Using the ENCODE regulatory data to interpret non-coding somatic variants in cancer

[JZ2MG]##### around 2500 word without abstract and removed

Long Abstract New version

[JZ2MG]##### long abstract 574 word

Though the impacts of somatic mutations within the very limited number of cancer-associated genes are well understood, the overwhelming number of mutations in cancer genomes fall within non-coding regions, rendering them far more difficult to evaluate. Data obtained from the new ENCODE release allows us to bridge these gaps in knowledge.

Here we collected comprehensive data from ENCODE and deeply integrated them to interpret cancer genomes. In particular, we combine computational predictions with EnhancerSeq experiments to generate high quality enhancer lists in several cell lines. To link the enhancers to genes, we performed enhancer target prediction by synthesizing evidences from expression profiles and chromatin status, and further pruned it using Hi-C data for higher accuracy. With this high quality linkage, we are able to define the extended genes by combining coding regions with key regulatory elements in enhancers and promoters for better functional interpretation. In addition, we also explored the full spectrum of the binding profiles from ENCODE and set up high confidence gene regulatory networks for both transcription factor (TF) and RNA binding proteins (RBPs).

We then integrated various signal tracks from comprehensive experiments to carry out rigorous recurrence analysis on the proposed extended gene regions. Specifically, we calibrated a genome-wide background mutation rate (BMR) by regressing out the effects from well-known confounders, such as replication timing and chromatin status. Then we performed a joint burden test on the extended genes to amplify mutation signals that might be weakly distributed in individual elements. Analyses show that our scheme could effectively remove false positives and discover meaningful burdened regions. In the context of leukemia, our analysis identified well-known drivers (such as TP53 and ATM) and key genes (BCL6) that has strong prognostic value but missed by coding region analysis.

We then explored the structure of TF-TF network by organizing it into a stratified hierarchy through comparison of outbound edges to inbound ones. We find that top-level TFs tend to be more associated with tumor-to-normal differential expression, and bottom-level TFs (e.g., EZH2 and NR2C2) tend to be enriched with burdened binding sites. We then rigorously compared TF regulatory networks between loosely matched tumor and

Jing Zhang 1/26/2017 5:34 PM
Comment [1]: This sentence reads a little bit weird since the "though" part is pretty long, but I am not a native speaker, so please suggest

Jing Zhang 1/26/2017 5:34 PM
Comment [2]: Need to confirm with Kevin Yip

Jing Zhang 1/26/2017 5:34 PM
Comment [3]: It is more coherent to say enhancers and promoters, since we just mentioned how to set up enhancers. But this will limit our extended gene definition fur further extension. Somehow, we could potentially say that the extended gene can be easily incorporated with other regulatory annotations

Jing Zhang 1/26/2017 5:34 PM
Comment [4]: Strong is not a good word here? Please suggest alternative.

normal cell lines and identified significantly rewired (i.e., target-changing) TFs, such as IKZF1 and MYC. By integrating large-scale chromatin features and whole genome sequencing (WGS) data, we demonstrate that such massive tumor-to-normal rewiring events may largely be explained by changes in chromatin structures, rather than direct mutational effects. Patient survival analyses reveal that the regulatory activity of our top rewired TF (IKZF1) is significantly associated with cancer progression.

We further integrated expression data from multiple cohorts into the more generalized TF/RBP network to prioritize key regulators that significantly drives the differential expression between normal and tumor cell lines in multiple cancer types. We identified ZNF687 as a key TF for breast cancer and SUB1 as a key RBP for liver and lung cancer. We further validated the effect of these TF/RBPs through different siRNA knockdown experiments.

Finally, we developed a step-wise scoring workflow to prioritize key variants in a cancer specific way. In particular, we identify several active enhancers in breast cancer pinpointed variants therein that potentially affect their downstream gene expressions. Experiments on both wild and mutant type sequences through luciferase assays confirmed their effects in MCF-7.

Our work demonstrates how careful integration of ENCODE resources offers unprecedented opportunities to accurately characterize oncogenic regulation and serves as a powerful tool to prioritize cell-type specific regulatory elements and variants in cancer.

Jing Zhang 1/26/2017 5:34 PM
Comment [5]: Is this correct in grammar?

