulate MYC expression in colorectal cancer \{cite 20065031}

Comment [j7]: In case Gaddy is the reviewer

Comment [j8]: MP: Any of Shirley Liu's work?

Formatted: Highlight

Deleted: the functional interpretation

Deleted: s

Comment [j9]: I this true? I think this is too strong

Deleted: five

Comment [j10]: MP: I would say cell types; there were plenty of primary cells in ENCODE 2, including NHEK, HMEC.

Deleted: lines

Comment [j11]: I don't know what you mean h ... [1]

Deleted: and

Deleted: and their relevance to cancer

Deleted: cancer-relevant

Formatted: Font:Italic, Underline

Deleted: companion

Formatted: Font:Italic, Underline

Deleted: the

Deleted: encyclopedia companion

Often

MULTI NORM

Comprehensive functional characterization by ENCODE data integration

The ENCODE top-tier cell lines provide good models not only for studying gene regulation in detail, but also for understanding cancers of the blood (K562), breast (MCF-7), liver (HepG2), lung (A549), and cervix (HeLa-S3). In different contexts, these top-tier cell lines can be "paired" with functional genomics data form normal tissue (often from epigenome roadmap) or another immortalized cell line from corresponding healthy tissue (Fig 1 A). Comparisons of these "TN-pairs" could help to model the differential gene regulation between tumor and normal tissues. It is worth noting both relating these cell lines to cancers and pairing the tumor-normal matches is approximate in nature and are not intended to substitute real tumor and normal tissues. However, cancer is such a heterogeneous disease that even the tumor cells from one patient usually shows 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. We believe that these "TN-pairs" still serve as good models for performing a wide variety of functional genomics profiles, perturbation assays, and experimental validations. Furthermore, many of these pairings have been used in previous analyses \{cite 25144821, 1975513}(Figure 1 A & supp Fig. s2).

To build the companion encyclopedia, we started by defining enhancers. Unlike the ENCODE encyclopedia, we used genomic signal tracks from a battery of 5 to 10 histone modification marks in combination with DNase-seq. These were used as input into CASPER, a machine learning predictor that we developed to integrate the signal shapes of these various signals. We then assembled these predictions with peaks called from STARR-Seq experiments, which directly read out candidate enhancers in the genome. Such an integrative approach gives accurate definitions of enhancers (see supplement). We then used RAMPAGE data to better define promoters, and further linked enhancers to putative promoters using a deep learning algorithm. These potential linkages were then further filtered through the results of Hi-C and ChIA-pet experiments to obtain high confidence enhancer target linkages. These enhancer-target linkages, refined promoters, and RNA binding sites from eCLIP experiments within genes constitute a so-called extended gene neighborhood (Fig1 C).

We further linked the enhancers and promoters with their predicted associated transcription factors (TF) to construct extended regulatory networks. First, we built cell-type-specific distal and proximal TF regulatory networks by linking TF to genes, either directly by TF-promoter interactions or indirectly via TF-to-enhancer-to-gene interactions (Fig1 B). We then pruned these networks to include only the strongest edges using another signal shape algorithm called TIP \{cite 22039215}. In paired "tumor-normal" cell lines, we measured the signed, fractional number of edges changing, the rewiring index, and ranked TFs by this. In addition, we merged our cell-type-specific networks to get a generalized network for pan-cancer analysis. For each network, we then arranged all regulators into a hierarchy. TFs are placed into different levels of the hierarchy to the degree which they directly regulate the expression of other TFs \{cite 25880651} or are in turn regulated by them. A final hierarchical network structure is shown in Fig1 D. This shows that the top layer TFs are not only enriched in cancer associated genes but also more significantly drive tumor-to-normal gene differential expressions. We also observe that highly mutated TFs tend to sit at the bottom of the hierarchy.

[JZ2MG: actually I personally feel a little bit uncomfortable of all using the present tense through the paper. I agree with Shirley that past tense is better. Please advise]

2

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

[JZ2MG: is this better?] Multi-level data integration benefits variant recurrence analysis in cancer?

One of the most powerful ways of identifying key elements and functional mutations in cancer is with recurrence analysis to discover regions that mutate more than expected. However, somatic mutational processes can be influenced by numerous confounding factors (in the form of both external genomic factors and local sequence context factors), which can result in many false positives or negatives without appropriate correction \{cite 23770567}. In addition, traditional methods often neglect the natural association of different annotation types (e.g. a gene body and its linked enhancer) and evaluate regions separately. Consequently, they sometimes fail to identify mutational signals from distributed yet biologically relevant genomic regions, thereby limiting their functional interpretation.

