An integrative ENCODE resource for cancer genomics

Jing Zhang*, Donghoon Lee*, Vineet Dhiman*, Peng Jiang*, William Meyerson, Matthew Ung, Shaoke Lou, Patrick Mcgillivray, Declan Clarke, Mengting Gu, Lucas Lochovsky, Lijia Ma, Grace Yu, Arif Harmanci, Koon-kiu Yan, Anurag Sethi, Qin Cao, Daifeng Wang, Gamze Gursoy, Jason Liu, Xiaotong Li, Michael Rutenberg Schoenberg, Joel Rozowsky, Lilly Reich, Juan Carlos Rivera-Mulia, Jie Xu, Jayanth Krishnan, Yanlin Feng, Jessica Adrian, James R Broach, Michael Bolt, Vishnu Dileep, Tingting Liu, Shenglin Mei, Takayo Sasaki, Claudia Trevilla-Garcia, Su Wang, Yanli Wang, Hongbo Yang, Chongzhi Zang, Feng Yue, David M. Gilbert, Michael Snyder, Kevin Yip, Chao Cheng, Robert Klein, X. Shirley Liu, Kevin White, Mark Gerstein

Abstract

ENCODE now comprises thousands of functional genomics data sets; it is possible to tailor these into a targeted resource for interpreting cancer genomes. In particular, this resource can be used, firstly, to measure the impact of non-coding mutations, constituting the bulk of the somatic variants. Moreover, by integrating advanced assays (e.g. STARR-seq) with many epigenetic features, we can make a more focused and refined genome annotation, thereby increasing the power for detecting recurrent somatic mutations in cohorts. Second, ENCODE signal data (e.g. replication-timing data) allow us to construct cell-type-matched models for background mutation rates considerably more accurate than previous models. Third, ENCODE data, incorporating new assays, such as Hi-C and RNA-binding protein assays, in addition to large-scale transcription-factor ChIP-seq, allows the construction of extensive regulatory networks. In some contexts, these networks reveal how connections "rewire" during oncogenesis, as well as how these changes relate to a stem-cell state. More generally, one can use ENCODE networks to prioritize regulators most associated with large-scale expression changes in cancer. Combining the networks with the refined annotations and background mutation models, one can develop a step-wise scheme for prioritizing non-coding mutations. Here, we show how this can be instantiated. We perform a number of focused experimental validations (i.e., luciferase assays and siRNA knockdowns) to demonstrate how the resource can highlight mutations with significant consequences in cancer.

Introduction

The initial ENCODE release in 2012 and other targeted functional genomic data have motivated many integrative studies, some of which have focused on cancer genomes[1-7]. Specifically, functional genomics data have been used to investigate cancer in three ways. First, they enable researchers to evaluate the molecular functional impact of non-coding mutations -- the vast majority of variants in cancer genomes -- and to identify non-coding annotation “elements” (e.g., enhancers)[6,8-11]. Secondly, by incorporating genome-wide features (such as replication timing, methylation, and expression), functional genomics data sets can be used to estimate background mutation rates (BMR), which vary widely over the genome[12-14]. Precise BMR calibration enables us to accurately identify recurrently mutated annotation elements across cancer cohorts for candidate drivers[15-17]. Finally, ENCODE data and other genomic data sets have been used to link non-coding elements and organize them into regulatory networks, which can be used to gain a systems-level perspective on cancer[18-20].

The new release of ENCODE data has a number of improvements over the last release, which was mainly focused on a limited number of cell types using RNA-seq, DNase-seq and ChIP-seq assays[21]. The new release has two new directions. First, it considerably broadened the number of cell types using the original assays. As such, the main ENCODE encyclopedia aims to utilize these to provide a general annotation resource applicable across many cell types. Second, ENCODE also expanded the number of

advanced assays on several "top-tier" cell types (e.g. STARR-seq, Hi-C, ChIA-PET, eCLIP and RAMPAGE). Many of these are associated with various types of cancer, including those of the blood, breast, liver, and lung (K562, MCF-7, HepG2, A549, see Fig. 1). Such rich functional assays and annotation resources in the new ENCODE release allow us to characterize these non-coding regions in depth and construct a customized *ENCODE* companion resource for *C*ancer genomics (which we call EN-CODEC). This resource consists of a set of annotation files and computer codes available online (encodec.encodeproject.org, see suppl.). It comprises three main parts: background mutation rate models, compact annotations, and regulatory networks. We detail each of these parts below and provide illustrations of how they may be used to interpret cancer genomes after combining mutation and expression profiles from large cancer cohorts, such as TCGA.

In particular, with a much wider selection of cell types, EN-CODEC provides substantially more functional genomics data that can be better matched to specific cancer types of interest, allowing a demonstrably improved background mutation rate estimation. In addition, for a number of well-known cancer cell types, it incorporates a large battery of data on histone marks with various more specialized assays. For example, in several model cell types, we incorporate STARR-seq data, which directly measures the genome-wide enhancer activities, to accurately define core enhancers and used Hi-C and ChIA-PET data for accurate enhancer-gene linkage prediction. Consequently, relative to generic annotations, it constructs more compact annotations to maximize statistical power in the determination of mutationally burdened regions.

Finally, our resource significantly extends TF regulatory networks with comprehensive ChIP-seq coverage across cell types and constructs additional networks from more recent assays such as eCLIP and Hi-C. For a few prominent cancers (e.g. blood and liver cancer), these provide cell type specific networks in model tumor and normal cells, thereby enabling direct measurement of potential regulatory changes in oncogenesis. Furthermore, a prevailing decades-old paradigm has held that at least a subpopulation of tumor cells has the ability to self-renew, differentiate, and regenerate, in a manner similar to stem cells[22]. Hence, the top-tier cell line H1-hESC can serve as a valuable comparison when investigating the degree to which an oncogenic transformation moves towards or away from a stem-cell-like state. More generally, our network can better explain cancer specific expression patterns in tumors from cancer resources such as TCGA, and it also helps reveal key regulators that drive large-scale tumor-to-normal expression changes.

We combined the ENCODE networks with the compact annotation sets and mutational burdening analysis (from the enhanced background model) to propose a step-wise prioritizing scheme that highlights key mutations associated with cancer progression. We validated the functional impact of prioritized mutations and elements using focused experiments such as siRNA RNA-seq and luciferase assays. Such prioritization serves as an illustration of how the new EN-CODEC resource can immediately be used to help analyze existing cancer mutation data and cancer-associated gene expression.

ENCODE data allows more accurate BMR estimation (for better cancer driver detection)

One of the most powerful ways of identifying key elements in cancer genomes is through mutation recurrence analysis to discover regions that harbor more mutations than expected. However, developing a null expectation for these analyses is non-trivial – the somatic mutation process can be influenced by numerous confounders (in the form of both external genomic factors and local sequence context factors), and these can result in false conclusions if not appropriately corrected[15]. Hence, we demonstrate how to integrate extensive ENCODE data to construct an accurate background mutation rate model in a wide range of cancer types.

