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

Long Abstract

[JZ2DL&MG: this is 450+ words, too long for an abstract!!!, I have a short version in the following session]

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 allows us to bridge these two facts. First, multiple layer of ENCODE data integration benefits somatic mutation burden analysis. On the raw signal level, we normalized raw signal tracks from comprehensive experiments to perform a genome-wide background mutation rate (BMR) calibration in a variety of tumors to separate the effects of well-known confounders, such as replication timing and chromatin status; then, on the annotation level, we create extended gene definitions, which link the ENCODE non-coding cis-acting regulatory elements (CREs) to annotated genes as a whole burden test unit after integrating large scale ChIP-seq, DNase-seq, Enhancer-seq, Hi-C, and ChIA-PET data. Our analyses on various cancer types demonstrate that they are more sensitive than coding regions in terms of sensible hyper-mutated regions discovery. In particular in leukemia, in addition to well-known drivers such as TP53 and ATM, it also picks 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 rigorously 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 performed hierarchical analysis on TF-TF networks and found that TFs on the top hierarchy are more associated with tumor-normal differential expression and TFs at the bottom hierarchy (e.g., EZH2 and NR2C2) are usually with more mutationally burdened binding sites. Third, using the ENCODE regulatory network, we developed a multi-scale scoring workflow to prioritize key CREs/SNVs according to their role in cancer and then validated these through different experimental assays. In particular, at the macro layer, 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; at the middle layer, we identified several key enhancers and validated their effect on gene

Formatted: Highlight
Formatted: Normal1, Left
Deleted: allows
Deleted: the new ENCODE data enables
Deleted: precise
Deleted: preform
Deleted: tissue-matched
Deleted: genome-wide
Deleted: by
Deleted: separating
Deleted: .
Deleted: Furthermore
Comment [JZ1]: I strongly suggest we give extended gene another name, like gene complex?
Deleted: ,
Deleted: by
Deleted: from ENCODE, we are able to define with high confidence distal and proximal regulatory elements and their linkages to annotated genes
Deleted: This enables us to create extended gene definitions, and we are able to
Deleted: show
Deleted: these
Deleted: burdening analysis
Deleted: allows us to
Comment [LD2]: We could say we identified highly rewired regulators (TFs) in “cell-type specific manner” and demonstrated their significances using patient survival analysis.
Deleted: y
Deleted: we also
Deleted: layer
Deleted: /
Deleted: layer
Deleted: sites (e.g., EZH2 and NR2C2) tend to be located at the bottom hierarchy of the TF regulation network
Deleted: integrative
Deleted: elements (and mutations in them)
Deleted: in small-scale
Deleted: studies
Deleted: master

expression through luciferase assays; and finally picked up the most important SNV in them in to check their mutational effect on these enhancers through luciferase assays. Our work demonstrates careful integration of the ENCODE resources offers unprecedented opportunity to accurately characterize oncogenic regulation and serves as a powerful tool to prioritize cell-type specific cancer variants.

Deleted: .
Deleted: Finally, we
Deleted: identified key enhancers and mutations
Deleted: breast cancer and then validated
Deleted: functional

Short Abstract

We understand the somatic mutations well in a very limited number of cancer genes; in contrast, the majority of mutations in cancer genomes occur in non-coding regions. The new release of the ENCODE data bridges these two facts. First, ENCODE data integration from multiple levels benefits mutation burden analysis. On the raw signal level, we normalized data from comprehensive experiments to precisely calibrate background mutation rate (BMR) calibration in a variety of tumors; on the annotation level, we created extended gene definitions to link the cis-acting regulatory elements (CREs) to genes. Results demonstrate that our scheme can find sensible hyper-mutated regions. For example in leukemia, in addition to well-known drivers such as TP53 and ATM, it also discovered BCL6, which is proven to be associated with disease prognosis. Second, we build up a high confidence TF-gene regulatory network in a cancer specific way and identified highly rewired TFs, such as NRF1 and MYC. We also demonstrated that such massive rewiring events are mainly attributable to chromatin structure changes instead of direct mutational effect. Furthermore, we performed hierarchy analysis on TF-TF networks and found that TFs on the top layer are more associated with tumor/normal differential expression and TFs at the bottom layer (e.g., EZH2 and NR2C2) are usually with more mutationally burdened binding sites. Third, using the ENCODE regulatory network, we developed a multi-resolution scoring workflow to prioritize key CREs/SNVs. At the macro layer, we prioritized ZNF687 for breast cancer and SUB1 for liver and lung cancer and validated them through siRNA knockdown experiments; at the middle layer, we identified several key enhancers and validated them through luciferase assays; and we finally zoom in on the single nucleotide resolution to prioritize SNVs with validations.

