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

Abstract

We understand the impact of somatic mutations well in a very limited number of cancer genes; in contrast, the overwhelming number of mutations in cancer genomes occur in non-coding regions. The new release of the ENCODE data allow us to bridge these two facts. First, the new ENCODE data enables precise tissue-matched genome-wide background mutation rate calibration in a variety of tumors by separating the effect of well-known confounders, such as replication timing and chromatin status. Furthermore, by integrating large scale ChIP-seq, DNase-seq, Enhancer-seq, Hi-C, and ChIA-PET data from ENCODE, we are able to define with high confidence distal and proximal regulatory elements and their linkages to annotated genes. This enables us to create extended gene definitions, and we are able to show these are more sensitive than coding regions in terms of burdening analysis. In particular in leukemia, in addition to well-known drivers such as TP53 and ATM, it allows us to pick up other key genes such as BCL6, which can then be associated with patient prognosis. Second, we integrated the ENCODE data to build up a high confidence TF-gene regulatory network. This enabled us to identify highly rewired (i.e. target changing) TFs, such as NRF1 and MYC by comparing tumor and normal samples. By integrating large-scale chromatin features, we demonstrated that such massive rewiring events between tumor and normal cell lines are mainly attributable to the chromatin structure changes instead of direct mutational effect. Furthermore, we also found that TFs with more mutationally burdened binding sites (e.g., EZH2 and NR2C2) tend to be located at the bottom hierarchy of the TF regulation network. Third, using the ENCODE regulatory network, we developed integrative scoring workflow to prioritize key elements (and mutations in them) according to their role in cancer and then validated these in small-scale studies. In particular, we prioritized ZNF687 as a key TF for breast cancer and SUB1 as a key RNA binding protein for liver and lung cancer and validated them through siRNA knockdown experiments. Finally, we identified key enhancers and mutations in them in breast cancer and then validated their functional effect through luciferase assays.

Introduction

[JZ: why cancer community need ENCODE data]

Recent developments of whole genome sequencing (WGS) and personal genomics have provided unprecedented opportunities to identify deleterious mutations that are important for tumor genesis, which in turn enable development of targeted therapies in clinical studies. However, although thousands of genomes' WGS data were provided through the collaborative effort of many consortia, the overwhelming number of mutations occur in noncoding regions, where functional impact remains difficult to characterize. Deciphering these noncoding regions interactions, and how they are perturbed in cancerous cells, are key to understanding cancer.

[JZ: challenges of directly using ENCODE data for the cancer community]

Since the inception of ENCODE project, deep sequencing of entire human genome allowed us to identify many noncoding regulatory regions and link these regions to the better understood coding regions to uncover the underlying biological mechanisms. The ENCODE resources may potentially bridge the gap in understanding between the fast growing set of discovered noncoding variants with unknown role, and the limited number of cancer genes with better interpretation for the cancer community. However, it is still challenging to directly incorporate the ENCODE data in an effective way for several reasons. First, while ENCODE provides comprehensive experiments focusing on various regulatory processes of the whole genome, the corresponding data sets are usually heterogeneous and require different levels of integration to better serve the cancer community. Second, due to the heterogeneous nature of various cancer types, it is important to harvest data from most relevant cell line when evaluating the variant effect in different cancer types. However, tissue matching is still a challenging problem. Lastly, none of the available cancer genomics data is complete in any cancer cell line. Hence, maximizing the utility of ENCODE data, and learning from other noncancerous tissue types, is an important topic. ,,,,,,manyoncancercell

[JZ: what we have done? Three layers of prioritization.]

We here presented an integrative framework to specifically tailor all ENCODE resources for cancer analysis. First, we integrated the comprehensive set of ENCODE data to better analyze recurrence events for cancer genomes. We 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) at local context. We also integrated the most comprehensive noncoding annotations and precisely linked them to well-known coding genes to better quantify the recurrence level for each protein coding gene. Second, we set up a loosely matched tumor and normal gene regulation network for key regulatory elements that undergo dramatic changes during the transition from tumor to normal cells. Additionally, we aggregated numerous sources of expression data to further prioritize the key elements that driver tumor and normal differential expression. Lastly, we scrutinized the single nucleotide variation (SNV) effect and prioritize those that potentially affect regulatory

events the most. Finally, experimental validation at different scales demonstrated the effectiveness of our scheme to pinpoint the key elements and variants in various cancer types.

Data summary

[JZ: how to link this section???]
[DL: emphasize how difficult to integrate missing/heterogeneous/incomplete tumor-normal data]

ENCODE includes extensive functional genomics data for cancer cell lines XXXXX. They include the most comprehensive sets of functional annotations of the human genome to date, from transcription level to chromatin and nuclear organization level, Despite this impressive coverage, the ENCODE resources are still far from perfect. First, some tumor-normal paired tissues might be only loosely connected to each other. For example, K562 can be loosely paired with GM12878 for CML, HepG2 to liver for liver cancer, A549 to IMR-90 for lung cancer, and MCF-7 to MCF-10A for breast adenocarcinoma. Second, no cancer type has the complete data set. There are lots of missing data from a certain experiment assay for some cancer type. In summary, we have a lot of issues in these cell lines and we further synthesized the most comprehensive dataset to our best and places resource list in Figure 1A.