We address this issue in a cancer-cohort-specific manner (see suppl.). Specifically, we separated the whole genome into bins (1Mb) and calculated bin-wise mutation counts. We used a negative binomial

regression of the mutation counts against 475 genomic features across 229 cell types, including replication timing, chromatin accessibility, histone modifications, methylation, Hi-C, and expression profiles. In contrast to methods that use data from unmatched cell types, our approach automatically selects the most relevant features, thereby providing considerable improvements in BMR estimation (Fig. 2A). For example, using matched replication timing data in multiple cancer types significantly outperforms an approach in which one restricts the analysis to replication timing data from the unmatched HeLa-S3 cell line. Moreover, combining many different genomic features significantly improves the estimation accuracy (Fig. 2B). The weightings of the features in the model are consistent with our expectations: for instance, for breast cancer, we observed elevated mutation rates in regions with the repressive mark H3K9me3 and a reduced mutation rate in regions with the activating, enhancer-associated mark H3K27ac[12-14]. Also, due to the correlated nature of genomic features across cell types, even approximate matching of a specific cancer type to a particular ENCODE cell line can still improve BMR estimation (see suppl.). Hence, our analyses may easily be extended to many cancer types.

A focused, compact annotation increases power for detecting cancer drivers

A second advantage of leveraging ENCODE data in determining recurrently mutated regions is provided by maximizing the statistical power of burden tests. In traditional genomic analyses, a comprehensive set of annotations (usually covering as many base pairs as possible) is considered to be optimal. However, testing every possible nucleotide in the genome greatly reduces the statistical power for variant recurrence detection (see suppl.). Here, we aim to increase the power of burden tests by creating a focused, compact annotation for a given cell type.

First, for a single burden test on an individual genomic element (e.g., an enhancer), focusing on a smaller, "core" region, enriched for true functional impact, significantly improves detectability (see suppl.). Hence, we trimmed the conventional annotations to key "functional territories" by using the well-known small territories of TF-binding sites and the shapes of various genomic signals (e.g., the well-known double-hump of H3K27ac around enhancers, see suppl.).

Second, repeated burden tests on a large number of elements would be subject to a large multiple-testing penalty. Thus, we tried to restrict our annotation set to a minimum number of high-confidence elements. With a particular focus on enhancers, we started by searching for regions supported by multiple lines of evidence in the data-rich top-tier cell types. We developed a machine-learning algorithm to combine DNase-seq experiments and a battery of up to 10 histone modification marks to predict enhancers (see suppl.). Using a second algorithm, we then combined these predictions with our processing of the STARR-seq experiments (see suppl.). These experiments provide a direct, albeit noisy, readout of enhancer activity in specific cell types. Such an "ensemble" approach enables us to define a minimal list of enhancers with as few false-positives as possible. We also reconciled and cross-referenced our "compact annotation" with the main encyclopedia annotations (see suppl.).

An extended gene annotation by linking genes to non-coding elements (for better cancer driver detection)

To increase statistical power, a final part of our "compact" annotation entails linking non-coding regulatory elements to protein-coding exons to form an extended gene region as a single test unit. Such a unified annotation enables joint evaluation of the mutational signals from distributed yet biologically connected genomic regions. Traditional methods for linking rely solely on the correlation of individual signals (e.g., between the activity of one histone mark at an enhancer and gene expression of neighboring

genes), and these may result in inaccurate extended gene definitions. Here, we use direct experimental evidence on physical interactions from Hi-C and ChIA-PET experiments, combined with a machine learning algorithm that takes into consideration the wide variety of histone modification marks and gene expression to delineate accurate enhancer-target gene linkages.

By integrating our compact annotation sets, BMR estimates, and accurate extended gene definitions, we were able to obtain maximal power for detecting genomic regions (coding and non-coding) that are mutationally burdened. Fig. 2C illustrates the greater power in detecting mutationally burdened non-coding regions in several well-known cancer cohorts. For example, in the context of chronic lymphocytic leukemia (CLL), our analyses identified well-known highly mutated genes (such as TP53 and ATM) that have been reported from previous analyses[23,24]. More importantly, the increased power provided by the extended-gene annotation allowed us to detect genes that would otherwise be missed by an exclusively coding analysis. An example of this is the well-known cancer gene BCL6, which may be associated with patient survival (Fig. 2D and refs. [25-27]).

Interpreting tumor expression profiles using ENCODE networks identifies key regulators

Building on the extended gene annotation, we provide detailed regulatory networks. Specifically, for TF networks, we incorporated both distal and proximal networks by linking TFs to genes, either directly by TF-promoter binding or indirectly via TF-enhancer-gene interactions in each cell type (see suppl.[1]). We then pruned these networks to include only the strongest edges using a signal shape algorithm[28] (see suppl.). In addition, we reconciled all our cell type specific networks to form a generalized pan-cancer network. Similarly, we also defined an RNA-binding protein (RBP) network from eCLIP experiments. (eCLIP is an enhanced CLIP protocol that provides single-nucleotide resolution of the RBPs binding signatures[29]). Compared to imputed networks derived from gene expression or motif analyses, our ENCODE TF and RBP networks provide much more accurate and experimentally based regulatory linkages between functional elements.

ENCODE networks are useful for interpreting gene-expression data from tumor samples. To enable this, we integrated 8,202 tumor expression profiles from TCGA, using a regression-based approach, to systematically search for the TFs and RBPs that most strongly drive tumor-specific expression (see suppl.). For each patient, we tested the degree to which a regulator's activity correlates with its target's tumor-to-normal expression changes. We then calculated the percentage of patients with these relationships in each cancer type and present the overall trends for key TFs and RBPs in Fig. 3A.

As expected, we found that the target genes of MYC are significantly up-regulated in numerous cancer types, consistent with its well-known role as an oncogenic TF[30,31]. We further validated MYC's regulatory effects using knockdown experiments in breast cancer (Fig. 3). Consistent with our predictions, the expression of MYC targets is significantly reduced after MYC knockdown in MCF-7 (Fig. 3B). We then used the regulatory network to investigate how MYC works with other TFs. We first looked at MYC's target genes co-regulated by a second TF, as shown in the triplets in Fig. 3C. In all cancer types, we found that the shared target genes' expressions are strongly positively correlated with MYC, while they showed only limited correlation with the second TF (as determined by partial correlation analysis, see suppl.). We further investigated the exact structure of these regulatory triplets. The most common one is the well-understood feed-forward loop (FFL). In this case, MYC regulates both another TF and a common target of both MYC and that TF (Fig. 3C). Since MYC amplification has been discovered in many cancers, understanding which TFs appear to further amplify its effects may yield insights for efforts aimed at MYC inhibition[31]. Most of the FFLs involve well-known MYC partners such as MAX and MXL1. However, we

[1] Details see section 4.3S for generalized network and 5.1S-5.2S for cell-type specific network

also discovered many involving NRF1. Upon further examination, we found that that the MYC-NRF1 FFL relationships were mostly coherent, i.e., "amplifying" in nature (Fig. 3C ii). We further studied these FFLs by organizing them into logic gates, in which two TFs act as inputs and the target gene expression represents the output[32] (see suppl.). We show that most of these gates follow either an OR or MYC-always-dominant logic gate. Thus, the ENCODE regulatory network not only helps identify key regulators, but also illustrates how these may work in combination.