Comment [JZ3]: I strongly suggest we give extended gene another name, like gene complex?

Deleted: master

Introduction

[JZ: why cancer community need ENCODE data: done!]

Recent developments of whole genome sequencing (WGS) and personal genomics have opened the 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

Deleted: provided unprecedented
Deleted:

of genomes' WGS data were provided through the collaborative effort of many consortia, the overwhelming number of mutations occur 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 of ENCODE project, deep sequencing of entire human genome from comprehensive functional characterization assays allowed us to identify numerous noncoding cis-acting regulatory elements (CREs). 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.

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

However, it is still challenging to directly incorporate the ENCODE data in an effective way for several reasons. First, cancer is such a highly heterogeneous disease that the variant impact evaluation should be carried in a tissue-specific way. However, despite the impressive coverage, the ENCODE resources are still far from perfect. Some tumor-normal paired tissues might be only loosely connected to each other, and no cancer type has the complete data set. There are lots of missing data from a certain experiment assay for some cancer type. Hence, it is key to harvest data from most relevant cell lines or even learning from other noncancerous tissue types; second, the various genome-wide signal tracks from different ENCODE experiments are in nature highly heterogeneous and it requires rigorous de-duplication and normalization to better serve the cancer community. Lastly, no all CREs annotations provided by ENCODE, for example the transcription factor binding sites (TFBSs) and enhancers, are directly linked to a specific gene, resulting in difficulty in immediate SNV effect interpretation therein.

[JZ: what we have done? Three layers of prioritization; Done!]

We here present an integrative framework to specifically tailor all ENCODE resources for cancer analysis and prioritize CREs and SNVs at multiple resolutions. 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 BMR under local context. We also integrated the most comprehensive noncoding annotations and precisely linked them to well-known coding genes as whole unit to better quantify the recurrence level for each protein-coding gene. Second, we set up a loosely matched tumor and normal gene regulation network in a cancer specific way to identify 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 to the single nucleotide variation resolution and prioritized those that potentially affect regulatory events the most. Finally, experimental validation at different scales demonstrated the effectiveness of our multi-resolution scheme to pinpoint the key elements and variants in various cancer types.

Data summary

Deleted: Deciphering
Deleted: ions,
Deleted: , are key to understanding cancer
Deleted: .
Moved (insertion) [1]
Deleted: many
Deleted: regions
Deleted:
Deleted: and link these regions to the better understood coding regions to uncover the underlying biological mechanisms
Deleted: in understanding
Deleted: role,
Deleted: with better interpretation
Moved up [1]: 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.
Deleted:
Deleted: First,
Deleted: while ENCODE provides comprehensive experiments focusing on various regulatory processes of the whole genome,
Deleted: corresponding
Deleted: data sets
Deleted: usually
Deleted: different levels of integration
Deleted: 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 car ... [1]
Comment [LD4]: We may want to emphasis th ... [2]
Deleted: .
Deleted: ed
Deleted: background mutation rate (
Deleted:)
Deleted: at
Deleted: protein
Deleted: for key
Deleted: (SNV)
Deleted: effect

[JZ2DL&MG: this data summary session has so much overlap with the introduction session. Need more polish to clarify these difference!]

[JZ: how to link this section???]

[DL: emphasize how difficult to integrate missing/heterogeneous/incomplete tumor-normal data]

ENCODE includes the most comprehensive sets of functional annotations of the human genome to date, from transcription level to chromatin and nuclear organization level. Up to xxx percent of the cell lines in ENCODE are actually cancerous cells, leaving it as valuable resource for the cancer research. Here we synthesized the most comprehensive raw dataset to our best and places resource list in Figure 1A. Besides, the noncoding annotations from ENCODE significantly enlarges our understanding of mutations from 1~2 percent of the whole genome in the coding regions to up to xx percent of the annotated regions, strongly benefits variant functional interpretation.

In this paper, we also customized two other resources of ENCODE annotation to facilitate cancer research. First, we proposed the extended gene concept by linking various noncoding CREs to genes with high confidence in various cell lines. In particular, we predicted high-quality enhancer targets using both computational methods and incorporating nuclear organization and 3D chromatin architecture using Hi-C and ChIA-PET. Then the distal (enhancers) and proximal (TFBS in promoters) CREs were assigned to each protein coding genes as a full category of regulation elements. Second, we built up gene expression regulatory network for in a cancer-specific way to facilitate regulatory network analysis.

