ENCODEC: An integrative ENCODE resource for cancer genomics

Introduction

With the initial ENCODE release in 2012, investigators began to systematically map functional elements in the human genome, such as transcription, chromatin accessibility, histone modifications, and transcription factor (TF) binding \{cite}. The most recent ENCODE data release offers much greater scope than that of the previous release. In addition to the much greater number of different cell types studied, the more recent datasets provide considerably more novel assays on several "top-tier" cell types. [[PDM2all: some short text could go here stating why there is an opportunity and challenge in building an integrative resource. E.g. “This increased scale of data represents an opportunity to ascertain the function of genomic regions at a closer resolution and with greater accuracy than ever. However, the scale of data also represents a technical challenge relative to its processing and integration.”]] Hence, focusing on top-tier ENCODE cell types, we performed deep integration of ENCODE data over tens of functional assays to deeply characterize the noncoding genome at higher resolution and greater accuracy than has previously been possible.

[JZ2ALL: did we introduce well what is the resouce? I am afraid not, but if I do the bullets, it will be too repeatative with why cancer is the best application. I am currently thinking to put the main resource into the conclusion part. But we have to let them know there are NETWORKS]

[[PDM2all: most of my suggested edits in the following paragraph are just softening the language a bit.]]

Our integrative ENCODE annotation allows accurate functional interpretation of germline and somatic genome variations in a tissue-specific manner and hence serve as a valuable resource for many disease studies. We consider interpretation of genomic variation in cancer one of the best applications of our resource for several reasons. First, many of the most data-rich cell types are are from cancer cells, including cell types from blood, breast, liver, and lung (Fig. 1). Second, the wealth of ENCODE data, such as replication, epigenetic, and transcriptional profiles, may be used to inform our understanding of cancer mutational processes for both single nucleotide variations (SNV) and structural variations (SV). Third, data in the most recent ENCODE release can be used to

shed light on epigenetic remodeling and cell-state transitions, which are implicated in oncogenesis. Lastly, ENCODE constructed [[PDM2all: did ENCODE construct these networks, or does ENCODE data allow for the construction of these networks?]] a variety of high-quality regulatory networks from thousands of experiments to provide a systems-level perspective of cancer. One may thus directly measure the perturbations of individual regulators and entire networks to better elucidate the biological mechanisms of cancer inititation and progression.

[JZ2MG: I know you may not like the novel findings in this para, but I feel we do have to list our results here to convince the referees. I only listed results that we are confident of. To disc.][[PDM2all: Notably extended gene, compact annotation, BMR, not mentioned directly. BMR is maybe least defensible (until we show improvement). Extended gene / compact annotation may deserve a sentence or two -- reviewers liked these, if we can substantiate claims.]]

[[PDM: This next paragraph is fairly important. Given that this is a resource, we might want to cover (roughly in order) 1. What resource we built / what it is. 2. Why it is valid / how we validated it. 3. Some examples of its applications w/ notable results.]]

Therefore, we present the integrative *ENCODE* companion resource for *C*ancer genomics (*ENCODEC*). This resource consists of various annotations, networks, and code bundles available online (encodec.encodeproject.org). Our resource is designed specifically for investigating cancer, but it may also be applied to the interpretation of genomes in a variety of disease and non-disease contexts. [[PDM32: Some text about validity could go here. The next two sentences could also be reframed as validation e.g. 'We validated our resource through the successful recovery of well-documented variations and regulators associated with cancer … as well as through statistical comparison to benchmark standards where appropriate.]] We applied ENCODEC to interpret and prioritize key regulators, SNVs, and SVs associated with cancer progression. Specifically, in addition to recovering well-documented variations and regulators associated with cancer, such those affecting the TERT promoter, TP53, and the oncogene transcription factor (TF) MYC,we also highlighted potentially novel noncoding SNVs in enhancers that change gene expression, SVs that can initiate oncogene expression, and key RNA binding proteins (RBPs) such as SUB1. We further validated our results using small-scale experiments such as luciferase assays, CRISPR-engineered deletions, and TF/RBP knockdowns. Our successful validations, serve as an illustration of how our new ENCODEC resource can immediately be leveraged for cancer research.