We analyzed the RBP network in a similar manner to the TF network, finding key regulators associated with cancer (see suppl.). For example, the ENCODE eCLIP profile for the RBP SUB1 has peaks enriched on the 3'UTR regions of genes, and the predicted targets of SUB1 were significantly up-regulated in many cancer types (Fig. 3C). As an RBP, SUB1 has not previously been associated with cancer, so we sought to validate its role. Knocking down SUB1 in HepG2 cells significantly down-regulated its targets (Fig. 3D), and the decay rate of SUB1 targets is significantly lower than those of non-targets (see suppl.). Moreover, we found that up-regulation of SUB1 targets may indicate a poorer patient survival in some cancer types (Fig. 3D).

We further analyzed the overall TF regulatory network by systematically arranging it into a hierarchy (Fig. 4A). Here, TFs are placed at different levels such that those in the middle tend to regulate TFs below them and, in turn, are more regulated by TFs above them[33] (see suppl.). In the hierarchy, we found that the top-layer TFs are not only enriched in cancer-associated genes but also more significantly drive differential expression in model cell types.

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

For the top-tier cell types with numerous TF ChIP-seq experiments, our resource contains cell type specific regulatory networks, which enable direct comparison 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.

In particular, in "tumor-normal pairs," we measured the signed, fractional number of edges changing (which we call the "rewiring index") to study how TF targets change in the oncogenic transformation. In Fig. 5A, we ranked TFs according to this index. In leukemia, well-known oncogenes (such as MYC and NRF1) were among the top edge gainers, while the well-known tumor suppressor IKZF1 is the most significant edge loser (Fig. 5A). Mutations in IKZF1 serve as a hallmark of various forms of high-risk leukemia[34,35]. 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. 5).

In addition to direct TF-to-gene connections, we also measured rewiring using a more complex gene-community model. Here, the targets within the regulatory network were characterized in terms of heterogeneous modules (so called "gene communities"), which come from multiple genes. Instead of directly measuring the changes in a TF's targets between tumor and normal cells, we determined the changes in its gene communities via a mixed-membership model (see suppl.). Similar patterns to the direct rewiring were observed using this model (Fig. 5A) and also in terms of a simpler co-binding approach (see suppl.).

We next tested whether the gain or loss events from normal-to-tumor transitions result in a network that is more or less similar to that in stem cells like H1-hESC. Interestingly, the gainer TF group tends to "rewire away" from the stem cell's regulatory network, while the loser group is more likely to rewire in such a way that it becomes more stem-like.

We found that the majority of rewiring events were associated with noticeable gene-expression and chromatin-status changes, but not necessarily with mutation-induced motif loss or gain events (Fig. 5A). For example, JUND is a top gainer in K562. The majority of its gained targets in tumor cells demonstrate higher gene expression, stronger active and weaker repressive histone modification mark signals, yet few of its binding sites are mutated. This is consistent with previous work that most non-coding risk variants are not well-explained by the current model[36]. With a few notable exceptions (see suppl.), we found a similar trend for the rewiring events associated with JUND in liver cancer and, largely, for other factors in a variety of cancers. On a related note, we organized the cell type specific networks into hierarchies, as shown in Fig. 4B. Specifically, in blood cancer, the more mutationally burdened TFs sit at the bottom of the hierarchy, whereas the TFs more associated with driving cancer gene expression changes tend to be at the top.

Step-wise prioritization scheme pinpoints deleterious SNVs in cancer

Summarizing the above, our companion resource consists of annotations in Fig. 1 and 6: (1) a BMR model with a matching procedure for the relevant functional genomics data and a list of regions with higher-than-expected mutational burdens in a diverse selection of cancers; (2) accurate and compactly defined enhancer and promotor annotation that is based on integrating many functional assays, including STARR-seq; (3) enhancer-target-gene linkages and extended gene neighborhoods that are obtained by integrating Hi-C and multi-histone-mark experiments; (4) tumor-normal differential expression, chromatin, and regulatory changes; (5) TF regulatory networks, both merged and cell type specific, based on both distal and proximal regulation; (6) for each TF, its position in the network hierarchy and its rewiring status; and (7) an analogous but less-developed network for RBPs. All the resources mentioned above are available online through the ENCODE website as simple flat files and computer codes (see suppl.).

Collectively, these resources allow us to prioritize key genomic features associated with oncogenesis. Our prioritization scheme is schematized as a workflow in Fig. 6A. We first search for key regulators that are frequently rewired, located in network hubs, sit at the top of the hierarchy, or significantly drive expression changes in cancer. We then prioritize functional elements associated with these regulators, are highly mutated in tumors, or undergo large changes in gene expression, TF binding, or chromatin status. Finally, on a nucleotide level, by estimating their ability to disrupt or introduce specific binding sites, we pinpoint impactful SNVs.

Small-scale validation experiments on prioritized genomic elements and variants

To demonstrate the utility of the ENCODE resources, we instantiated our 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 experiments to validate their regulatory effects (Fig. 3D). 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 the wild-type in multiple biological replicates.

One particularly interesting example, illustrating the value of ENCODE data integration, is in an intronic region of CDH26 in chromosome 20 (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 confirmed by STARR-seq (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[37]. We further observed strong binding of many TFs in this region in MCF-7. Motif analysis predicts that the particular mutation from a breast cancer patient significantly disrupts the binding affinity of several TFs, such as FOSL2, in this region (Fig. 6D). Luciferase assays demonstrated that this mutation introduces a 3.6-fold reduction in expression relative to the wild-type, indicating a strong repressive effect on enhancer functionality.

Conclusion

This study highlights the value of ENCODE data as an aid to interpreting cancer genomes. It presents the EN-CODEC companion resource, which tailors the ENCODE annotation to cancer. This has three parts: 1) cancer-specific BMR models with significantly increased accuracy; 2) compact annotations that by maximize statistical power for recurrent-mutation detection; and 3) various regulatory networks and hierarchies for both pan-cancer and cancer-specific studies.

A key caveat related to our resource concerns the rewiring in cell type specific networks. The utility of these networks is based on associating them to particular cancer types and then pairing a specific cancer network with a composite normal. Both of these "correspondences" are approximate. Nevertheless, we feel that the EN-CODEC networks currently provide the best available view of the regulatory changes in oncogenesis. No other system has this scale of TF-ChIP data. Moreover, the heterogeneous nature of cancer means that even tumor cells from a given patient usually show distinct molecular, morphological, and genetic profiles[38]. It will be difficult to obtain a "perfect" match even from real tumor and normal tissues taken from a single patient.

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 in BMR estimation with additional genomic features; we expect that this accuracy will increase further still with more features. We successfully formed compact annotations and regulatory networks for model systems 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 greater numbers of TFs. (In fact, the re-wiring formalism actually provides a way of selecting candidate TFs and cell types.) This will give us a greater sense of which regulators 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 linking it with the large cohorts currently being assembled (e.g., PCAWG, pancaner.info) will more fully utilize EN-CODEC and provide even greater value.

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.

Reference

1 Cai, Q. *et al.* Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat Genet* **46**, 886-890, doi:10.1038/ng.3041 (2014).