To address these limitations, we adopt a two-pronged approach for better recurrence analysis. First, we predict an accurate local background mutation rate (BMR) by removing effects of confounding factors in a cancer-specific manner. Specifically, we separated the whole genome into bins (1Mb) and calculated mutation counts under each local context category. For each category, we used a negative binomial regression of the mutation counts against features like replication timing, chromatin accessibility, Hi-C signal, and expression profiles for BMR prediction. In contrast to methods that use unmatched data \{cite 23770567}, our approach automatically selects the most relevant features, thereby

IMPROVES

Comment [j12]: SLiu:HepG2 a good model for liver cancer? Shirley thinks it depends on reviewer

Deleted: , as these matches are not intended to substitute d

Deleted: from

Deleted: Nonetheless, they

Comment [j13]: MP: I think there is probably a better reference for the idea that MCF7 and MCF10 are a tumor/normal pair

Deleted: 26241649

Comment [j14]: MP: ENCODE encyclopedia?

Deleted: general

Deleted: Casper

Deleted: .

Deleted: Casper

Deleted: s

Deleted: around enhancers

Deleted: ensembled

Deleted: prediction

Deleted: based on gene expression

Comment [j15]: MP: Provide a map; the functional data don't map physical interaction, only functional interaction that is presumed to work through physical interaction

Deleted: , which provide a more accurate map of physical interactions

Deleted:

Comment [j16]: MP: With predictions of their associated TFs

Deleted: tissue

Deleted: -

Deleted: We use a simulated annealing procedure to search for the placement that maximizes the number of downward flowing edges.

Deleted: This shows how

Deleted: and those associated with large expression changes at the top

Deleted: better

Deleted: recurrent

Deleted: variant

Deleted: with greater mutational burden

Deleted: effects

Deleted: mutations among

Deleted: types of linked annotations providing noticeable improvements in BMR estimation, which significantly benefits recurrence analyses (Fig 2A). Notably it requires the combination of many different genomic features to get such an accurate estimation (Fig 2 B)

[JZ2MG: do you think we should talk about AML, CML, CLL, ALL very clearly in the main MS? I have mixed feeling of that]

Second, rather than separately testing standalone annotation categories, we used our extended gene neighborhoods as joint test units that contain both the coding exons and non-coding regulatory elements (Fig 1C). Such a scheme allows for the accumulation of weak mutational 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 higher-than-expected mutation counts (Fig 2C). For example, in the context of chronic lymphocytic leukemia (CLL), our analysis identifies well-known highly mutated genes, such as TP53 and ATM, which has been reported from previous coding region analysis. It also discovered genes that are missed by the exclusive analysis of coding regions, such as BCL6. Note that BCL6 has strong prognostic value with respect to patient survival (Fig. 2D), indicating that the extended gene neighborhood could be used as an annotation set for recurrence analysis. In addition, we can easily generalize this BMR calibration approach for other cancer types beyond the five discussed here, as our model will pick an appropriately matched ENCODE feature type.

Extensive rewiring events in regulatory network during normal to tumor transition

the

[JZ2MG: see MP comment here. Why he is thinking we are prioritizing motifs, not TFs?]

We then investigated the transcriptional regulatory network in a cell-type specific way to highlight the key regulators in cancer. Here, we utilized 4 main tumor-normal cell line pairings described earlier to study how the targets of each common TF changed (i.e., rewired) over the course of oncogenic transformation. We first ranked TFs according the "rewiring index" (Fig. 3 A). In leukemia, well-known oncogenes such as MYC and NRF1 are among the top edge gainers, while the well-known tumor suppressor IKZF1is the most significant edge loser (Fig 3A). Mutations in this later factor serve as a hallmark of various forms of high-risk leukemia \cite{26202931, 26713593, 26069293 }. Interestingly, IKZF1 loss has been found to be associated with 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 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 such as BHLHE40, JUND, and MYC in lung, liver, and breast cancers (Fig 3).

Our rewiring index only considers direct connections associated with a given TF. However, the targets within the TF regulatory network are 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 these gene communities via a mixed-membership model. This enabled us to evaluate each TF's overall association changes to these gene communities in tumor and normal cells. Similar patterns are observed using this model to using the rewiring index (Fig 3A).

[JZ2MG: the newly added sentence is actually a trouble maker. If we add H1, we should skip this.]

We find that the majority of rewiring events are associated with noticeable gene expression and chromatin status changes, but not necessarily with variant-induced motif loss or gain events (Fig. 3A). 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. We found a similar trend for the rewiring events associated with JUND in liver cancer.