[JZ2MG: I prefer to merge some parts breadth=BMR/SV, depth=enhancer and E2G linkage, extended gene is just extended genes]

The breadth and depth of the ENCODEC resource

Our ENCODEC resource takes advantage of the breadth and depth of the ENCODE3 resource and customizes it for cancer research. We collected 2069 epigenetic and replication timing experiments from 229 cell types to facilitate the interpretation of cancer mutational process for both SNVs and SVs. Such features have been proven to facilitate background mutation rate (BMR) estimation through many mathematical models \{cite}. We demonstrated that the inclusion of ENCODE data brings additive improvements in BMR estimation that scales with the number of included features (Fig.1) \{cite}. Further, the modelling of BMR using ENCODE feature data simulatenously improves computational speed and interpretability with equivalent accuracy to established non-parametric BMR models. . We also demonstrated that the richness of the ENCODE data may help to provide genomics data that is strictly matched by cell type or tissue, to uncover the underlying mutational mechanism for SVs. For instance, we found that breakpoints in K562 are associated with H4K20me1, which is an activation histone marker, only in K562 but not in others.

[[PDM: I may have over-edited this next paragraph. It just wasn't initially clear to me how the first sentence claiming 'accurate and compact annotations' connected with subsequent sentences. I think it may be clearer now. However, until the edits are accepted/rejected, it's hard to read, and needs to be checked for accuracy.]]

We further leveraged the novel assays in ENCODE3 tobuild accurate and compact annotations. For example, in order to build a compact set of regulatory elements enriched for functional significance, we were able to restrict a genome-wide list of potential proximal and distal regulatory elements using ENCODE3 data. In particular, xxx ChIP-seq experiments were used to refine a list of XXX million proximal regulatory elements of likely functional significance. ~~explored xxx ChIP-seq experiments on top tier ENCODE cell types to define millions of transcriptional level proximal regulatory elements. We integrated the full catalogue of ENCODE assays and proposed an ensemble method to define high quality distal elements such as enhancers. Specifically~~ developed a machine-learning algorithm called match-filter to combine DNase-seq with up to 10 histone modification marks to predict enhancers. Then, using a multi-resolution peak-calling algorithm called ESCAPE, we combined our match-filter predictions with STARR-seq data. This ensemble-based approach enabled us to define a minimal list of enhancers with fewer false-positives. Our annotation of 112 RNA binding proteins (RBPs)

built using post-transcriptional eCLIP experiments is significantly more compact than previous transcriptional level annotations (xxx. VS xxx for TF, P=… two-sided Wilcoxon test), while also demonstrating higher cross-population and cross-species conservation consecration [[PDM: Consecration??]] [[PDM: Higher conservation compared to what?]] that suggests the accuracy of our annotation. RNA binding proteins play key roles in various diseases including cancer, but their role is largely ignored in many previous annotation efforts \{cite}. We believe such smaller and "core" region definitions (which are enriched for true functional impact) significantly improve key variant detection.

An extended gene annotation and its applications for interpreting variants and differential expression

Much current knowledge of disease has been derived by focusing on protein-coding regions. To broaden the scope of elements studied, we also linked our above noncoding annotations to genes in order create a gene-centric annotation (which we call the extended gene). Our extended gene annotation includes both proximal and distal, transcriptional and post-transcriptional level annotations (Fig. 2A). Specifically, in contrast to most previous efforts that rely solely on correlating individual genomic signals, we used a machine learning algorithm that integrates a wide variety of histone modifications, gene expression signals, and physical interactions from Hi-C and ChIA-PET to link distal regulatory elements to genes with greater accuracy [[PDM: Greater accuracy compared to what standard?]] (suppl. sect. xxx). We then demonstrated the value of our extended gene annotations by applying it for to somatic, germline, and expression analyses.

