Integrating ENCODE data to interpret regulatory changes in cancer

Abstract

#/*=== abstract section 345 words ===*/

Articles have a summary, separate from the main text, of up to 150 words, which does not have references, and does not contain numbers, abbreviations, acronyms or measurements unless essential.

In cancer, the impact of mutations in a limited number of genes is well understood; in contrast, the preponderance of mutations found in cancer genomes occur in non-coding regions. The new release of the ENCODE data allows us to bridge these two facts for a number of well-studied cancers of the blood, liver, lung, breast and cervix, for which a diverse number of genome-wide assays (e.g., ChIP-seq, DNase-seq, Enhancer-seq, Hi-C, and ChIA-PET) are applied to matched tumor-normal cell lines. First, the new data enabled precise, tissue-matched non-coding background mutation-rate calibration, taking into account the effect of confounders, such as replication timing. Furthermore, by integrating diverse ENCODE data, we were able to define high-confidence regulatory elements and their linkages to genes. This allowed us to create definitions of extended gene neighborhoods, which were more sensitive than just coding regions for recurrent-mutation, burden analysis. Using this approach, we identified additional key genes, beyond well-known drivers, associated with patient prognosis (e.g., *BCL6* in leukemia). Second, we integrated the ENCODE data to build a hierarchical regulatory network, including both transcription factors (TFs) and RNA-binding proteins. We found that more mutationally burdened TFs tended to be located at the bottom of the hierarchy (e.g., EZH2 and NR2C2), whereas those associated with the largest oncogenic gene-expression changes tended to be at the top. Furthermore, by comparing tumor and normal samples, the network enabled us to identify highly "rewired" (i.e. target changing) TFs, such as IKZF1 and MYC. Our results indicated that such rewiring events are mainly attributable to chromatin changes instead of direct mutational effects. Third, we developed a scoring workflow to prioritize key non-coding elements (and mutations in them) according to their roles in cancer and positions in the regulatory network 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 cancers and validated them through siRNA knockdown experiments. Finally, we identified key enhancers and mutations in them in breast cancer and then validated their functional effects through luciferase assays.

Articles have a summary, separate from the main text, of up to 150 words, which does not have references, and does not contain numbers, abbreviations, acronyms or measurements unless essential.

Articles are typically 3,000 words of text, beginning with up to 500 words of referenced text expanding on the background to the work (some overlap with the summary is acceptable), before proceeding to a concise, focused account of the findings, ending with one or two short paragraphs of discussion.

In total 2673 words,

#/*=== Introduction section 402 words ===*/

#/*=== Data section 407 words ===*/

#/*=== recurrence section 501 words ===*/

#/*=== Rewiring section 644 words ===*/

#/*=== expression section 250 (+111) words ===*/

#/*=== expression section 326 words ===*/

#/*=== expression section 143 words ===*/

Introduction

#/*=== Introduction section 402 words ===*/

Despite the discovery tens of thousands of mutations, only a very small fraction is readily interpretable in terms of their effects on known cancer-associated genes. A uncertainty lies in the degree to which these tens of thousands of variants contribute to cancer. Are they simply neutral passengers created as byproducts of the genomic dysregulation known to be intrinsic to cancer genomes? Alternatively, do key cancer-driving variants remain lurking among this newly discovered pool of mutations? Or perhaps these non-coding variants affect the regulation and expression of key known cancer genes? The new data release of ENCODE Consortium may bridge these knowledge gaps by providing accurate non-coding annotations and precisely linking these annotations to well-characterized protein coding genes.

The ENCODE data resource provides an annotation of the human genome. Annotations are assigned in a cell type-specific manner, and many of the cell lines used are cancerous. Admittedly, this data is imperfect in the context of cancer research: some cell lines are suboptimal for actual tumor samples, and matching data from these cell lines with that from appropriate normal samples is often challenging. However, because of the tremendous richness of assays available in this new ENCODE release, comparisons of tumor-like and normal-like cell lines provide an unprecedented and accurate window into the regulatory and chromatin-related changes associated with oncogensis.

Here, we endeavor to make the ENCODE resource as useful as possible for cancer research. We mainly focused on blood, breast, liver, lung, and cervix cancers where

ENCODE has super data richness in their corresponding cancer cell lines (details see supplementary file). We match these cell lines with known cancers to better integrate relevant expression profiles and somatic mutations from known cohorts. We then develop methods to integrate comprehensive ENCODE signal tracks to calibrate an accurate background mutation rate (BMR), which is then provided as a resource. This allows us to accurately find burdened regions in many cancers. We use the wealth of ENCODE assays to accurately determine non-coding elements in each cell line (enhancers in particular). It also enables us to delineate regulatory networks involving transcription factors (TFs) and, to a lesser extent, RNA-binding proteins (RBPs). We represent these regulatory networks in a variety of ways, including hierarchical models, wherein master regulators occupy the top of the hierarchy. For each regulator in the network, we then calculate a rewiring score that represents the degree to which a regulator differs between normal and cancerous cells.