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

[JZ2DL&MG: shortened significantly, removed local context effect, we need to emphasize there is multi-level integration here!, I think it is more clean]

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. Two main difficulties in such analysis are: 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; second, traditional burden tests are performed on separate groups based on their regulatory categories, such as promoter or enhancer list, which completely ignores the interaction among different groups and impairs immediate functional interpretation of the results.

Formatted: Highlight

Formatted: Highlight

Deleted: includes extensive functional genomics data for cancer cell lines XXXXX. They

Comment [LD5]: [suggestion?] Because this is data summary section, we may want to emphasis that we have the best enhancers (CREs) to begin with, by integrating both EnhancerSeq experiments and match-filter based method. Then, based-on our best enhancer sets, we can talk about how we further predicted target genes using xxx methods.

Deleted:

Deleted:

Deleted: , 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.

Deleted: Recurrence

Formatted: Highlight

Formatted: Highlight

Formatted: Highlight

Deleted: One of the tricky parts of such analysis is that

Deleted: ly

As a contrast, here we integrated the ENODE resources at two levels for better recurrence analysis. We first normalized and summarized data from comprehensive experiments 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 extended gene 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. It allows us to pickup weak mutation signals from individual CREs/coding exons and jointly evaluates the mutation burden for improved statistical power and better functional interpretation. 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.

Tissue specific network analysis helps to pinpoint key CREs for cancer

The human regulatory network specifies the combinatorial control of gene expression states 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 seek for master regulators for multiple cancer types.

[JZ: Peng and Loregic part here!]

Deleted: 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

Comment [LD6]: [suggestion] maybe use actual number to emphasize, say we integrated xxx number of experiments across xxx cell-types, then reference to figure how using more data and matched cell-type can improve BMR prediction

Deleted:

Deleted: H

Deleted: i

Comment [LD7]: maybe we want to state in one summary sentence before this, saying "based on our burdening analysis, we identified xxx significant genes and among them xxx were found to be insignificant using CDS-based analysis. For example, BCL6 is highly recurrent gene in CLL cohort and has high prognostic value (fig. x)"

Deleted: ... [3]

Formatted: Font:(Default) Arial, 20 pt

Deleted: [JZ: Should have rabbit here a one para to serve as the macro prioritization? Strongly suggest to put rabbit here],,,,,later Our regulatory network incorporates both distal and proximal interactions among TFs and genes. [JZ: put how to build up network details into supplementary?]

Formatted: Font:(Default) Arial, 20 pt

Formatted: Font:(Default) Arial, 20 pt

Formatted: Font:(Default) Arial, 20 pt

Formatted: Font:(Default) Arial, 20 pt, Font color: Black

Deleted: t

Deleted: [JZ: Should have rabbit here a one para to serve prioritization? Strongly suggest to put rabbit here],,,,,later Our regulatory network incorporates both distal and proximal interactions among TFs and genes. [JZ: put how to build up network details into supplementary?]

Deleted: [DL: i.e. switch]

Deleted: set up regulatory network to study the combinatorial and co-association relationships of transcription factors

Deleted: Our regulatory network incorporates both distal proximal interactions among TFs and genes. [JZ: put how to build up network details into supplementary?]

Formatted: Font color: Black, Highlight

Deleted: hierarchy of the network

Formatted: Font:(Default) Times New Roman, 12 pt

Deleted: [DL: TF as master regulator of other genes]

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. 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.

Regulatory network rewiring between 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 investigated 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. There's a variety of ways of formulating this rewiring. First we counted the number of edges that remain the same to rank the TF with the raw gain/loss events and found that different TFs show rewiring disparate patterns. For example, several oncogenes, such as RCOR1, REST, and ZBTB33, were among the top gainer TFs; Some other TFs, such as the tumor related gene 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 RAD21 and YY1 maintained most common edges in the network of K562 and GM12878 (as show in Fig3 X). 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. Survival analysis showed that the highly rewired TFs are significantly associated with tumor progression [JZ: CC's data to come in!].

Upon further investigation, we aim to explore the reason of network 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 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 limited role of mutational effect during the transition from normal to cancer cells.~~

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

Deleted: clustered the TF-TF regulatory network into different layers based on their regulatory hierarchy.

Deleted: We found unique properties of the TF in each layer.