First, we used the extended gene annotation as a single test unit for recurrence analysis, rather than testing all regions separately. Such a unified scheme enables joint evaluation of the mutational signals from distributed yet biologically connected genomic regions. Fig. 2B illustrates the larger number of known cancer-related genes detected in several cohorts, [[PDM: Do we really do a relative comparison to 'traditional approaches' in one of our figures (2B)? What are 'traditional' approaches? Single region recurrence testing?]] relative to those derived from traditional approaches. For instance, in the context of chronic lymphocytic leukemia (CLL), our joint detection approach identified well-known highly mutated genes (such as TP53 and ATM) \{cite}, in addition to genes that would otherwise be missed by exclusively focusing on coding regions. [[PDM: Not sure about 'case-in-point', maybe just return to 'an example'.]]As a case-in-point, we identified the well-known cancer gene BCL6, which may be associated with patient survival (Fig. 2B and refs. [1-3]). Secondly, our extended gene definitions

include many tissue-specific proximal and distal noncoding regulatory elements that are useful for interpreting cancer-associated GWAS variants. To illustrate this, we calculated the enrichment of cancer GWAS SNPs with respect to various annotations. We observed a positive relationship between GWAS SNP enrichment and the number of annotation categories included [[PDM: I'm not sure I understand this terminology/analysis. Annotation categories == sets of features? Why not just say features?]] (Fig 2C). We note that, in contrast to the unified gene definitions across different cell types, the tissues-specific experiments allow us to build a highly dynamic extended gene definition that is unique to specific cancer types. Indeed, the greatest enrichment of GWAS SNPs is achieved using tissue-matched samples [[PDM: need a pointer to the evidence for this here. Fig 2C?]].

Thirdly, our extended gene annotation can better [[PDM: better compared with what?]] stratify gene expression signals of cancer patients by their mutational status. For instance, we combined the mutational and expression profiles from large cohorts and found that mutational status in our extended gene definition can explain the expression differences for a larger [[PDM: larger than what?]] number of genes. One example [[PDM: Example of? Perhaps: 'example of a gene where an extended gene annotation helps provide perspective on the relationship of variation and gene expression'?]] is the splicing factor SRSF3, which has been shown to affect liver cancer progression. Aggregate mutational burden falling within its extended gene annotation in HepG2 exhibits greater significance relative to gene expression, compared with to any single annotation category (p=xxx, one sided Wilcoxon test).

Finally, we tried to interpret the impact of variations by showing an example of SV introduced extended genes dynamics that leads to oncogene activation. ERBB4 is a well-known oncogene in many cancer types \{cite}, and is significantly upregulated in tumor [[PDM: It's an oncogene (vs. tumor suppressor) so do we need to say this? If so, is there a particular tumor type or cancer type? Also, grammar needs fixing.]] (Fig. 2D). We noticed a 130Kb heterozygous deletion (about 45Kb downstream from the TSS) that potentially merges two Hi-C TADs and links a distal enhancer to the ERBB4 promoter in T47D cells but not in normal cells (xxx color track from 4C experiment). We therefore hypothesized that the heterozygous deletion disrupts the insulation of ERBB4 from distal regions, thereby activating its allele-specific expression . We tested this hypothesis by CRISPR editing to excise the 86bp sequence on the wild-type allele in T47D cells and found elevated ERBB4 expression (as measured by PCR). Our results suggest that ERBB4 activation in T47D is at least partially due to the 130 kb deletion that disrupts its insulation.

[JZ is done one all parts, please revise all sections! Thank you!]

Leveraging ENCODEC networks to prioritize regulators

We compiled xxx ChIP-seq and xxx eCLIP experiments to build comprehensive and accurate proximal regulatory networks. Compared to networks derived from gene expression or motif analyses, our ENCODE TF and RBP networks are built using experimentally-defined regulatory linkages between functional elements (suppl. sect. xxx), thereby enabling us to more accurately capture interactions (suppl. sect. xxx).