Jing Zhang 1/26/2017 5:34 PM
Comment [6]: We should differentiate regulatory elements from regulators. RE means for example TFBS, but regulators means TF/RBP

Short Abstract

#/*= requirement from Nature **Articles:** an abstract of approximately 150 words =*/
#/*= right now 271 words =*/

While the majority of somatic mutations occur in non-coding regions, we only understand mutations well in very limited cancer-associated genes. Data obtained from the new ENCODE release allows us to bridge the gaps. Here we collected comprehensive data from ENCODE and deeply integrated them to interpret cancer genomes.

In particular, we provided high quality enhancer list and their gene target linkage to define the extended genes and then set up gene regulatory networks for both transcription factors (TFs) and RNA binding proteins (RBPs). Based on these data, we performed rigorous recurrence analysis on the proposed extended gene regions after integrating comprehensive signals. We identified well-known drivers (such as TP53 and ATM in CLL) and key genes (BCL6) that has strong prognostic value but missed by coding region analysis. We then explored the hierarchy of TF-TF network and find that top-level TFs tend to be more associated with tumor-to-normal differential expression, and bottom-level TFs (e.g., EZH2 and NR2C2) tend to be with more frequent burdened binding sites. We investigated the edge gain and loss events in matched tumor/normal networks and identified significantly rewired TFs (e.g., IKZF1 and MYC) that are highly associated with cancer progression. We further integrated expression data into the more generalized TF/RBP network and identified ZNF687 and SUB1 as key regulators in breast and lung

cancer and validated their effects through knockdown experiments. Finally, we developed a step-wise scoring workflow to pinpoint key variants and confirmed their effects through luciferase assays. Our work demonstrates that data from ENCODE after careful integration may serve as a powerful resource for the cancer community to investigate and prioritize key regulators and variants.

Introduction

The advent of whole genome sequencing (WGS) and personal genomics have opened the opportunities to identify key regulatory elements and deleterious mutations therein that are important for carcinogenesis, which in turn enables development of targeted therapies in clinical studies. Despite the collaborative efforts of many consortia to sequence tens of thousands of mutations, only a very small fraction of them are easily interpretable in terms of their effect to known cancer-associated genes. An open question is to what degree do the many thousands of non-coding mutations contribute to cancer? Are they simply neutral passengers created because of the dysregulation of the few known drivers or within them are lurking some key cancer-forming mutations? The new release of ENCODE resources may potentially bridge the gap by providing accurate non-coding annotation and precisely linking them to well-characterized protein coding genes.

Jing Zhang 1/27/2017 3:54 PM

Comment [7]: JZ2MG: I carefully read your text sent to me but select some to insert into this para. The key reason is that I need to de-amphasize the variants, because both rewiring and rabit part is reall NOT about SNVs

The ENCODE data resource provides a fundamental annotation on the human genome. It is done in a cell line specific manner with decent number of cell lines to be cancerous ones. It still imperfect data for cancer research because some cell lines are suboptimal for actual tumor samples and matching them with appropriate normals is often challenging. However, because of the tremendous richness of assays available here, comparison of tumor-like and more normal-like cell lines provides an unprecedented and accurate window into the regulatory and chromatin changes in cancer.

Here we endeavor to try to make the ENCODE resource as useful as possible for cancer research. We match the cell lines as best as possible with known cancers to better integrate relevant expression profiles and somatic mutations from known cohorts. Then, we show how we can develop methods to integrate comprehensive ENCODE signal tracks to calibrate an accurate background mutation rate (BMR) and provide this BMR as a resource. This allows us to accurately find burdened regions in many cancers. Besides for each of the known cell lines we use the wealth of ENCODE assays to accurately determine non-coding elements, particularly enhancers, and regulatory networks involving transcription factors (TFs) and actually, to a lesser degree, RNA binding proteins (RBPs). We how these regulatory networks in a variety of ways, particularly as hierarchies with master regulators at top. Then, we are able to, for each of the regulators in the regulatory network, calculate a rewiring score that represents the degree to which these regulators change in the course of cancer. We have provided as a resource, for each of the main ENCODE cell lines then, a list of accurately determined enhancers, a list of burdened regions, the regulatory network for the TFs determined and the most rewired and changed TFs in this regulatory network.

Jing Zhang 1/27/2017 4:21 PM

Comment [8]: JZ2MG: BMR is not a resource since it is data specific. I think the data matrix is a resource

These resources allow us to prioritize a few key elements as being associated with oncogenesis. On one hand some are large-scale elements such as TF and RBPs. Others

are also smaller scale elements such as particular enhancers and even mutations in these. We validate a number of these prioritizations with small-scale studies such as involving luciferase assays or TF knockdowns, finding most of these prioritizations are in fact accurate.