2 Jager, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* **6**, 6178, doi:10.1038/ncomms7178 (2015).

3 Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* **46**, 1258-1263, doi:10.1038/ng.3141 (2014).

4 Cancer Genome Atlas Research, N. *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378-384, doi:10.1038/nature21386 (2017).

5 Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944, doi:10.1016/j.cell.2014.06.049 (2014).

6 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).

7 Torchia, J. *et al.* Integrated (epi)-Genomic Analyses Identify Subgroup-Specific Therapeutic Targets in CNS Rhabdoid Tumors. *Cancer Cell* **30**, 891-908, doi:10.1016/j.ccell.2016.11.003 (2016).

8 Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827-841, doi:10.1093/nar/gks1284 (2013).

9 Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-1797, doi:10.1101/gr.137323.112 (2012).

10 Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55-60, doi:10.1038/nature22992 (2017).

11 Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-1165, doi:10.1038/ng.3101 (2014).

12 Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* **16**, 213-223, doi:10.1038/nrg3890 (2015).

13 Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360-364, doi:10.1038/nature14221 (2015).

14 Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504-507, doi:10.1038/nature11273 (2012).

15 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

16 Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat Genet* **47**, 710-716, doi:10.1038/ng.3332 (2015).

17 Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. & Gerstein, M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res* **43**, 8123-8134, doi:10.1093/nar/gkv803 (2015).

18 Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* **10**, 1108-1115, doi:10.1038/nmeth.2651 (2013).

19 Jacobsen, A. *et al.* Analysis of microRNA-target interactions across diverse cancer types. *Nat Struct Mol Biol* **20**, 1325-1332, doi:10.1038/nsmb.2678 (2013).

20 Mutation, C. & Pathway Analysis working group of the International Cancer Genome, C. Pathway and network analysis of cancer genomes. *Nat Methods* **12**, 615-621, doi:10.1038/nmeth.3440 (2015).

21 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

22 O'Connor, M. L. *et al.* Cancer stem cells: A contentious hypothesis now moving forward. *Cancer Lett* **344**, 180-187, doi:10.1016/j.canlet.2013.11.012 (2014).
23 Zenz, T. *et al.* Detailed analysis of p53 pathway defects in fludarabine-refractory chronic lymphocytic leukemia (CLL): dissecting the contribution of 17p deletion, TP53 mutation, p53-p21 dysfunction, and miR34a in a prospective clinical trial. *Blood* **114**, 2589-2597, doi:10.1182/blood-2009-05-224071 (2009).
24 Guarini, A. *et al.* ATM gene alterations in chronic lymphocytic leukemia patients induce a distinct gene expression profile and predict disease progression. *Haematologica* **97**, 47-55, doi:10.3324/haematol.2011.049270 (2012).
25 Jantus Lewintre, E. *et al.* BCL6: somatic mutations and expression in early-stage chronic lymphocytic leukemia. *Leuk Lymphoma* **50**, 773-780, doi:10.1080/10428190902842626 (2009).
26 Cardenas, M. G. *et al.* The Expanding Role of the BCL6 Oncoprotein as a Cancer Therapeutic Target. *Clin Cancer Res* **23**, 885-893, doi:10.1158/1078-0432.CCR-16-2071 (2017).
27 Capello, D. *et al.* Identification of three subgroups of B cell chronic lymphocytic leukemia based upon mutations of BCL-6 and IgV genes. *Leukemia* **14**, 811-815 (2000).
28 Cheng, C., Min, R. & Gerstein, M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* **27**, 3221-3227, doi:10.1093/bioinformatics/btr552 (2011).
29 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514, doi:10.1038/nmeth.3810 (2016).
30 Dang, C. V. MYC on the path to cancer. *Cell* **149**, 22-35, doi:10.1016/j.cell.2012.03.003 (2012).
31 McKeown, M. R. & Bradner, J. E. Therapeutic strategies to inhibit MYC. *Cold Spring Harb Perspect Med* **4**, doi:10.1101/cshperspect.a014266 (2014).
32 Wang, D. *et al.* Loregic: a method to characterize the cooperative logic of regulatory factors. *PLoS Comput Biol* **11**, e1004132, doi:10.1371/journal.pcbi.1004132 (2015).
33 Cheng, C. *et al.* An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. *Genome Biol* **16**, 63, doi:10.1186/s13059-015-0624-2 (2015).
34 de Rooij, J. D. *et al.* Recurrent deletions of IKZF1 in pediatric acute myeloid leukemia. *Haematologica* **100**, 1151-1159, doi:10.3324/haematol.2015.124321 (2015).
35 Boer, J. M. *et al.* Prognostic value of rare IKZF1 deletion in childhood B-cell precursor acute lymphoblastic leukemia: an international collaborative study. *Leukemia* **30**, 32-38, doi:10.1038/leu.2015.199 (2016).
36 Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343, doi:10.1038/nature13835 (2015).
37 Masterson, L. *et al.* Deregulation of SYCP2 predicts early stage human papillomavirus-positive oropharyngeal carcinoma: A prospective whole transcriptome analysis. *Cancer Sci* **106**, 1568-1575, doi:10.1111/cas.12809 (2015).
38 Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328-337, doi:10.1038/nature12624 (2013).

BIOSAMPLE →

ASSAY ↓

EN-CODEC

		K562	HepG2	A549	MCF-7	HeLa-S3	H1-hESC	Caco-2	HCT116	Panc1	LNCaP	PC-3	PC-9	SK-N-MC	DND-41	SK-N-SH	...
		CML	LIHC	LUAD	BRCA	Cervix	ESC	COAD+READ	PAAD	PRAD			LUAD	SARC	LAML	NB	
Chromatin Accessibility DS	DNase-seq	●	●	●	●	●	●	●	●	●	●	●	●	●		●	
Histone Modifications HM	Histone	12	11	11	5	11	11	3	12	6	1	13	13	7	11	13	
Transcription TX	RNA-seq	●	●	●	●	●	●		●	●		●				●	
	RAMPAGE	●															
RNA-binding Proteins RP	eCLIP	89	70														
shRNA/siRNA Knockdown KD	TF KD	85	61		2												
	RBP KD	234	225														
3D Chromatin Structure 3D	ChIA-PET	3	1		4	2											
	Hi-C	▲		●	▲		▲										
Enhancers SS	STARR-seq	●	▲		●												
Methylation ME	WGBS	●	●		▲		●										
	RRBS	●	●	●	●	●	●										
Replication Timing RT	Repli-seq/chip	●	●	●	●	●	●										
Transcription Factors TF	TF Total	207	95	31	52	59	49										
	TFSS	125	69	23	34	34	32										
	Chromatin Remodeller	31	13	3	7	9	9										
	Cofactor	20	7	3	4	6	3										
	General (GTF)	17	4	2	2	10	5										
	Other	14	2		5												
Cell Line WGS WG	SNV	▲			▲	▲											
	SV	▲			▲	▲											
Patient Cohort PC	Mutation	150	82	197	116			115+45	352	122	197	40	50				
	Expression	173	373	515	1100	546		459+166	178	497	515	259	179				

Cell type-spec. Analysis

528 ENCODE Cell Types → **229** Deduplicated & Selected Human Biosamples

- **143** Composite Normal
- **86** Cancerous (40 Cancer Types)