CONNECT BETW BOTH

We analyzed the overall TF and RBP regulatory network by systematically arranging it into a hierarchy (Fig. 3A) in which TFs at top levels tend to regulate those in the levels below (suppl. sect.). We found that top-level TFs are not only enriched in cancer-associated genes (P=xxx, Fisher's exact test), but they also more significantly drive differential expression in model cell types (P=xxx, one sided Wilcoxon Test).

[JZ2all: I feel this part is too wordy… but difficult to shrink though]

Our networks also enable gene-expression analyses in tumor samples. We used a regression-based approach to integrate 8,202 tumor expression profiles from TCGA and searched for TFs and RBPs that most strongly drive tumor-specific expression (see suppl.). We tested the degree to which a regulators' activities correlate with tumor-to-normal expression changes of their respective targets to prioritize key TFs and RBPs in cancer (Fig 3B).

DIFF

As expected, we found that many previously reported cancer-associated TFs show high regulatory potential and are associated with patient survival (Fig. 3B). For instance, we found that MYC targets are significantly up-regulated in numerous cancer types. We performed MYC knockdowns in MCF-7 and confirmed that MYC targets exhibit significantly greater expression reduction upon knockdown (Fig. 3C). Similarly, in the RBP network, we found that SUB1 upregulates its target genes in many cancer types and SUB1's upregulation may indicate poor patient survival in lung and liver cancer (Fig. 3D). We confirmed the positive regulatory role of SUB1 through SUB1 knockdown in HepG2 cells. SUB1 knockdown resulted in significantly reduced target expression (Fig. 3D).

SUB1 has not previously been associated with cancer as an RBP, so we sought to explore its role in oncogenesis. We found that SUB1 tends to bind to distal 3'UTR regions, and its targets are enriched in CGC genes (p=xxx, Fisher's exact test). We also find that decay rates of SUB1 targets are significantly lower

than those of non-targets (see suppl.), and hence hypothesize that it stabilizes the transcripts of its targets.

We further investigated how key regulators can interact with others during regulatory processes in tumors. For MYC, we found that, with the exception of its well-known co-regulators MAX and MXL1, NRF1 is the most frequent co-regulator that forms a feed-forward loop (FFL). Upon further examination, we found that MYC-NRF1 FFLs were mostly coherent (i.e., "amplifying" in nature; in supplement). We further studied these FFLs by organizing them into logic gates, in which two TFs act as inputs and the target gene's expression represents the output[4] (see suppl.). We show that most of these gates follow either an OR or MYC-always-dominant logic gate. Similarly, with respect to RBPs, MYC is the top co-regulator with MYC after correcting for many potential confounding factors, such as GC content and expression (see suppl. sect.). Interestingly, we found that SUB1 is a direct target of MYC in many cell types (see supplements), forming many FFLs in the regulatory network. We hypothesize that MYC can bind to the promoter regions of key oncogenes to initiate their transcription and SUB1, and it binds to 3UTRs to stabilize such genes at the level of RNA transcripts. This collaboration between MYC and SUB1 results in overexpression of several key oncogenes and leads to proliferation of cancer cells. To validate this hypothesis, we knocked down MYC and SUB1 separately in HepG2 and used qPCR to quantify changes in gene expression. As expected, the expression of oncogenes (such as MCM7, BIRC5, and ATAD3A) is significantly reduced.

Cell-type specific regulatory networks highlight extensive rewiring events during oncogenesis

For the top-tier cell types that have numerous associated TF ChIP-seq experiments, we were able to build cell-type specific regulatory networks, thereby enabling direct comparisons with networks built from their paired normal cell types. To achieve the best pairing (given the existing data), we construct a "composite normal" by reconciling multiple related normal cell types (see suppl.). Although the pairings are only approximate, many of them have previously been widely used in the literature (see suppl.). Furthermore, they leverage the extensive functional characterization assays in ENCODE to provide us with a unique opportunity to study the regulatory alterations in cancer on a large scale for the first time.