[[MG(LONG): need to emphasize to expl. Assay richness]]

Data for comprehensive functional characterization in ENCODE

#/*=== Data section 407 words ===*/

[[MG(LONG): very repetitive to **1**. Made some cuts. Need to Shorten by 50% + mention that we focus on 5 cancers as models]]

Despite the comprehensive catalog of functional characterization assays in ENCODE, integrating its associated data into cancer research remains challenging for two main reasons. First, cancer is such a heterogeneous disease that it is necessary to use data from optimally-matched cell lines. ENCODE is imperfect for such analysis. We observe that there are only loosely matched tumor-normal pairs for some cancer types, and most cell lines lack data from certain experimental assays (Fig 1A). Therefore, it is necessary to create biologically relevant tumor-normal pairs, as well as to develop appropriate algorithms to learn from this sub optimally matched data. The second challenge arises as a result of the heterogeneous nature of the raw data from various experimental assays. The data must undergo de-duplication, unified processing, and proper normalization before accurate large-scale integration can be achieved.

Specifically, to tackle the heterogeneity amongst data types, we constructed a comprehensive data matrix by normalizing raw signals of genomic features that severely confound somatic mutagenic processes (see Supp. File/Section(?) X). In contrast to previous approaches to annotation (many of which use only histone modification and chromatin accessibility data \{cite chromHMM}), we proposed an ensemble method to accurately pinpoint active enhancers. It combines large-scale Enhancer-Seq experimental data with computational predictions based on pattern recognitions of histone marks (see Supp. File/Section(?) X). To link these enhancers to genes, we perform enhancer target predictions by integrating evidence from expression profiles and chromatin status (see Supp. File/Section(?) X), and we further prune these predictions using Hi-C data for greater accuracy (see Supp. File/Section(?) X). [[MG: we need more here on etg links, I'd say 2 sentences]]To achieve improved functional interpretation, we use these high-quality linkages to construct what we term "extended genes" – coding regions matched with key regulatory elements in enhancers and promoters (Fig1 B). In addition, we also explore the

Comment [JZ1]: May remove this effort to emphasize the enhancer and gene linkage and ext gene. To disc

full spectrum of the binding profiles in ENCODE data, and construct high-confidence gene regulatory networks for both transcription factor (TF) and RNA binding proteins (RBPs; Fig 1C and Fig X in Supp. File/Section(?) X).

Finally, for each of the main ENCODE cell lines, our publically-disseminated resource consists of a list of accurately determined enhancers, a list of burdened regions, the regulatory TF network, as well as the most rewired TFs in this regulatory network (see suppl.). Collectively, these resources allow us to prioritize a few key elements as being associated with oncogenesis, some of which have been validated through small-scale experiments (see table S1).

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

#/*=== recurrence section 501 words ===*/

One of the most powerful ways of identifying key elements and deleterious mutations in cancer is through recurrence analysis, which attempts to discern which regions of the genome are more heavily mutated than expected. There are two challenges associated with such analysis. First, the mutation process introduces confounding factors (in the form of both external genomic factors and local context effects), which can result in many false positives or negatives (see Supp. File/Section(?) X). Secondly, traditional burden tests often neglect the interplay among annotation categories, and they thus test regions separately. Consequently, these tests are sometimes unable to identify distributed mutation signals from biologically relevant regions, thereby limiting the functional interpretation of the burdened regions.

In contrast, we integrate the ENCODE resources at two levels for better recurrence analysis. First, we predict an accurate local BMR by regressing out the confounding effects of features in a cancer-specific manner (see Supp. File/Section(?) X). Specifically, we prepared a covariate matrix by normalizing 475 features from ENCODE to remove the external confounding effects to BMR. We then separated the whole genome into 64 categories according to the local 3 mers and run separate regression models to further eliminate the internal context effect. In contrast to methods that use unmatched data \{cite MutsigCV}, our regression-based approach demonstrates that matched data usually provides higher BMR prediction precision (Fig 2A, see also Supp. File/Section(?) X). For example, in breast cancer, the correlation between observed and predicted mutation counts over 1-megabase bins (ρ) using replication-timing signals (from MCF-7) increases from XX to XXX relative to that using data from HeLa-S3. Furthermore, despite the possibly high correlations in signal tracks, various functional characterization assays from ENCODE usually represent different biological mechanisms that affect mutagenic progresses (see Supp. File/Section(?) X). Thus, it is important to integrate these features to infer BMR (Fig 1B). For example, ρ only ranges from xxx-xxx using matched replication timing, but its range increases to xxx-xxx by adding 1 PC from the remaining covariates. It progressively increases to the xxx-xxx regime by adding PCs to the full model through forward selection (Fig 1B, see Supp. File/Section(?) X). Such noticeable improvements in BMR estimation significantly improve burden analyses (see below).