Depth Approach ("Top-tier")

Breadth Approach ("Encyclopedia")

- **3** Core Assays
- **10** Advanced Assays
- TF ChIP-seq
- WGS & Patient Cohort

Cell type-specific Analysis

Merged Analysis

Merged Analysis

Compact Annotations
- Cis-Regulatory Element ♦♦
  - Enhancer Prediction ♦ (ESCAPE+CASPER)
  - Gene-Enhancer Linkage ♦ (JEME+Hi-C)
- Extended Gene Definitions ♦

Mutation Burden
- BMR Model ♦♦
- Burdened Regions ♦♦

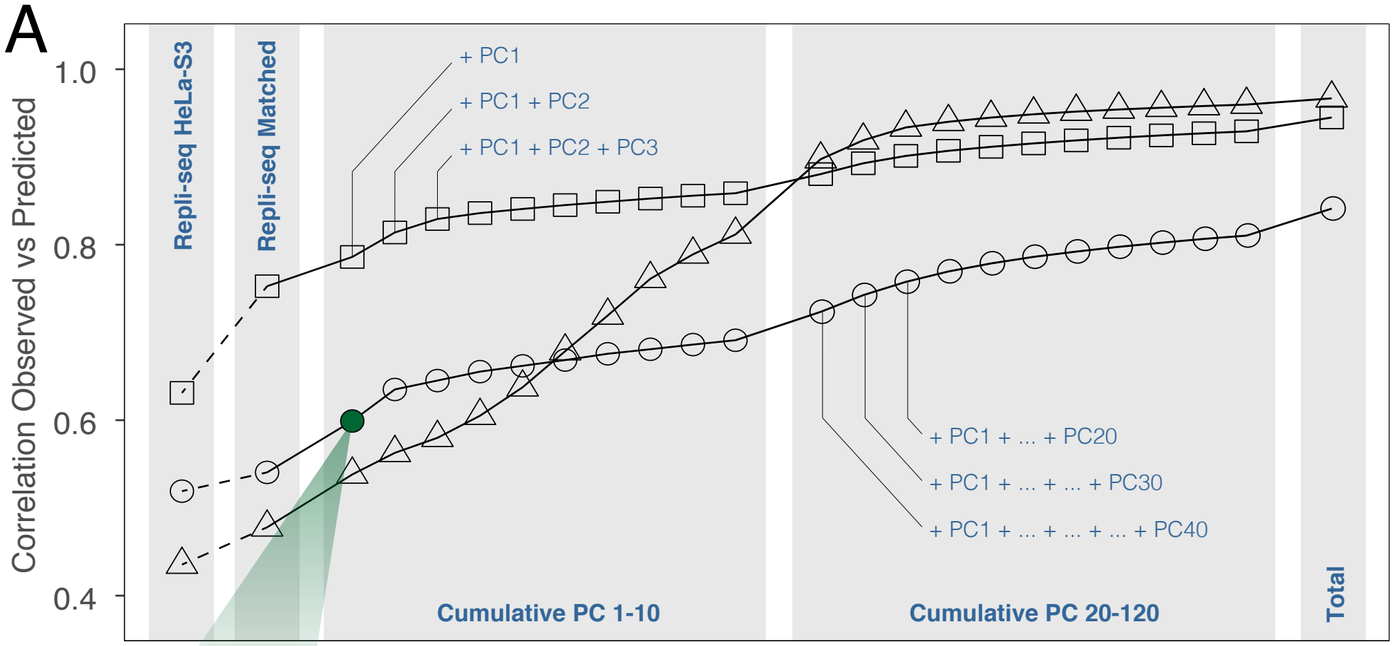
encodec.encodeproject.org

Regulatory Network
- Regulation Power ♦
- Chromatin Changes ♦
- Rewiring ♦
- Hierarchy ♦♦
- Network Motif ♦♦
- Seq Motif Gain/Loss ♦♦

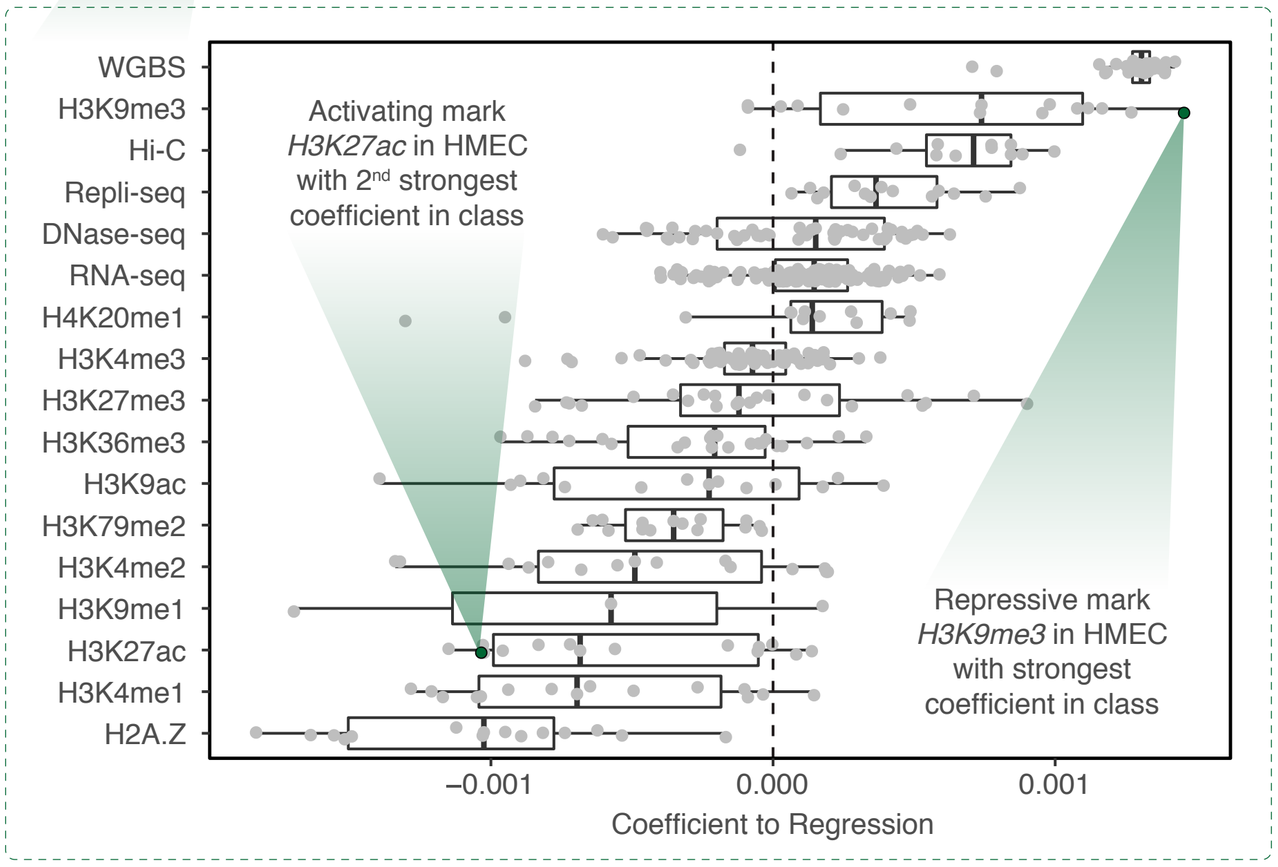
Available from
- ● ENCODE
- ▲ External Resource
- ◆ EN-CODEC