In particular, we measured the signed, fractional number of edge changes (which we call the "rewiring index") in "tumor-normal pairs" to evaluate how TF targets may change over the course of oncogenic transformation. In Fig. 4A, we

used this index to rank TFs. In addition to direct TF-to-gene connections, we also measured rewiring using a more complex gene-community model, where targets within the regulatory network were characterized in terms of heterogeneous modules from multiple genes (so called "gene communities") (see suppl. sect.). Similar patterns to direct rewiring were observed using this model (Fig. 4A). In leukemia, well-known oncogenes (such as MYC and NRF1) were among the top edge gainers, while the well-known tumor suppressor IKZF1 wass the most significant edge loser (Fig. 4A). We observed a similar rewiring trend using distal, proximal, and combined networks (details in suppl.). This trend was also consistent across a number of cancers: highly rewired TFs (such as BHLHE40, JUND, and MYC) behaved similarly in lung, liver, and breast cancers (Fig. 4).

We found that the majority of rewiring events were associated with noticeable gene-expression and chromatin-status changes, but not necessarily with SNV and SV introduced motif gain or loss events (Fig.4D). For example, JUND is a top edge gainer in K562. Many of its gained targets in tumor cells exhibit higher gene expression (as well as stronger active and weaker repressive histone modification mark signals), yet few of its binding sites are mutated or affected by structural variations.

Stemness measurement during oncogenic transformation through regulatory networks

A prevailing decades-old paradigm has held that at least a subpopulation of tumor cells have the ability to self-renew, differentiate, and regenerate in a manner similar to that in stem cells. We projected the xxx RNA-seq data by RCA and observed that tumor cells (green in Fig. 5A) are more similar to the stem group (blue) than its normal counterpart (yellow), which is consistent with recent discoveries \cite{TCGA stemness}. Furthermore, we explored the extensive proximal and distal regulatory networks in ENCODE and observed a consistent (or even more obvious) pattern: tumor cells tends to cluster together around stem cells and stay away from the normal ones.

It is also well-known that dysregulation of key oncogene TFs are hallmarks of tumor progression. Key genes, such as MYC, initiate overexpression of other oncogenes in tumor cells. To test our hypothesis that tumor cells are more similar to stem cells, we measured the perturbations introduced by oncogenic TFs through expression comparisons before and after TF knockdowns. Interestingly, the overall expression profiles reverted slightly back toward normal state upon oncogene knockdowns.

A step-wise prioritization scheme pinpoints key SNVs in cancer

In sum, our companion resource consists of the annotations in Figs. 1 and 6: (1) thousands of uniformly processed genomic features for genome variation interpretation; (2) accurate and compact annotations and their linkage to genes to form extended gene definitions; (3) tumor-to-normal genetic, epigenetic, and high-dimensional structural changes; (4) accurate regulatory networks for TFs and RBPs with comprehensive regulator attributes, such as a network hierarchy, overall disruptiveness, regulatory potential, and rewiring status. Together, these resources are made available online through the ENCODE website as flat text files as well as code bundles (suppl. sect. xxx).

[JZ2DL: could we change the Fig. 6A to incorporate regulator and SNVs?]

Collectively, these resources allowed us to prioritize key regulators, SVs, and SNVs associated with oncogenesis. Our prioritization scheme is presented in Fig. 6A. We first searched for key regulators that are frequently rewired, located in network hubs, sit at the top of the network hierarchy, or significantly drive expression changes in cancer. We then prioritized functional elements associated with these regulators that are highly mutated in tumors, or undergo large changes in gene expression, TF binding, or chromatin status. Finally,at a finer scale we pinpoint impactful SNVs by estimating their ability to disrupt or introduce specific binding sites..