Comprehensive functional characterization data in ENCODE

Efforts of the ENCODE project has led to a surge in functional annotation data of the human genome, from transcription level to chromatin and nuclear organization level. Since over XXX percent of the cell lines provided by ENCODE are cancer cell lines, the raw data from ENCODE may serve as an invaluable resource for cancer research (see table S1). Here we created a comprehensive list of raw datasets (Figure 1A) from ENCODE to interpret cancer genomes. In addition to the raw data being highly relevant to cancer, ENCODE annotations also demonstrates great breadth, expanding genomic insight from only the coding region to over xxx percent of the noncoding annotated regions of the genome (Table S2). This significant increase in data provided by ENCODE could benefit functional interpretation in cancer.

Despite the comprehensive catalog of functional characterization assays in ENCODE, it is still challenging to conveniently integrate its data into cancer research for three reasons. First, cancer is such a heterogeneous disease that it is necessary to use data from optimally matched cell lines. However, ENCODE is imperfect for such analysis. We observe there is only loosely matched tumor-normal pairs for some cancer types, and most cell lines may lack data from a certain experimental assays (Fig 1A). Therefore, it is necessary to create biologically relevant tumor-normal pairs and develop appropriate algorithms to learn from data with suboptimal matching. The second issue arises due to the heterogeneous nature of the raw data from various experimental assays. It requires rigorous de-duplication, unified processing, and proper normalization before accurate large-scale integration can be achieved. Lastly, the noncoding annotations in ENCODE, such as TF binding sites and enhancers, are provided as standalone regions in the genome and lack in linkage to protein-coding genes. Hence, direct mutation function interpretation still remains elusive.

In this paper, we address these issues described above to maximize the usage of ENCODE data as a resource for cancer research (Figure 1 B-C). In order to tackle the heterogeneity amongst data types, we made a comprehensive data matrix by normalizing raw signals of genomic features that severely confound the somatic mutagenesis process (details in supplementary file). The resultant data matrix can be immediately used for background mutation rate correction. On the annotation side, different from previous efforts using only histone modifications and chromatin accessibility \{cite chromHMM}, we directly combined large-scale Enhancer-Seq experiments with computational predictions to generate high quality enhancer lists in several cell lines (details in supplementary file). To link these enhancers to genes, we performed enhancer target prediction by synthesizing evidences from expression profiles and chromatin status (details in supplementary file), and further pruned it using Hi-C data for higher accuracy (details in supplementary file). With such high quality linkage, we are able to build up the extended genes by combining coding regions with key regulatory elements in enhancers and promoters for better functional interpretation (Fig1 B). In addition, we also explored

Jing Zhang 1/27/2017 10:14 AM
Comment [9]: We could refer to papars that mis-use ENCODE data, but I am trying to avoid finger pointing. Please advise

Jing Zhang 1/27/2017 10:47 AM
Comment [10]: May remove this effort to emphasize the enhancer and gene linkage and ext gene. To disc

Jing Zhang 1/27/2017 10:42 AM
Comment [11]: Need to confirm with Kevin Yip

the full spectrum of the binding profiles from ENCODE and set up high confidence gene regulatory networks for both transcription factor (TF) and RNA binding proteins (RBPs) (Fig 1C and Fig Sxx in supplementary file).

Multi-level data integration from ENCODE benefits variants recurrence analysis in cancer

One of the most powerful ways of identifying key elements and deleterious mutations in cancer is by recurrence analysis, which attempts to discern which regions in the genome are more heavily mutated than expected. There are two challenges associated with such analysis. The mutation process is severely confounded by both external genomic factors and local context effects, which will result in numerous false positives and negatives if uncorrected (details in supplementary file). In addition, traditional burden tests usually ignore the interplay among annotations categories and separately test on standalone regions. Consequently, they are sometimes unable to pick up distributed mutation signals from biologically relevant regions and constraint functional interpretation of the burdened regions.