Fig 1

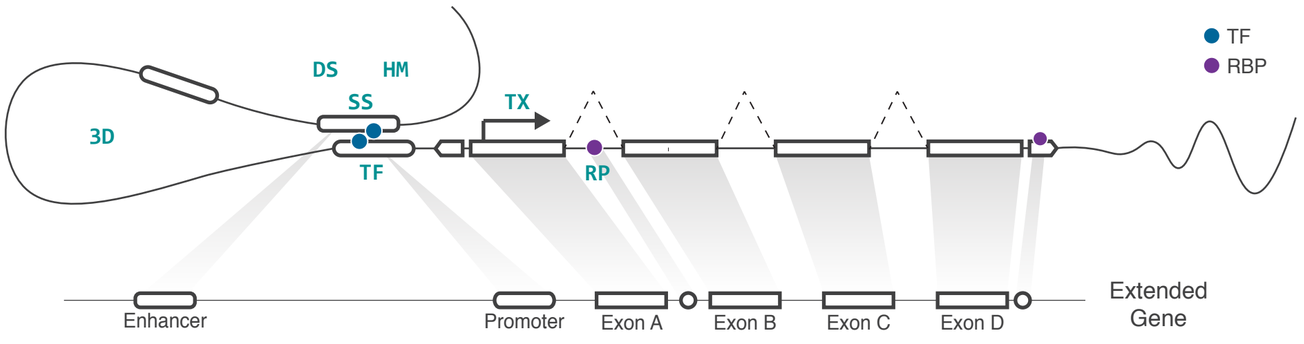
A


Correlation Observed vs Predicted
1.0
0.8
0.6
0.4
Repli-seq HeLa-S3
Repli-seq Matched
+ PC1
+ PC1 + PC2
+ PC1 + PC2 + PC3
+ PC1 + ... + PC20
+ PC1 + ... + ... + PC30
+ PC1 + ... + ... + ... + PC40
Cumulative PC 1-10
Cumulative PC 20-120
Total

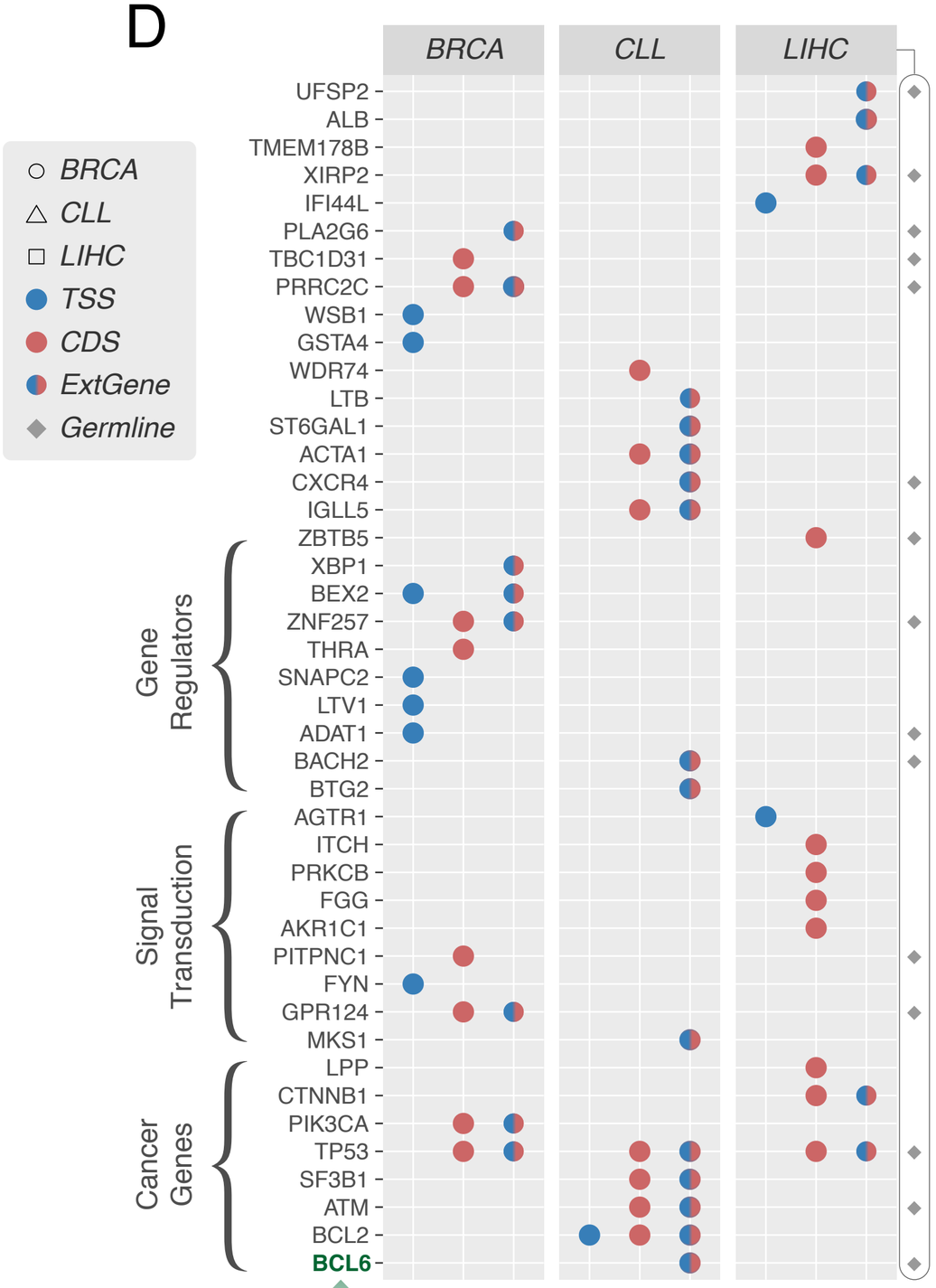
B


WGBS
H3K9me3
Hi-C
Repli-seq
DNase-seq
RNA-seq
H4K20me1
H3K4me3
H3K27me3
H3K36me3
H3K9ac
H3K79me2
H3K4me2
H3K9me1
H3K27ac
H3K4me1
H2A.Z
Activating mark
H3K27ac in HMEC
with 2nd strongest
coefficient in class
Repressive mark
H3K9me3 in HMEC
with strongest
coefficient in class
−0.001
0.000
0.001
Coefficient to Regression

C


DS
HM
SS
TX
3D
TF
RP
TF
RBP
Enhancer
Promoter
Exon A
Exon B
Exon C
Exon D
Extended
Gene

