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

[JZ2MG]##### around 2800 word without abstract

Long Abstract

Short Abstract

Introduction

Recent developments of whole genome sequencing (WGS) and personal genomics have opened the opportunities to identify 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 catalogue human genome at multi-dimensional scale, the overwhelmingly large number of mutations are found in noncoding regions, where their functional impact remains difficult to characterize. Hence, it is important to decipher how these noncoding regions interact and how they are perturbed in cancerous cells to better dissect the somatic mutational landscape and provide personalized therapy for cancer patients. Since the inception, the ENCODE project has provided unprecedented opportunity to identify numerous noncoding cis-acting regulatory elements (CREs) through deep sequencing of entire human genome from comprehensive functional characterization assays. The ENCODE resources may potentially bridge the gap between the fast growing set of discovered noncoding variants with unknown functional impact and the limited number of well-known cancer genes for the cancer community.

[DL2MG: I thought there were a gap between previous paragraph and this. We may include previous examples of how non-coding elements/functional assays can help analyze oncogenesis.]

We here present an integrative framework to specifically tailor all the ENCODE resources for cancer analysis and prioritize CREs and SNVs related to tumorigenesis at multiple resolutions. In the large scale, we first set up a loosely matched tumor and normal gene regulation network in a cancer specific way to identify transcription factors (TF) that undergo dramatic changes during the transition from tumorous to normal cells. Patient survival analysis demonstrated that the top rewired TFs is closely associated with cancer prognoses. We then integrated ENCODE ChIP-seq and eCLIP data with patient-specific expression profiles from numerous sources to further prioritize the TFs or RNA binding proteins (RBPs) that drives tumor and normal differential expression. At the middle scale, we further consolidated highly heterogeneous genomic features that

confound the mutation process in cancer genomes to dissect the somatic mutational landscape and predict the true background mutation rate (BMR) under local sequence context. By integrating the most comprehensive non-coding annotations to coding genes, we can precisely quantify the recurrence level for each protein-coding gene as whole. Lastly, we scrutinized to the single base pair resolution to prioritize mutations that potentially disrupt regulatory events the most. Experimental validation at different scales demonstrated the effectiveness of our multi-resolution scheme to pinpoint the key regulatory elements and variants in various cancer types.

Comprehensive functional characterization data in ENCODE

[JZ2MG: logic: 1st para: table1 is what ENCODE has as raw data, 2nd is OUR effort to further process the ENCODE data and release to the cancer community]

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. Here we created a comprehensive list of raw datasets (Figure 1A) for use in further analysis. 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 (1-2% of genome) to over xxx percent of the noncoding annotated regions of the genome. This significant increase in data provided by ENCODE could benefit variant functional interpretation.

Despite the impressive coverage of ENCODE data, it is still challenging to integrate this data directly in cancer research. Cancer is a heterogeneous disease and functional characterization data usually changes from different cancer types, so it is important to use the optimally matched cell line in a cancer specific way. 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. Therefore, it become necessary to create biologically relevant tumor-normal pairs and create learning algorithms to get information from tissues/cell line with suboptimal matching. Another issue arises due to the heterogeneous nature of the raw data from various experimental assays, which requires rigorous de-duplication, unified processing, and normalization before accurate large-scale integration can be achieved. Finally, the CRE annotations, such as transcription factor binding sites and enhancers, from the ENCODE data are provided as standalone regions in the genome, lacking linkage to protein-coding genes, leaving the biological interpretation of mutations in CRE annotation regions still challenging.