As a contrast, here we integrated the ENODE resources at two levels for better recurrence analysis. We first precisely predict an accurate local BMR by regressing out the confounding effects from features collected above in a cancer specific way (see details in supplementary file). Different from methods using unmatched data \{cite MutsigCV}, our regression based approach demonstrates that matched data usually provides higher BMR prediction precision (Fig 2A, details in supplementary). For example, in CLL, the correlation between observed and predicted mutation counts over 1mb bins (ρ) using replication-timing signals from K562 increases from XX to XXX relative to that using data from HeLa-S3. Further, various functional characterization assays from ENCODE, despite their possibly high correlations in signal tracks, usually represent different biological mechanisms that affect the mutation genesis progress (details see supplementary file), so it is important to integrate these features to collaboratively predict BMR (Fig 1B). For example, ρ is only from xxx to xxx using matched replication timing, but increased to xxx to xxx by adding 1 PC from the remaining covariates. It progressively increased to the xxx to xxx by adding up PCs to the full model through forward selection (Fig 1B, details in supplementary). Such noticeable improvement in BMR estimation will significantly benefit the following burdening analysis.

Instead of separately testing standalone annotation categories, we employed the abovementioned extended genes as joint test units (details in supplementary). Such scheme allows accumulation of weak mutation signals distributed across multiple biologically relevant functional elements, which might be lost during individual tests (Fig. Sx in supplementary file). Analyses show that our scheme could effectively remove false positives and discover meaningful burdened regions (Fig 2C). For example, in the context of leukemia, our analysis identified well-known drivers (such as TP53 and ATM) and other genes (BCL6) that missed by coding region analysis. In addition, BCL6 demonstrates strong prognostic value by differentiating patient survivals (Fig. 2D), indicating that the extended gene should be used as an annotation set for recurrence analysis when biological relevance is desired.

Extensive rewiring events in several transcription factors in cancer

We formed the TF network into hierarchies by comparing the inbound and outbound edges to investigate the topology of TF regulation (Fig 1E, details in supplementary file). TFs in different level in this hierarchy reflect the degree to which they directly regulate expression of other TFs \{cite 25880651}. For example, TFs in the top layer have more outbound edges than inbound edges in TF-TF network, representing larger roles in regulating other TFs rather than being regulated (supplementary Fig. xx). In this representation, we can see two patterns readily emerge. The top-level TFs often more strongly influence the tumor/normal differential expression. For instance, the average Pearson correlation of TFs binding events tumor-normal expression fold changes increases from 0.125 in the bottom layer to 0.270 in the top layer (Table Sx). TFs at the bottom layer were more frequently associated with burdened binding sites in general, perhaps reflecting their increased resilience to cellular mutation (details in supplementary file, Table Sx).

The human regulatory network specifies the combinatorial control of gene expression states from various regulators and constitutes the wiring diagram for a cell. Edge gain and loss analysis in matched tumor-normal TF networks could help to identify cancer-associated dysregulation. Hence, we investigated such rewiring events in TF networks through multiple formulations (details see supplementary file). Specifically, out of the 69 common TFs in K562 and GM12878 from ENCODE, we removed the general TFs and focused on the rewiring analysis on the remaining 61 TFs. (details in supplementary). In the simple counting methods, we first ranked TFs according to their number of loss/gain edges (Fig3 A, details in supplementary file). For example, several oncogenes (such as MYC and NRF1) were the top gainer TFs. On the contrary, IKZF1, where its somatic mutations serve as a hallmark of high-risk acute lymphoblastic leukemia (ALL), is the top loser with up to xxx percent of edges lost in K562 (Fig 3A). On the contrary, some ubiquitously distributed TFs such as YY1 retained their regulatory linkages (as show in Fig 3A). We observed similar trend of TFs using distal, proximal and combined network. Besides, we further used more complicated mixed membership model to look more abstractly at the local gene communities to re-rank the TFs (details in supplementary), and similar pattern was found and the well-known oncogene MYC become the top gainer (Fig 3A). To study the consequence of network rewiring, we performed the survival analysis on xxx AML patients and also found IKZF1 as significantly associated with tumor progression (details in supplementary).

Upon further investigation, we aim to explore the contributing factors to rewiring in tumor/cancer pairs and check the degrees of direction mutational effect during this process. We found that the majority of rewiring events were due to chromatin status change rather than from motif loss or gain events due to mutations (Fig. 3A). For example, JUND is a top gainer in K562 (xxx gain and xx loss). We found that up to 30.5 and 58.1 percent of the gain/loss events are associated with at least 2-fold expression change, and xxx percent has huge chromatin changes. Among those edges, only xxx variants were found in 100 CLL samples and among these up to xxx motif gain/loss variants could potentially affect rewiring events. All these analysis indicates the limited role of direct motif changing effect from mutations during the transition from normal to cancer cells.