D


BRCA
CLL
LIHC
TSS
CDS
ExtGene
Germline
UFSP2
ALB
TMEM178B
XIRP2
IFI44L
PLA2G6
TBC1D31
PRRC2C
WSB1
GSTA4
WDR74
LTB
ST6GAL1
ACTA1
CXCR4
IGLL5
ZBTB5
XBP1
BEX2
ZNF257
THRA
SNAPC2
LTV1
ADAT1
BACH2
BTG2
AGTR1
ITCH
PRKCB
FGG
AKR1C1
PITPNC1
FYN
GPR124
MKS1
LPP
CTNNB1
PIK3CA
TP53
SF3B1
ATM
BCL2
BCL6
Gene Regulators
Signal Transduction
Cancer Genes

E

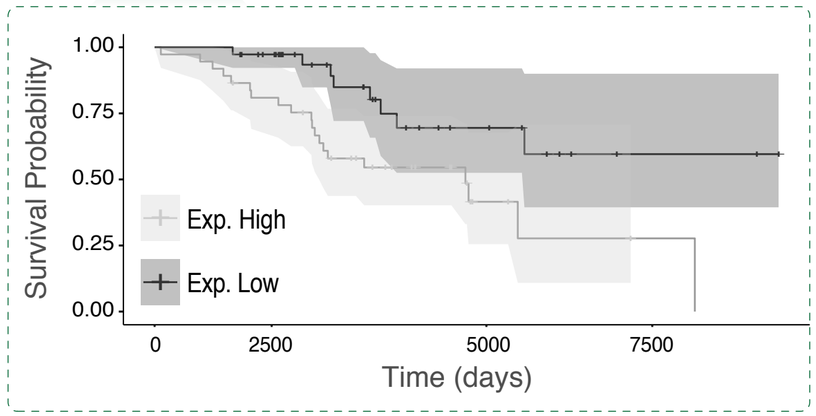

Survival Probability
1.00
0.75
0.50
0.25
0.00
Exp. High
Exp. Low
0
2500
5000
7500
Time (days)

Fig 2

A

TF

RBP

Aggregative Target Expression

UP

DN

BLCA
BRCA
ESCA
GBM
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
OV
PRAD
STAD
THCA
UCEC

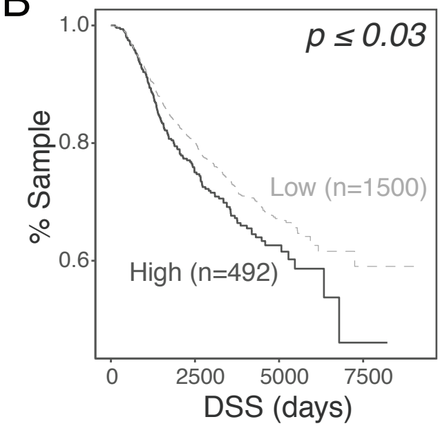
MYC

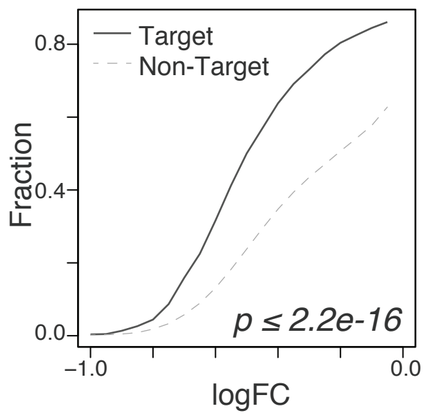
RAD21 NR2F2 FOXK2 YY1 FOXJ2 GLIS2 HLF HMBOX1 RFX1 TEAD4 ESRRA CHAMP1 MYNN TRIM28 HNF4G ATF3 SAP130 TAF1 HBP1 E2F4 ZNF623 STAT2 CTCF IKZF1 TFDP1 KDM6A ZNF121 PAF1 REST E2F1 ELF1

SUB1

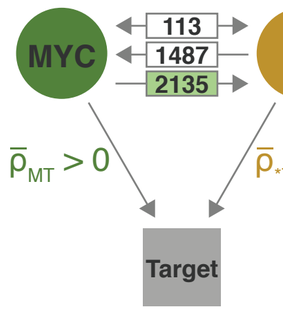
TIA1 LSM11 PUM2 GNL3 GTF2F1 TAF15 DDX55 NONO NPM1 SLBP LARP4 NKRF UPF1 KHDRBS1 SAFB2 HNRNPC HNRNPM CSTF2 HNRNPUL1 DHX30 CSTF2T AARS RBFOX2 EIF3D TARDBP CDC40 RPS5 HNRNPK HNRNPU HNRNPA1 FUBP3

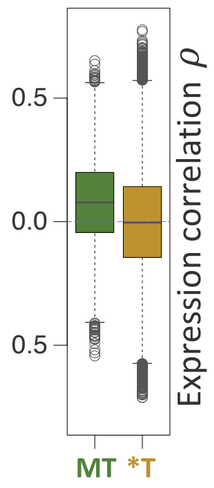
B



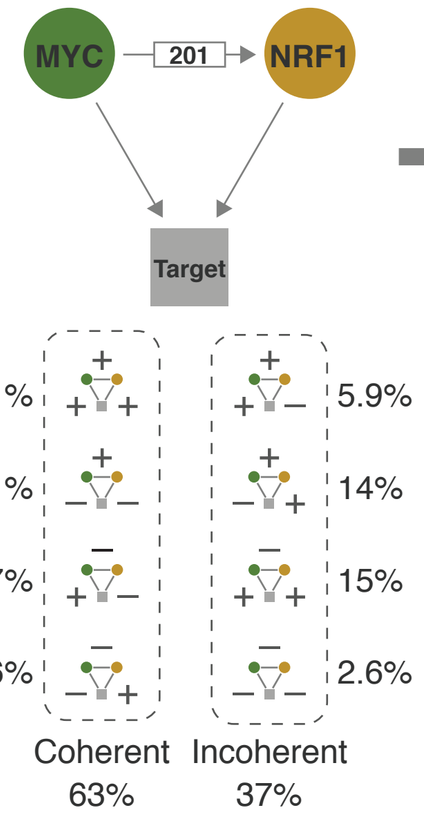


C(i)

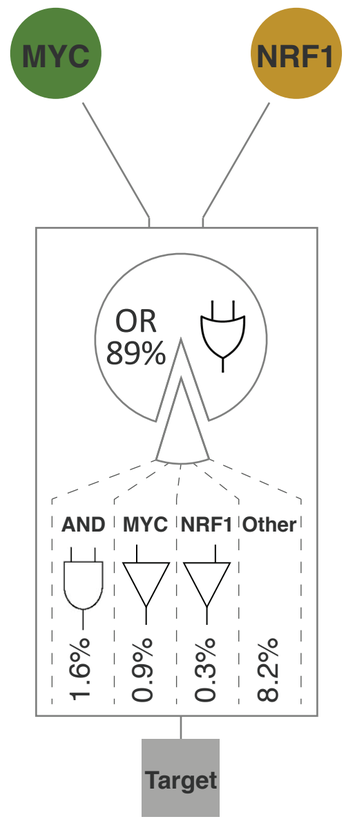




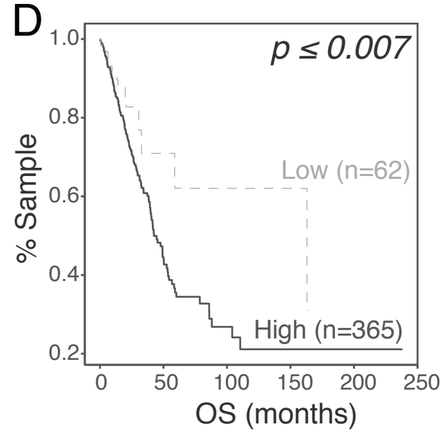
C(ii)