Recurrence analysis

Recurrence analysis, which properly looks for regions mutated more frequently than expected, is one of the most powerful ways to identify key elements and deleterious mutations for cancer. One of the tricky parts of such analysis is that the mutation process is severely confounded by both external genomic factors and local context effects. As a result, the underlying background mutation rate across different regions over the genome could change up to several orders of magnitude even within one sample. Hence it is necessary to carefully calibrate such background mutation rate (BMR) to rigorously control the false positive and negative rate during the recurrence analysis.

[JZ: matched tissue -> use many features joint estimation -> local context effect]

Numerous references mention the importance of using matched tissue information to predict BMR. But state-of-the-art burdening analysis tools usually adopts a very limited number of features from seemingly distant cell lines, which largely limits the BMR estimation precision. The large cohort of functional genomics data in more than XXX tissue/cell lines in ENCODE provided us a chance to build up an integrative model to better dissect the somatic mutational landscape. Specifically, we collected the most comprehensive features from ENCODE and

processed these heterogeneous data into a covariate matrix to predict the local mutation rate with high precision through regression. We first demonstrated the advantages using matched data from ENCODE as compared with non-matched data. 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 cell lines. In addition, despite the possibility of high inter-correlation, various functional characterization assays usually uncover different biological mechanisms of mutation genesis progress, so it is important to integrate these features to collaboratively predict BMR for better precision. Specifically, the correlation among expected and observed mutation counts per 1mb bins ranges from 0.88 to 0.95 in various types of cancers. It is noticeably higher than those from a single feature such as replication timing and significantly benefit the following burdening analysis. In addition, the mutational frequency varies substantially due to the confounding effects of the local genome context. The BMR could range from xxx to xxx across different categories of local effect which further complicates the mutation burdening analysis. Our model takes these local context effects into consideration to better decompose the mutational signature effect. ,,,,butdoesencodedata It allows the contribution of different genomic features to vary across different mutational categories, and provides a context-specific mutation rate–even for regions with similar genomic features–to allow more accurate burden analysis.

In addition, previous methods are mainly focused on investigating various regulatory elements separately, which does not only limit the statistical power of burdening gene detection but also impair our interpretation of the discoveries. As a contrast, we proposed the extended gene definition to perform burdening analysis jointly on extensive cis-regulatory elements (CRE) together with the coding regions. Our model enables us to pickup weak mutation burden signals from individual CREs/coding exons and jointly evaluate the mutation burden for improved statistical power and better functional interpretation. In particular, we customize the most comprehensive noncoding annotations from ENCODE and link them to the well characterized protein coding genes to defined the extended gene region with high confidence. Burdening analysis on the extended gene regions demonstrates that our model is more sensitive to pick up reasonable burdened genes. For example, in CLL the extended gene analysis not only detects almost all genes burdened by either CDS or TSS regions, but also pinpoint other gene candidates with burdening on key regulatory regions. Among them, BCL6, which is missed using either TSS and CDS analysis, is identified as burdened in our method and its expression in CLL has been demonstrated to be significantly associated with patient survival. We also performed burdening analysis in breast and liver cancers and burdened regions were given in Fig 2.

Network analysis

[JZ: Should have rabbit here a one para to serve as the macro prioritization? Strongly suggest to put rabbit here],,,,,later

Logic: how to build -> global view of the network (hirNet) -> regulation activity changes -> co-association changes

[JZ: buildup gene expression regulation network]

The human regulatory network specifies the combinatorial control of gene expression states [DL: i.e. switch] from various regulatory elements, and constitute the wiring diagram for a cell. To examine the principles of the tumor transcriptional regulatory network, and decipher the consequences of rewiring events during the transition from normal to tumor cells, we integrated xxx transcription-related factors in over xxx distinct experiments and xxx cell lines for different cancer types to set up regulatory network to study the combinatorial and co-association relationships of transcription factors. Our regulatory network incorporates both distal and proximal interactions among TFs and genes. [JZ: put how to build up network details into supplementary?]

[JZ: hierarchy of the network]

[DL: TF as master regulator of other genes]

To investigate the network topology of TF regulation, we first clustered the TF-TF regulatory network into different layers based on their regulatory hierarchy. We found unique properties of the TF in each layer. For example, we found the general trend is that 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. [JZ: to check the middle layer co-association of TFs?]

[JZ: evaluate the TF classification according to rewiring events]

Regulatory network rewiring among tumor and normal cells suggest changes in control of gene expression status, which could result in massive gain or loss functions during the cell cycle. Here, we carefully defined edge loss and gain events by comparing the regulatory network in loosely matched tumor and normal for different cancers. Across all tumor types, we observed frequent rewiring events relative to each reference (normal) state. To assess the regulation potential of different TFs, we quantified their differential binding events in the network as a regulatory score and classified the TFs into three major groups: the gain, loss, and common

group. For example, several oncogenes, such as RCOR1, REST, and ZBTB33, were among the top TFs that gained massive binding events in promoter and enhancer regions; Some other TFs, such as the tumor suppressor HDGF, lost 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 CTCF and MAZ [DL: this MYC associated gene, RAD21 or YY1 are good alternatives] maintained most common edges in the network of K562 and GM12878, showing less differential regulation changes in these two cell lines. We propose to prioritize the TFs that showed huge rewiring events in the regulatory network.