[JZ2MG: logic of this para: WE further processed the raw signals and then the annotations. Although we use first use annotation and then data matrix in our following session]

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-D). In order to tackle the heterogeneity amongst data types, we summarized the raw signals of genomic features that confounds the somatic mutagenesis into a covariate matrix, which can be

immediately used for background mutation rate correction. On the annotation side, we pruned the computationally predicted enhancer list by adding large-scale Enhancer-seq results, and then provided accurate enhancer-gene linkage by integrating various experimental assay data, such as ChIP-seq, DNase-seq, Hi-C and ChIA-pet. Furthermore, we used uniformly processed peaks and signal tracks from xxx ChIP-seq experiments to build a TF-gene regulatory network, which includes both proximal and distal CREs. Across different cell lines, we also integrate noncoding CREs and protein-coding genes to generate a high confidence extended gene annotation, known as the epiGene. These publicly accessible annotations we created could greatly benefit cancer research due to their widespread coverage and accuracy, as well as due to their application in discerning key biological CREs and SNVs that play explanatory roles in tumorigenesis.

[DL2MG: We need to disc. I like the definition, but "epigene" is used as other meanings in some literature, the hereditary unit. http://www.biopolymers.org.ua/content/en/12/6/005/]

Extensive rewiring events in several transcription factors in cancer

[JZ2MG: where to put this para? Not suitable in data summary and also not quite suitable here]

To investigate the network topology of TF regulation, we first form the transcriptional regulatory network into hierarchy with TFs at different levels reflecting the degree to which they regulate 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 regulator TFs more strongly influence the tumor/normal differential expression than others. The average Pearson correlation of the binding events of TFs and gene expression changes was as high as 0.270 in the top layer, but it drops to 0.125 in the bottom layer. In contrast, the TFs at the bottom layer of the hierarchy were more frequently associated with burdened binding sites in general, perhaps reflecting their increased resilience to cellular mutation.

The human regulatory network specifies the combinatorial control of gene expression states from various regulatory elements constitutes the wiring diagram for a cell. Changes in the regulatory network during the transition from normal to tumor cells could help to decipher to discover the deregulation in cancer. Hence, we investigated the gain and loss of TF in cis-regulatory regions, the so-called rewiring events, in matched tumor and normal networks through multiple formulations. In the simple counting methods, we first ranked 61 TFs according to their number of loss/gain edges in the network of K562 and GM12878 (Fig3 A). For example, several oncogenes, such as RCOR1, REST, and ZBTB33, were among the top gainer TFs. Transcription factor IKZF1, whole mutant form serves as a hallmark of high-risk acute lymphoblastic leukemia (ALL), loses up to xxx percent of edges during the transition from tumor to normal cells. On the contrary, some non-specific cell type TFs such as RAD21 and YY1 retained their regulatory linkages (as show in Fig3 X). We observed similar trend of TFs using distal, proximal and combined network. Besides, we further used a mixed membership model to look more abstractly at the local neighborhood of all the

connections to re-rank the TFs, and similar pattern was found and the well-known oncogene MYC become the top gainer and IKZF1 as the top loser. To show the consequence of network rewiring, we analyzed the survival analysis showed that the highly rewired TF IKZF1 is significantly associated with tumor progression

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. For example, JUND is a top rewiring TF that gained a large number of targets in K562. 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 sample and among these up to xxx motif gain/loss variants could potentially affect rewiring events. All these analysis indicates the limited role of mutational effect during the transition from normal to cancer cells.

Integrating ENCODE data and patient expression data helps to identify key CREs

In order to systematically search for TFs and RNA binding proteins (RBP) that drive tumor specific expression patterns, we utilized a computational framework RABIT, developed to integrate cancer genomics data with regulatory profiling data[1]. We first collected 762 ChIP-Seq profiles from ENCODE representing 445 TFs, and 159 eCLIP profiles representing 112 RBPs. For a given TF ChIP-Ssq profile, candidate target genes are identified by weighting the number of binding sites by their distance to the transcription start site of genes. For a RBP eCLIP profile, we counted the number of binding sites on mRNA 3' UTR region to identify candidate target genes. Then, between each pair of regulator and cancer type, RABIT estimates the fraction of patients with target genes differentially regulated. For example, in liver and lung cancer, the target genes of RBP *SUB1* are significantly up regulated in most tumors. In contrast, the targets of RBP *RBFOX2* are significantly down regulated for most brain tumors (Figure 3B).