To demonstrate the utility of ENCODEC , we instantiated our prioritization workflow in a few select cancers and experimentally validated the results. In particular, as described above, we subjected some key regulators (such as MYC and SUB1) to knockdown in order to validate their regulatory effects (Fig. 3B and 3D), and validated the effects of SVs on oncogene activation through CRISPR deletions. We also identified several candidate enhancers in noncoding regions associated with breast cancer and validated their ability to influence transcription using luciferase assays in MCF-7. Finally, we selected key SNVs, based on mutation recurrence in breast-cancer cohorts and motif disruption scores within these enhancers that are important for controlling gene expression. Of the eight motif-disrupting SNVs that we tested, six exhibited consistent up- or down-regulation (relative to wild-type cells) in multiple biological replicates.

Specifically, CDH26 (an intronic region in chromosome 20) serves as an in interesting example to illustrate the value of ENCODE data integration (Fig. 6C). The signal shapes for both histone modification and chromatin accessibility (DNase-seq) data indicate its active regulatory role as an enhancer in MCF-7. This was further validated using STARR-seq assays (Fig. 6C). Hi-C and ChIA-PET linkages indicated that the region is within a topologically associated domain (i.e.,

a "TAD") and validated a regulatory connection to the breast-cancer-associated gene SYCP2[8]. We further observed strong binding of many TFs in this region in MCF-7. Motif analyses predict that the particular mutation from a breast cancer patient significantly disrupts the binding affinity of several TFs, such as FOSL2, in this region (Fig. 6C). Luciferase assays demonstrated that it introduces a 3.6-fold reduction in expression relative to that in wild-type cells, thereby indicating a strong repressive effect on enhancer functionality.

Conclusion

This study highlights the value of deep data integration over many novel assays to annotate noncoding elements of the genome. We provided accurate tissue-specific extended gene annotations and extensive regulatory networks through integration of thousands of experiments. We find the one of the best application of our resource is in cancer since there are many of cell types associated with cancer.

A key caveat related to part of our resource, such as rewiring in cell-type specific networks, is based on associating a particular cancer type with a composite normal. Such "correspondences" may be approximate. Another limitation is that most of the current release is performed over many cells. However, heterogeneity in tumor cells and their microenvironments (e.g., immune cell infiltrates, hormonal factors, normal cell populations, etc.) significantly affect tumor growth and development. We therefore believe that the development of single-cell sequencing technologies may better capture tumor biology at a higher resolution and provide new insights in cancer.

Nevertheless, we feel that ENCODEC currently provides the most comprehensive view of oncogenic regulatory landscapes available. No other system has this scale of functional characterization data. Moreover, the heterogeneous nature of cancer means that even tumor cells from a given patient usually show distinct molecular, morphological, and genetic profiles[9]. It is difficult to obtain a "perfect" match even from real tumor and normal tissues taken from a single patient.

[[PDM2all: Just to point it out -- rather than acting as a synthesis of the paper, this final paragraph suggests something new: that what we developed is an approach or method that has not yet realized it's full potential. It's a fair point to make, but it's also a new point in this final paragraph (not mentioned elsewhere in the text that we are building a method/approach). It is also somewhat different than the framing of ENCODEC as a resource with results and validations.]]

In general, our study underscores the value of large-scale data integration, and we note that expanding the scale of our approach in a number of dimensions is straightforward. For example, we observed increased accuracy of BMR estimation with additional genomic features; we expect that this accuracy will increase further still with more features. We successfully identified extended gene annotations and regulatory networks for model systems that are already replete with advanced functional assays like eCLIP and STARR-seq; our methods can be readily extended to further model systems when they are similarly assayed in the future. Given the rewiring formalism presented here, it should be straightforward to expand the analysis to a greater number of TFs. (In fact, we note that the re-wiring formalism actually provides a way of selecting candidate TFs and cell types). We anticipate that this will provide a clearer and more accurate picture of the spectrum of regulators that are affected by extensive chromatin changes, and thus help prioritize research efforts in cancer. Finally, we demonstrated the utility of our resource for assisting in the detection of potential cancer drivers in limited publically available cohorts; we anticipate that integrating ENCODEC with the large cohorts currently being assembled (e.g., PCAWG, pancancer.info) will provide even greater value.