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

Next, we extended our network analysis in a pan-cancer fashion by merging the cell-type-specific networks for both TFs and RBPs. Then 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 to the degree a regulators' regulation potentials are sufficiently correlated with their targets' tumor-to-normal expression changes. We then calculated the percentage of patients with these relationships in each cancer type and presented the overall trends for key TFs and RBPs in Fig. 4A.

We find that the target genes of MYC are significantly up-regulated in numerous cancers, which is consistent with its well-known role as an oncogenic TF and a transcription activator \cite{22464321}. We further validate MYC's regulatory effect through external knock down experiments (Fig 4). Consistent with our predictions, the expression of

Deleted: Almost n

Deleted: signatures

Comment [j17]: MP: Is this published? In this manuscript, you don't claim to have a CLL tumor. If you mean K562, that is CML, not lymphoid

Deleted: signal

Deleted: in cell-type specific

Deleted: in cancer

Comment [j18]: MP: You're really ranking motifs predictive for TFs, or TF families, right? IKZF1-5 bind the same sequence, right? Do you filter out the family members that are not expressed?

Comment [j19]: RK: IKZF1 is a lymphoid TF

Comment [j20]: MP: Good news if you have a model for AML; are they linked to CML?

Deleted: acute lymphoblastic

Comment [j21]: This gene is interesting, it is not only a marker for AML, but also for ALL. So I think it is not purely due to myloid and lymphoid difference

Comment [j22]: MP: Again, do you mean these specific factors, or do you mean TF family members that bind to these motifs?

Comment [j23]: MP: You have seen PMID: 25363779, PMID: 24121437, PMID: 24760698, PMID: 26298065, PMID: 28137873?

Deleted: cell

Deleted: tumor-to-normal expression changes

Deleted: expression

Deleted: changes

Comment [j24]: And transcriptional activator

Deleted: CRISPi RNA-seq

MYC targets is significantly reduced after MYC knockdown (Fig 4A). After confirming the importance of MYC, we use the regulatory network to understand how MYC works with other TFs. We first looked at all triplets involving MYC by requiring that a second TF both interacts and shares a common target with MYC. In all cancer types, we found that MYC's expression levels are positively correlated with the expressions of most of its targets, while the second TF shows only a limited influence as determined from partial correlations. We then investigated the exact structure of such regulatory relationships. The most common triplet interaction type is a well-understood feed-forward loop (FFL) structure in which MYC regulates both the common target and the second TF. Most of these FFLs involve ~~with~~ well-known MYC partners such as Max and Mxl1. However, we also discovered that many involve another factor called NRF1. Upon further study, we found that that the MYC-NRF1 FFL relationships were mostly coherent ("amplifying") FFLs. We further studied these FFLs by forming these triplets into a logical gate, in which the two TFs act as inputs and the target gene expression represents the output \{cite 25884877}. We can show that the predominant number of these gates follow either OR or MYC-always-dominant logic. Thus, the ENCODE regulatory network not only helps find key regulators, but also to really demonstrate how they work in combination with other regulators.

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

[JZ2MG: MP questioned that some RBPs in Figure4 are not RBP… But somehow they are eCLIPped. Not sure whether we should go into so much details]

REWORD

3UTR

Step-wise prioritization schemes pinpoint deleterious SNVs in cancer

Summarizing the analysis described above, ~~our~~ EN-codec ~~resource~~ consists of number annotation resources: (1) a BMR model with matching procedure and a list of regions with higher-than-expected mutations in various cancers, (2) accurately determined enhancers, promotors and enhancer-target-gene linkages by integrating tens of different functional assays and their comparison with those in ENCODE encyclopedia; (3) extended gene neighborhoods, (4) tumor-normal differential expression and chromatin changes, (5) a regulatory network of TFs; (6) based on the network, for each TF position in the network hierarchy and 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. 5A describes this prioritization scheme in a systematic fashion. We first search for key regulators that are frequently rewired, located in network hubs or at top of the network hierarchy, or significantly driving expression changes in cancer. We then prioritize functional elements that are associated with top regulators, undergo large regulatory changes in terms of expression levels, 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 create specific binding sites, or which occur in positions under strong purifying selection.