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 Supplementary Figure S2), consistent with the known role of *MYC* as an oncogenic TF[2]. 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). *SUB1* was previously considered as a TF[3], however the ENCODE eCLIP experiments have pulled down many *SUB1* peaks over gene 3'UTR regions (Supplementary Figure 1A and B), and these targets are predicted to be up regulated through the RABIT integration analysis. 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 TCGA 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.

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

One of the most powerful ways of identifying to identify key elements and deleterious mutations in cancer is by employing 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. In addition, traditional burden tests ignore the linkage among different noncoding annotations and simply apply burden test on individual annotation categories. Hence it is sometimes difficult to interpret the function of the burdened regions.

As a contrast, here we integrated the ENODE resources at two levels for better recurrence analysis. We first normalized and summarized data from XXX experiments in XX cell lines in ENCODE into a covariate matrix to precisely predict the local BMR through regression in a cancer specific way. Different from other methods that use the same data for all cancer types, our result indicates that matched data usually provides better BMR prediction. For example, in CLL, using Repli-seq signals from K562 increases the correlation of predicted vs observed mutation counts over 1mb bins from XX to XXX relative to using data from HeLa-S3 cell lines (XX to XXX in HeLa). In addition, despite the possibility of high inter-correlation, various functional characterization assays usually represent different biological mechanisms of mutation genesis progress, so it is important to integrate these features to collaboratively predict BMR. For example, the correlation among expected and observed mutation counts per 1mb bins is only from xxx to xxx using one replication timing, but increased to 0.88 to 0.95 by adding other feature in various types of cancers, which will significantly benefit the following burdening analysis.

Second, instead of testing noncoding annotation categories separately, we proposed an epiGene concept. We deeply integrated the ENCODE noncoding annotations and provided their high confidence gene linkage by integrating evidence for various experiment assay, such as ChIP-seq, Hi-C, and ChIA-PET. It incorporates both protein-coding exons and noncoding CREs for a gene as the burden test unit. Recurrence analysis performed on these novel epigene regions can reveal biological relevance when discovered to be heavily burdened. Another benefit of this is to amplify mutation signals that may potentially be lost in individual regulatory elements. Because the epigene can

consist of multiple discrete regions, a joint burden test is employed, allowing for better signal detection. Notably, in CLL the recurrence analysis performed using the epigene annotations allows for detection of novel gene candidates that are not found in recurrence analysis performed only on TSS or CDS regions. Among these candidates, BCL6 is identified using the epigene annotation analysis, but not with methods just using TSS or CDS annotations. In addition, BCL6 demonstrates strong prognostic value (patient survival), indicating that the epigene should be used as an annotation set for recurrence analysis when biological relevance is desired.

Step-wise prioritization schemes pinpoint deleterious SNVs in cancer

Here we proposed a multi-resolution prioritization scheme to pinpoint from the key CREs to SNVs that are important for tumor genesis (flowchart shown in Fig.5 A). We first search at a larger scale for key CREs, such as TF or RBP that are massively rewired or drives tumor/normal differential expression. Then we investigate functional elements across the genome regulated by the prioritized CREs through burden analysis. At last, we zoomed into sequences of the prioritized functional elements by utilizing comparative genomic features like conservation scores and motif gain/loss events to pinpoint the impactful SNVs for functional characterization of cancer.

Under this framework, we identified several enhancers in the noncoding regions and validated their potential to initiate the transcription process using luciferase assay. In addition, we further selected key SNVs within the functional cis-regulatory elements that are key for gene expression control. Of 8 motif-disrupting SNVs we tested, we observed 6 variants that were consistently up or down-regulated activity relative to the wildtype. One particularly interesting region is chromosome 6, 13.5xxx. The enhancer region nearby is in the intergenic region and has been predicted as strong enhancers both in normal (HMEC) and tumor cells (MCF-7) in breast. It has been shown to be regulating an upstream oncogene SGK1, which is key to the tumor genesis in breast cancer. The SNV we selected in this region has strong motif breaking effect for a series of TFs such as xxx, and we observed various TF binding sites overlapping it.

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.