Deleted: For example, we found the general trend is that the t

Deleted: [JZ: to check the middle layer co-association of TFs?] ... [4]

Deleted: among

Deleted: defined

Comment [LD10]: After our discussion, we may change this to IKZF1. We can keep HDGF for now (HDGF is interesting too), but let's remind ourselves to update later

Formatted: Highlight

Deleted: 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.

Comment [LD11]: "Reason" maybe too strong statement. We can soften this tone by saying "contributing factors to rewiring" or "characteristic of network rewiring".

Deleted: direct mutational effect

Deleted: minimum [DL: indirect?]

Deleted: [JZ: evaluate the TF classification according to rewiring events] ... [5]

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

Here we proposed a multi-resolution prioritization scheme to pinpoint from the key CREs to SNVs that are important for tumorigenesis. At the large scale, we selected the key CREs, such as transcription factor and RNA binding proteins, that underwent either dramatic regulatory network rewiring and differential expression aggregation between tumor and normal states. At the middle level, we use mutation-burdening analysis to find those important ones with more mutations for prioritization. At the micro level, we zoomed into base pair resolution, utilizing comparative genomic features like conservation scores and transcription factor binding profiles to assess motif gain and loss, to pinpoint the impactful SNVs for functional characterization of cancer.

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 (Fig 5A, p=xxxx for two sided t-test). Similarly, we found the RNA-binding protein SUB1 to significantly up-regulate various target genes' expression in both lung and liver cancers. siRNA knockdown RNA-seq experiments also validated its regulatory role (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 identified several candidate enhancers that are based on computational predictions and targeted Enhancer-seq experiment in the noncoding regions and validated their potential to initiate the transcription process using luciferase assay. Of xxx candidate regions we identified as functional, 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. Of 8 motif-disrupting SNVs we tested, we observed 6 variants that were consistently up or down-regulated activity relative to the wild type. 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 tumorigenesis in breast cancer. The SNV we selected in this region has strong motif

Formatted ... [7]
Deleted: -level
Formatted ... [8]
Deleted: /
Formatted ... [9]
Deleted:
Formatted ... [10]
Deleted: At the macro layer
Deleted: First a
Formatted ... [11]
Deleted: through network rewiring analysis and ex ... [12]
Deleted: proposed to
Formatted ... [13]
Deleted: are
Formatted ... [14]
Deleted: experienced …ramatic regulatory networ ... [15]
Formatted ... [16]
Deleted: significantly drives tumor/normal…iffere ... [17]
Formatted ... [18]
Deleted: Then among the many functional regions ... [19]
Formatted ... [20]
Deleted: last
Formatted ... [21]
Deleted: …o base pair resolution, use ... [22]
Formatted ... [23]
Deleted: other
Formatted ... [24]
Deleted: , motif
Formatted ... [25]
Deleted:
Formatted ... [26]
Deleted: /
Formatted ... [27]
Deleted: analysis
Formatted ... [28]
Deleted: small-scale
Formatted ... [29]
Deleted: . ... [30]
Formatted ... [31]
Deleted: figure xxx…ig 5A, p=xxxx for two sided ... [32]
Formatted ... [33]
Deleted: [DL: we need more details]
Formatted ... [34]
Deleted: also use middle-scale assays to validate t ... [35]
Deleted: noncoding a match-filter based cis-regula ... [36]
Comment [LD13]: Is this part about MCF-7 ... [38]
Formatted ... [37]
Formatted ... [39]
Deleted: In order to evaluate the effect of mutation ... [40]
Deleted: a
Formatted ... [41]

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

Deleted: (need biology of SGK1, actually upregulation of SGK1 is good for tumor growth? Opposite to our point?, https://www.karger.com/Article/FullText/374008).

Formatted: Font color: Black

Formatted: Font:(Default) Times New Roman, 12 pt

Formatted: Font:(Default) Times New Roman, 12 pt

Page 3: [1] Deleted **Jing Zhang** **1/15/17 11:57:00 AM**

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

Page 3: [2] Commented **Lee, Donghoon** **1/15/17 3:25:00 PM**

We may want to emphasis that fact that some of these data have different length-scale, thus others may have difficult times to integrate various sources into one. (maybe between 1st and 2nd points here)

Page 5: [3] Deleted **Jing Zhang** **1/15/17 12:21:00 PM**

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

Page 6: [4] Deleted **Jing Zhang** **1/15/17 12:46:00 PM**

[JZ: to check the middle layer co-association of TFs?]