Integrating ENCODE data with patient expression profiles identifies key regulators in cancer

To maximize the usage of ENCODE data into various types of cancers, we extended our network analysis from strictly matched tumor-normal cell lines to more generalized networks by a regression based learning method called RABIT (details in supplementary file). We integrated thousands of patient expression profiles from multiple cohorts to systematically search for TFs and RBPs that drive tumor specific expression patterns (Table Sx). In particular, for each pair of regulator and cancer type we selected the best explanatory binding profile and estimated the fraction of patients with target genes differentially regulated (details see supplementary file). The overall trend for the discovered key TFs and RBPs were given in Fig. 4A. We found that the impact of regulators on tumor gene expression predicted by our integration is highly consistent with previous knowledge. For example, RABIT predicted the target genes of *MYC* to be significantly up regulated in numerous cancers (star in Fig Sx), consistent with the known role of *MYC* as an oncogenic TF. Besides capturing knowledge from previous studies, our analysis also predicted previously unidentified functions for regulators in cancer. For example, the predicted targets of RBP *SUB1* were significantly up regulated in many cancer types (Figure 3C). As another example of novel predictions in our integration analysis, the predicted targets of TF *ZNF687* were significantly up regulated in breast and prostate tumors (star in Supplementary Figure 2). Thus, the integration analysis between ENCODE and expression data has revealed many previously unidentified regulators with possible roles in driving the cancer specific expression patterns.

[JZ2MG: loregic to be here!]

The combinatorial regulation of many TFs jointly determines the ON and OFF states of all genes to maintain the correct biological processes of normal cells. The disruption of co-regulatory relationships of key elements in cancer cell lines will result in erroneous gene expression pattern. We quantified the co-association status of each TF and observed huge co-association changes in some of the key TFs when comparing the regulatory network of K562 and GM12878. For example, ZNFXXX is a suppressor TF that shows only marginal co-binding events in GM12878. However, it not only increases its binding sites from xxx to xxx in K562, but also up to xxx percent of its binding sites co-bind with other TFs. Such unique patterns of co-association in cancer cell lines indicate differential combinatorial code.

Jing Zhang 1/27/2017 3:18 PM

Comment [12]: May completely move to the supplementary unless vineet's experiment is back

Step-wise prioritization schemes pinpoint deleterious SNVs in cancer

Here we proposed a multi-resolution prioritization scheme to pinpoint key regulatory elements and single nucleotide variants (SNVs) that are important for carcinogenesis (workflow in Fig.5 A). We start from searching for key regulators, such as TF or RBPs, which are either massively rewired or drives tumor-normal differential expression. Then we prioritize functional elements (such as enhancers and TF binding sites) governed by above key regulators through recurrence analysis. Lastly, we scrutinize each nucleotide therein by synthesizing features from annotation, conservation, and motif gain/loss effects to pinpoint the impactful ones for small-scale functional characterization.

Under this framework, we identified several active enhancers in the noncoding regions and validated their potential to initiate the transcription process using luciferase assay (details in supplementary file). In addition, we further selected key SNVs within these enhancers that are key for gene expression control (table Sx). Of 8 motif-disrupting SNVs we tested, we observed 6 variants with consistent up or down-regulated activity relative to the wild type (Fig. 5B and details in supplementary file). One particularly interesting region is chromosome 6, 13.5xxx (Fig. 5C). This enhancer is located in the noncoding region and both histone modification and DHS signals all indicate its active regulatory role (Fig. 5C). Our xxx based enhancer prediction method identified it as an enhancer and captured EnhancerSeq experiment further validated it (Fig. 5D). Hi-C data indicates this region is regulating an upstream gene XXX, which is xxxx (DL to fill in). xx out of the XXX Chip-Seq experiment showed significant binding events and the C to G mutation strong breaks the FOLS2 binding affinity (details in supplementary). It has been demonstrated in the luciferase that this mutation introduced xx fold expression reduction as compared to the wild type, indicating strong repressive effect on the enhance activities.

Jing Zhang 1/27/2017 3:45 PM
Comment [13]: Is this correct DL?

Conclusion

In this paper, we demonstrated the effectiveness of using ENCODE data to prioritize key regulatory elements/SNVs at different scales that are important for oncogenesis. Our scheme can be immediately applied to interpret the noncoding variants from large cohorts to pinpoint key elements for detailed functional characterization.