As oppose to separately testing standalone annotation categories, we employed our extended gene (detailed above) as joint test units (see Supp. File/Section(?) X). Such a

scheme allows for the accumulation of weak mutation signals distributed across multiple biologically relevant functional elements, which may otherwise be lost if evaluated under individual tests (Fig. Sx in Supp. section X). We demonstrate that our scheme can 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) as well as other genes (such as BCL6) that were missed by the analysis of coding regions. In addition, BCL6 demonstrates strong prognostic value with respect to patient survival (Fig. 2D), indicating that the extended gene should be used as an annotation set for recurrence analysis.

Extensive rewiring events of several transcription factors in cancer

#/*=== Rewiring section 644 words ===*/

In each cell type, we organized the TF regulatory network into a hierarchy by comparing the inbound and outbound edges of each factor, thereby enabling us to investigate the global topology of TF regulation (Fig 1E, see also Supp. File/Section(?) X). TFs in different levels of the hierarchy reflect the extent to which they directly regulate the expression of other TFs \{cite 25880651}. For example, TFs in the top layer have more outbound than inbound edges in the network, and thus play larger roles in regulating other TFs (Supp. Fig. xx). In this representation, two patterns readily emerge. For instance, in leukemia, top-level TFs tend to more strongly influence the differential expression between tumor and normal cells. The average Pearson correlation between TF binding events and tumor-normal expression changes increases from 0.125 in the bottom layer to 0.270 in the top layer (Table Sx). TFs in the bottom layer are more frequently associated with burdened binding sites in general, perhaps reflecting their increased resilience to mutation (see Supp. Section X, Table Sx).

Next, we compare the common TFs in matched tumor and normal samples, edge gain and loss analysis may help to identify cancer-associated dysregulation. Hence, we investigated such rewiring events in TF networks using multiple formulations (see Supp. File/Section(?) X). Specifically, for leukemia,out of the 69 common TFs in K562 and GM12878 from ENCODE, we removed the general TFs and restricted our rewiring analysis to the remaining 61 (see Supp. File/Section(?) X). We first ranked TFs according to their respective number of lost and gained edges (Fig 3 A, see also Supp. File/Section(?) X). For example, several oncogenes (such as MYC and NRF1) are among the top edge gainers. In contrast, IKZF1 (somatic mutations in which serve as a hallmark of high-risk acute lymphoblastic leukemia, or ALL) is the most significant edge loser, with up to xxx% of lost edges in K562 (Fig 3A). On the other hand, several ubiquitously distributed TFs (such as YY1) retain their regulatory linkages (as shown in Fig 3A). We observe a similar trend in TFs using a distal, proximal and combined network (see details in supplementary file). We also observe a similarly highly rewired TFs in lung and liver cancer (see fig XX) though we do not have as many common TFs between tumor and normal in these systems.

In addition, we also used a more complicated mixed-membership model to look more abstractly at local gene communities to re-rank the TFs (see Supp. File/Section(?) X). Similar patterns were observed using this model. We also observed that MYC (a well-known oncogene) became a top gainer (Fig 3A). To study the consequences of

network rewiring under this model, we performed the survival analysis on xxx AML patients, in which we found IKZF1 to be significantly associated with tumor progression (see Supp. File/Section(?) X).

One question is what gives rise to this rewiring. Is it the direct effect of mutations, which could, for instance, knock out a binding site, or is it the indirect of chromatin changes, which could cover and uncover binding sites? Hence, here we investigate the potential underlying causes of this rewiring, and find that the majority of rewiring events result from changes in chromatin status, rather than from variant-induced motif loss or gain events (Fig. 3A). For example, JUND is a top gainer in K562 (with xxx gains and xx losses). We find that up to 30.5% (58.1%) of the gain (loss) events are associated with substantial expression changes (of at least 2-fold), and that xxx% have large 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. We see a somewhat similar pattern in liver cancer for JUND, though it is not as pronounced as this factor does not rewire as much in this cancer (Fig 3D).

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

#/*=== expression section 250 (+111) words ===*/

To optimally leverage ENCODE data for studying various types of cancers, we extended our network analysis from strictly matched tumor-normal cell lines to more generalized networks. Here we can use both TF networks and those derived from RBPs, which are new data type now in ENCODE but for which we have no matched tumor normal pairs (see suppl. for description of merged networks).