Using this framework, as we described above, we subject a number of key regulators, such as MYC and SUB1, to knockdown experiments to validate their regulatory effects in particular cancer contexts (Fig 4D). Next here, we also identified several candidate enhancers in noncoding regions, associated with breast cancer, and validated their ability to influence transcription using luciferase assays in MCF7. We selected key SNVs, based on significantly recurrent mutations in breast cancer cohorts, within these enhancers that are important for controlling gene expression. Of the eight motif-disrupting SNVs that we tested, six showed 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 the intronic region of CDH26 in chromosome 20 (Fig. 5C). Both histone modification and chromatin accessibility (DNase-seq) signals indicated an active regulatory role in MCF7, which was further confirmed as an enhancer by both CASPER and ESCAPE (STARR-seq) (Fig. 5D). Hi-C and ChIA-PET data indicated that the region is within a topologically associated domain (TAD) and validated a regulatory linkage to the downstream breast-cancer-associated gene SYCP2 \cite{26334652, 24662924}. We observed massive binding events from TFs in this region in MCF-7. Motif analysis predicts that the particular mutations found in the cohorts can significantly disrupt the binding affinity of several TFs, such as FOSL2, in this region (Fig. 5D). Luciferase assays demonstrate that this mutation introduces a ~~3.58~~-fold reduction in expression relative to wild type expression levels, indicating a strong repressive effect on this enhancer's functionality.

3.6x

Conclusion

This study highlights the value of our companion to the encyclopedia as a resource for cancer research. First, we show that, by integrating many different types of assays on a large scale, we can achieve a very accurate annotation of

Deleted: 3,735

Comment [j25]: MP: What does this mean?

Deleted: expressions

Deleted: , which has not been previously associated with MYC

Deleted: RBPs

Deleted: Sub1

Deleted: [[to Shirley: is this true? There are other papers mentioned it from other aspects in cancer. need to confirm with Peng?]]

Comment [j27]: MP: ig 4A RBPs includes TAF15 and GTF2F1; these are part of the Pol II initiation complex, they associate with DNA prior to initiation of transcription, and facilitate initiation.

Deleted: companion encyclopedia

Deleted: and

Deleted: mutational frequencies

Comment [j28]: MP: Doesn't the ENCODE encyclopedia have these things? Are your calculations the same or different? Shouldn't this be explained either way?

?

Deleted: of particularly high conservation

Comment [j29]: MP: candidate

Deleted: active

Deleted: further

Comment [j30]: MP: Consistent with oncogenesis model? Reproducible? Consistent with neighboring SNV?

Deleted: Casper

Comment [j31]: MP: Every locus lies in a TAD; do mean the candidate enhancer and predicted target promoter are in the same TAD?

Comment [j32]: MP: Several bound factors?

Comment [j33]: MP: FWIW, there are both sequence specific TFs and cofactors bound at this location, in Fig 5D, nicely marked by DHS and flanking H3K27ac; one known function for GATA3 is a key factor in mammary epithelia

Deleted: L

Deleted: n

ENCODE top-tier cell lines and relate them to cancer to build up extensive regulatory networks. Then we show how comparisons within this resource itself can illuminate potential regulatory changes in cancer (e.g. key rewiring TFs). Next, we show how the resource can be generalized into a pan-cancer regulatory network and BMR framework to help interpret patient data from cancer cohorts, both gene expression and mutation data. Finally, we show how we can leverage the companion resource to provide a prioritization scheme to pinpoint key regulatory elements and SNVs for small-scale follow-up. This study underscores the value of large-scale data integration, and we note that expanding this approach (either by integrating additional data types and/or using tumor mutation and expression data on a larger scale) is straightforward. We also anticipate that an additional step would be to carry out many of the ENCODE assays on specific tissues and tumor samples. Though volume of material needed for such analyses may present challenges, we show that such a framework is technically feasible and provides further opportunities for the future.

HETERO

Comment [j34]: MP: Figure 6, which I don't think you refer to, appears to have TBP, TAF1, and RNA Pol II in the network. As these are required for transcription of all protein-coding genes (and many non-coding genes), what does this mean? There is no connection between them, yet TBP and TAF1 are part of a complex that recruits Pol II to promoters, what does that mean? RAD21 and SMC3 are part of the cohesion complex yet there aren't any links connecting them, what does this mean?

Deleted: ie

Comment [j35]: Comments from SLiu:

Conclusion, we plan to write more
Mention actual tissue??? More limitation about matching
Nobody is perfect, talk about heterogeneity
Micro environment
Whole genome is still limited? Power analysis about whole genome, extended gene is important
Splicing is also noncoding regulation
kinase / signaling network
Kinase activity change can drive TF (phospholyation?)

Deleted: for

Deleted: ing

Page 1: [1] Commented	jingzhang.wti.bupt@gmail.com	3/30/17 3:24:00 PM

I don’t know what you mean here by “main” assays; worth noting that the assays covering the most human cell types are DNase-seq and RNA-seq (about 250 each).