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

Page 6: [5] Deleted **Jing Zhang** **1/15/17 1:08:00 PM**

[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]

Page 7: [6] Deleted **Jing Zhang** **1/15/17 1:10:00 PM**

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

Page 7: [7] Formatted **Jing Zhang** **1/15/17 1:10:00 PM**

Font color: Black

Page 7: [8] Formatted **Jing Zhang** **1/15/17 1:10:00 PM**

Font color: Black

Page 7: [8] Formatted **Jing Zhang** **1/15/17 1:10:00 PM**

Font color: Black

Page 7: [9] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [9] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [10] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [11] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [12] Deleted	Lee, Donghoon	1/15/17 4:13:00 PM

through network rewiring analysis and expression aggregation,

Page 7: [13] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [14] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [15] Deleted	Lee, Donghoon	1/15/17 4:15:00 PM

experienced

Page 7: [15] Deleted	Lee, Donghoon	1/15/17 4:15:00 PM

experienced

Page 7: [16] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [16] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [17] Deleted	Lee, Donghoon	1/15/17 4:16:00 PM

significantly drives tumor/normal

Page 7: [17] Deleted	Lee, Donghoon	1/15/17 4:16:00 PM

significantly drives tumor/normal

Page 7: [18] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [19] Deleted	Lee, Donghoon	1/15/17 4:20:00 PM

Then among the many functional regions regulated by these key CREs

Page 7: [20] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [21] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [22] Deleted	Lee, Donghoon	1/15/17 4:21:00 PM

Page 7: [22] Deleted	Lee, Donghoon	1/15/17 4:21:00 PM

Page 7: [23] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [24] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [24] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [25] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [26] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [27] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [27] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [28] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [29] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [29] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font color: Black

Page 7: [30] Deleted	Jing Zhang	1/15/17 1:50:00 PM

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.

Page 7: [31] Formatted	Jing Zhang	1/15/17 1:10:00 PM

Font:(Default) Times New Roman, 12 pt

Page 7: [32] Deleted **Jing Zhang** **1/15/17 1:50:00 PM**

figure xxx

Page 7: [32] Deleted **Jing Zhang** **1/15/17 1:50:00 PM**

figure xxx

Page 7: [33] Formatted **Jing Zhang** **1/15/17 1:10:00 PM**

Font color: Black

Page 7: [34] Formatted **Jing Zhang** **1/15/17 1:10:00 PM**

Font:(Default) Times New Roman, 12 pt

Page 7: [35] Deleted **Jing Zhang** **1/15/17 1:51:00 PM**

also use middle-scale assays to validate the functionality of regulatory elements. For example, after combining various chromatin status data, we used

Page 7: [36] Deleted **Lee, Donghoon** **1/15/17 4:28:00 PM**

noncoding a match-filter based cis-regulatory element prediction method to find the key

Page 7: [36] Deleted **Lee, Donghoon** **1/15/17 4:28:00 PM**

noncoding a match-filter based cis-regulatory element prediction method to find the key

Page 7: [36] Deleted **Lee, Donghoon** **1/15/17 4:28:00 PM**

noncoding a match-filter based cis-regulatory element prediction method to find the key

Page 7: [36] Deleted **Lee, Donghoon** **1/15/17 4:28:00 PM**

noncoding a match-filter based cis-regulatory element prediction method to find the key

Page 7: [36] Deleted **Lee, Donghoon** **1/15/17 4:28:00 PM**

noncoding a match-filter based cis-regulatory element prediction method to find the key

Page 7: [36] Deleted **Lee, Donghoon** **1/15/17 4:28:00 PM**

noncoding a match-filter based cis-regulatory element prediction method to find the key

Page 7: [36] Deleted **Lee, Donghoon** **1/15/17 4:28:00 PM**

noncoding a match-filter based cis-regulatory element prediction method to find the key

Page 7: [37] Formatted **Jing Zhang** **1/15/17 1:10:00 PM**

Font color: Black

Page 7: [38] Commented **Lee, Donghoon** **1/15/17 4:33:00 PM**

Is this part about MCF-7 validation?

Page 7: [39] Formatted **Jing Zhang** **1/15/17 1:10:00 PM**

Font:(Default) Times New Roman, 12 pt

Page 7: [40] Deleted **Jing Zhang** **1/15/17 1:53:00 PM**

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.

Page 7: [40] Deleted **Jing Zhang** **1/15/17 1:53:00 PM**

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.

Page 7: [41] Formatted **Jing Zhang** **1/15/17 1:10:00 PM**

Font color: Black