Using a regression-based learning method (see Supp. File/Section(?) X). 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 regulator-cancer type pair, we selected the best explanatory binding profile and estimated the fraction of patients with differentially regulated target genes (see Supp. File/Section(?) X). The overall trends for the key TFs and RBPs discovered are given in Fig. 4A. The predicted impacts of regulators on tumor gene expression are highly consistent with previous findings. For example, we found that the target genes of *MYC* to be significantly up-regulated in numerous cancers (star in Fig Sx), which is consistent with the known role of *MYC* as an oncogenic TF. In addition to recapitulating existing knowledge from previous studies, our analysis also predicted previously unidentified functions for regulators in cancer. For example, the predicted targets of the RBP *SUB1* were significantly up-regulated in many cancer types (Figure 3C). As another example, the predicted targets of the TF *ZNF687* were significantly up-regulated in breast and prostate tumors (star in Supp. Fig. 2).

[JZ2MG: loregic to be here!]

The combinatorial regulation of many TFs jointly determines the "ON" and "OFF" states of all genes as part of maintaining homeostasis in healthy cells. The disruption of co-regulatory relationships for key elements in cancer cell lines ultimately result in erroneous gene expression patterns. We quantified the co-association status of each TF

and observed major co-association changes in some of the key TFs when comparing the regulatory network of K562 to GM12878. For example, ZNFXXX is a suppressor TF that shows only marginal co-binding events in GM12878. Its number of binding sites increases from xxx to xxx in K562. In addition, up to xxx% of its binding sites co-bind with other TFs.

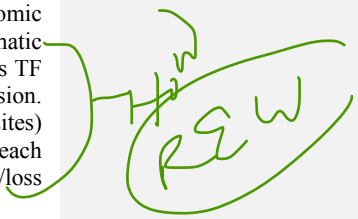
Comment [DC2]: ?

Comment [JZ3]: May completely move to the supplementary unless vineet's experiment is back

Step-wise prioritization schemes pinpoint deleterious SNVs in cancer

#/*=== conclusion section 326 words ===*/

Using the results from the precise, tissue-matched burdening analysis and the regulatory network and chromatin analysis, we can prioritize a variety of genomic features as being significant for oncogensiss. We describe this priorization in a systematic fashion in the workflow in Fig.5 A. We start by searching for key regulators (such as TF or RBPs) that are either massively rewired or drive tumor-normal differential expression. We then prioritize the functional elements (such as enhancers and TF-binding sites) governed by the key regulators through recurrence analysis. Lastly, we scrutinize each SNV therein by synthesizing features from annotation, conservation, and motif gain/loss events to pinpoint the impactful variants for small-scale functional characterization.



Using this framework, we subject a number of key regulators to knock downs, validating their important effect (fig XXX). We then identified several active enhancers in noncoding regions, and validated their ability to initiate transcription using luciferase assays (see Supp. File/Section(?) X). In addition, we further selected key SNVs within these enhancers that are important for gene expression control (table Sx). Of the 8 motif-disrupting SNVs that we tested, we observed 6 variants with consistent up- or down-regulated activity relative to the wild type (Fig. 5B and Supp. File/Section(?) X). One particularly interesting region is in chromosome 6, 13.5xxx (Fig. 5C). This enhancer is located in the noncoding region. Both histone modification and DHS signals implicate its regulatory role as being active (Fig. 5C). Note that both our histone-shape based enhancer prediction method and the EnhancerSeq experiment found this as an enhancer (Fig. 5D). Hi-C data indicate that this region is regulating an upstream gene XXX (DL to fill in). xx out of the XXX Chip-Seq experiments demonstrate significant binding events here, and the C to G mutation strongly disrupts the FOLS2 binding affinity (see Supp. File/Section(?) X). A luciferase assay demonstrates that this mutation introduces an xx-fold reduction in expression relative to wild type expression levels, indicating a strong repressive effect on this enhancer's functionality.

Conclusion

[[MG(long): how to improve? Better tn match, tissues? Deeper hic? Bigger cohorts… strength in numbers!!!!]]

#/*=== conclusion section 143 words ===*/

In the context of oncogenesis, this study comprehensively demonstrates the effectiveness of using ENCODE data to prioritize key regulatory elements and SNVs at different scales. Our scheme integrates hundreds of experiments from ENCODE to interpret the large number of noncoding variants from massive cohorts and pinpoint key regulatory changes for detailed functional characterization. It is worth mentioning our

interpretation could be noticeable improved with the release of more data. For example, with better normal-tumor paired functional characterization data, our analysis could be extended to cancer types other than the five cancer mentioned above. Deeper Hi-C data could provide enhancer gene linkages at better resolution and larger external cohorts data would potentially improve the statistical power to identify these key elements/SNVs. We believe that our scheme could readily handle the fast growing number of functional characterization data and further improve our understanding of cancer.