Upon further investigation, we found that the majority of rewiring events were due to chromatin status change rather than direct mutational effect from motif loss or gain events. For example, JUND is a top rewiring TF that gained a huge 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 is 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 minimum [DL: indirect?] role of mutational effect during the transition from normal to cancer cells.
[JZ: evaluate the TF classification according to rewiring events]

While it may be rare to have mutations that directly affect TF binding sites, we hypothesized that mutations could have an indirect effect on key regulators of cancer. To further assess the mutational effects on regulatory element rewiring and selection bias, we focused on K562 and GM12878 pair and compared the pool of real CLL mutations to simulated sets of randomized mutations of the same size. Relative to random mutations, CLL mutations were more likely to cause motif loss in CEBPG, IRF1, MAX, and NR2F1. In contrast, the real mutations were found to increase the likelihood of TF binding in JUND, MAFF, MAFG, and NRF1.,,,,,todisc

[JZ: Prioritize the TFs with sharp co-association changes]
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 indicates differential combinatorial code.

Validation results

[JZ2DL: need a little bit of set up in the 1st para]
[DL: we should emphasis validation results implication on cancer],,,,,describelflow

Here we proposed a multi-level prioritization scheme to pinpoint the key CRE/SNVs that are important for tumor genesis. At the macro layer, through network rewiring analysis and expression aggregation, we proposed to select the key CREs, such as transcription factor and RNA binding proteins, that are either experienced dramatic change in tumor/normal cells or significantly drives tumor/normal expression. Then among the many functional regions regulated by these key CREs, we use mutation-burdening analysis to find those important ones with more mutations for prioritization. At last, we use other features like conservation score, motif gain/loss analysis to pinpoint the SNVs for small-scale functional characterization.

We have so far integrated extensive ENCODE annotations to define key regulatory elements to impactful noncoding SNVs in these regions based on our multi-level prioritization scheme. To asses the performance, we selected several examples at different scales and used various experimental assays to validate our predictions. At macro-level, we identified key transcriptional regulators (TFs) that drive tumor-normal differential expression. Specifically, we predicted ZNF687 and SUB1 as the most impactful regulators in MCF-7 and both HepG2 and A549, respectively, and we validated their significance using RNAi-based knockdown experiments. At micro-level, we validated 10 motif-breaking noncoding SNVs in key regulatory regions of MCF-7 using luciferase assay.

First, shRNA RNA-seq experiments were used to evaluate the gene expression level change before and after knocking down key transcriptional or RNA-level regulators. Specifically, the TF ZNF678 was discovered to significantly drive the tumor and cancer differential expression in the majority of breast cancer samples (figure xxx). After its knockdown, we discovered that its target genes were remarkably down-expressed compared to the non-target genes (p=xxxx for two sided t-test). Similarly, we found the RNA-binding protein SUB1 to significantly upregulate various target genes' expression in both lung and liver cancers. siRNA knockdown RNA-seq experiments also validated its regulatory role [DL: we need more details] (Figure 5 A). In addition, we found that the activity level of SUB1 is closely associated with patient survival data, further indicating its prognostic role in liver and lung cancers.

Second, we also use middle-scale assays to validate the functionality of regulatory elements. For example, after combining various chromatin status data, we used a match-filter based cis-regulatory element prediction method to find the key noncoding regions, and used a

luciferase assay to validate their potential to initiate the transcription process. Out of the nine predictions, a decent amount of expression has been observed, demonstrating the effectiveness of our method.

In addition, we further selected key SNVs within the functional cis-regulatory elements that are key for gene expression control. In order to evaluate the effect of mutation on regulatory region, we used luciferase reporter assay to quantify the activity of cis-RE containing motif-breaking mutation relative to wildtype in MCF-7. Of 8 motif-disrupting SNVs we tested, we observed 6 variants that were consistently up or down-regulated activity relative to the wild type. This results prove two points that the cis-regulatory region we tested are highly functional and the single-base nucleotide change that we selected can completely alter the effect of the regulatory region. We further characterized the validated regulatory regions by predicting target genes using both computational methods (ref. Kevin's Yip's enhancer target prediction) and incorporating nuclear organization and 3D chromatin architecture using Hi-C and ChIA-PET (ref. needed). We investigated each of the selected variants in detail (supplementary figure xxx - xxx). 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 a upstream oncogene SGK1, which is key to the tumorigenesis 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. (need biology of SGK1, actually upregulation of SGK1 is good for tumor growth? Opposite to our point?, https://www.karger.com/Article/FullText/374008).

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 cancer genesis. Our scheme can be immediately applied to interpret the noncoding variants from large cohorts, and pinpoint key elements for detailed functional characterization.