========== JZ2All: forget about figure legend at this round ==========

Figure Legend

Figure 1

Schematic of the EN-CODEC resource. Columns list cell types and rows list assays. **Pink box**: "Top-tier" cancer-associated resources in ENCODE highlighting the depth of the resource. **Yellow box**: Cell types with several assays in the main ENCODE Encyclopedia highlighting the breadth of the resource. **Green box**: Cell-type specific analyses based on deep annotations of top-tier cell lines. **Blue box**: Merged analyses based on wide-coverage of many cell types. The actual content of our resources (annotations, background mutation rate, networks) are shown in the dotted black box.

Figure 2

BMR modeling and mutation burden analysis. (A) Improvement of BMR estimation by accumulation of principal components of multiple genomic features.

(B) In breast cancer, regression coefficients of remaining features after incorporating MCF-7 replication timing. **(C)** Schematic of extended gene definition. **(D)** Significantly burdened genes using noncoding elements (TSS), coding regions (CDS) and extended genes, alongside germline mutational status in liver cancer. **(E)** Expression of BCL6, which is only identified as recurrently mutated using extended genes, is correlated with patient survival.

Figure 3

Integration of ENCODE networks with expression profiles. (A) Heatmap of regulatory potentials of TFs/RBPs to drive tumor-to-normal expression changes; red and blue indicate up- and down- regulation. **(B)** Elevated MYC regulation activity is associated with reduced disease specific survival (DSS) in breast cancer (top); MYC knockdown in MCF-7 leads to significantly larger expression reduction in MYC target genes (bottom). **(C) (i)** MYC expression is more positively correlated with its target genes as compared to other TFs; **(ii)** MYC frequently form FFLs with NRF1, and these are mostly coherent; **(iii)** In the MYC-NRF1 FFLs, OR-gate logic predominates. (**D**) Elevated SUB1 regulation activity is associated with reduced overall survival (OS) in lung cancer (top); SUB1 knockdown in HepG2 leads to reduced target gene expressions (bottom).

Figure 4

Regulatory network hierarchies. TFs are organized into layers such that top layer TFs tend to regulate others, while bottom layer TFs tend to be regulated by others. **(A)** Generalized network: top layer TFs are enriched with cancer associated genes and demonstrate larger regulation potentials to drive tumor-to-normal gene expression changes. **(B)** Cell-type specific network using K562 and GM12878: top layer TFs significantly drive tumor-normal differential expression; bottom layer TFs are more often associated with burdened binding sites.

Figure 5

TF-Gene network rewiring. Green and red arrows designate edge gain and loss, respectively. **(A)** Rewiring index in a model for CML by direct edge counts using both proximal and distal networks (top) and by gene community analysis (bottom). TFs that gain edges tend to rewire away from stem cell-like state while TFs that lose edges tend to rewire toward stem cell-like state. **(B)** Examples of network rewiring for specific TFs in multiple cancer types. **(C)** Conceptual schematic for

rewiring towards or away from a stem cell-like state. **(D)** Genomic features associated with gained or lost edges.

Figure 6

Variant prioritization and validation. (A) Stepwise variant prioritization scheme utilizing EN-CODEC resources. We prioritize large-scale regulators based on network and expression analysis; regulatory elements based on mutation burden; then single nucleotide by motif gain/loss and conservation score. **(B)** Small-scale validation of prioritized variants using luciferase reporter assay. **(C)** Multiscale integrative analysis on Sample 5 with assorted functional genomics data. We start from large-scale Hi-C linkages, and then zoom into element level by highlighting signal tracks of histone modification marks and DNase hypersensitivity together with various TF binding events. At the nucleotide level, FOSL2 motif is disrupted.