C(iii)



D



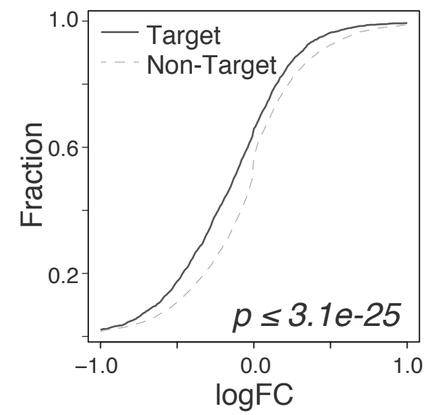
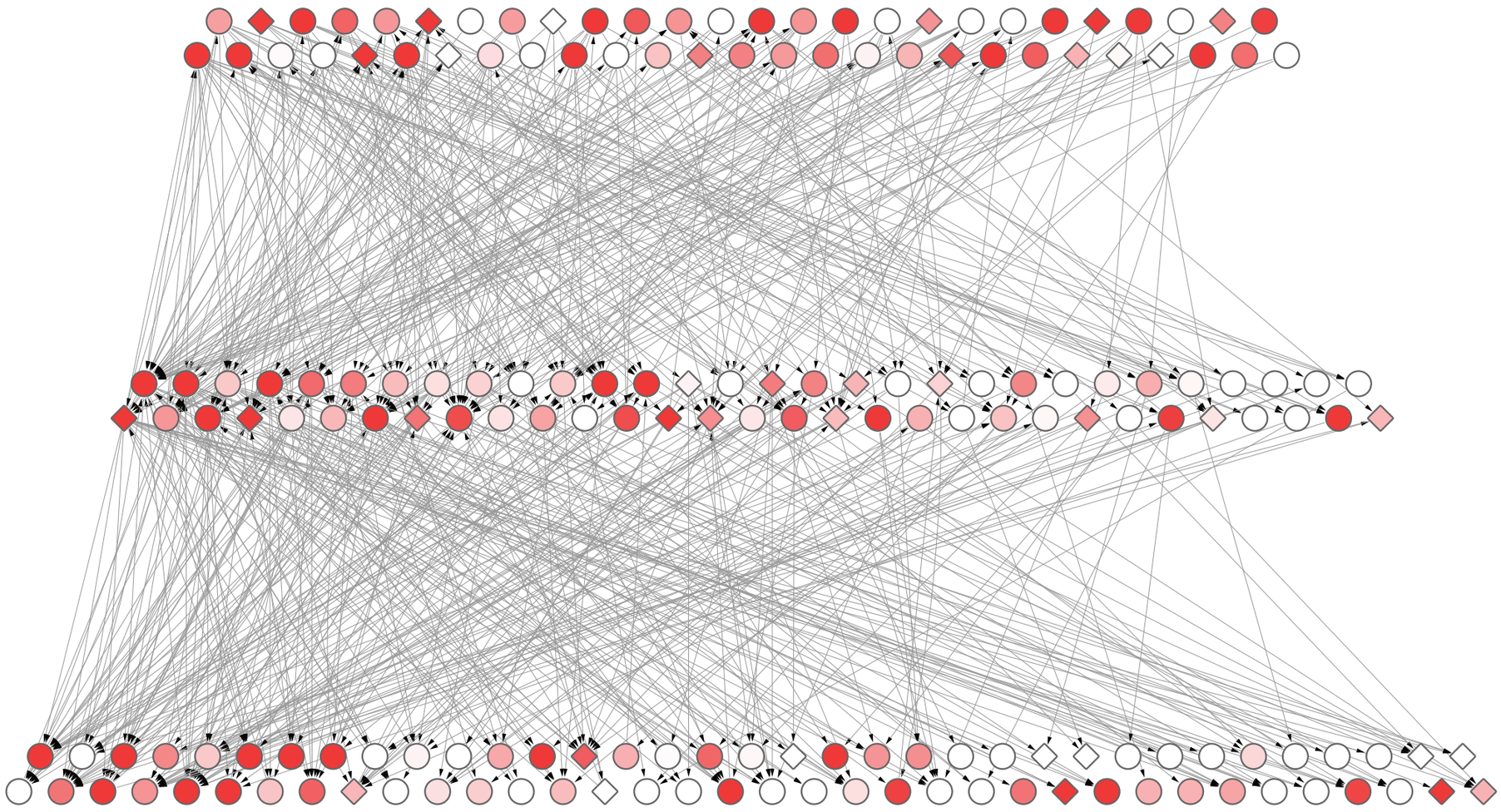
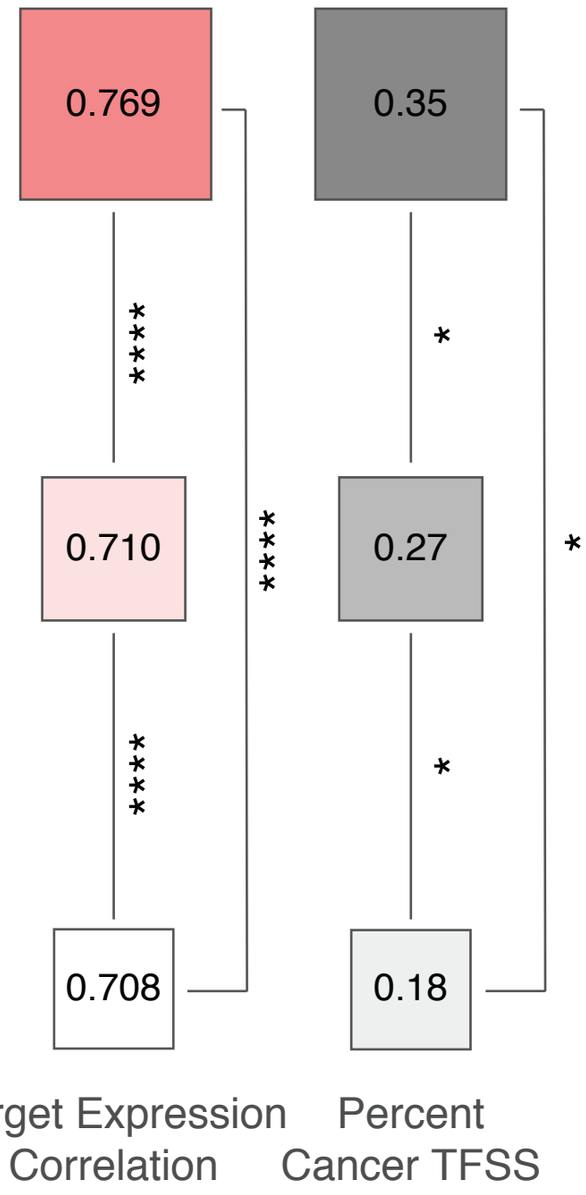


Fig 3

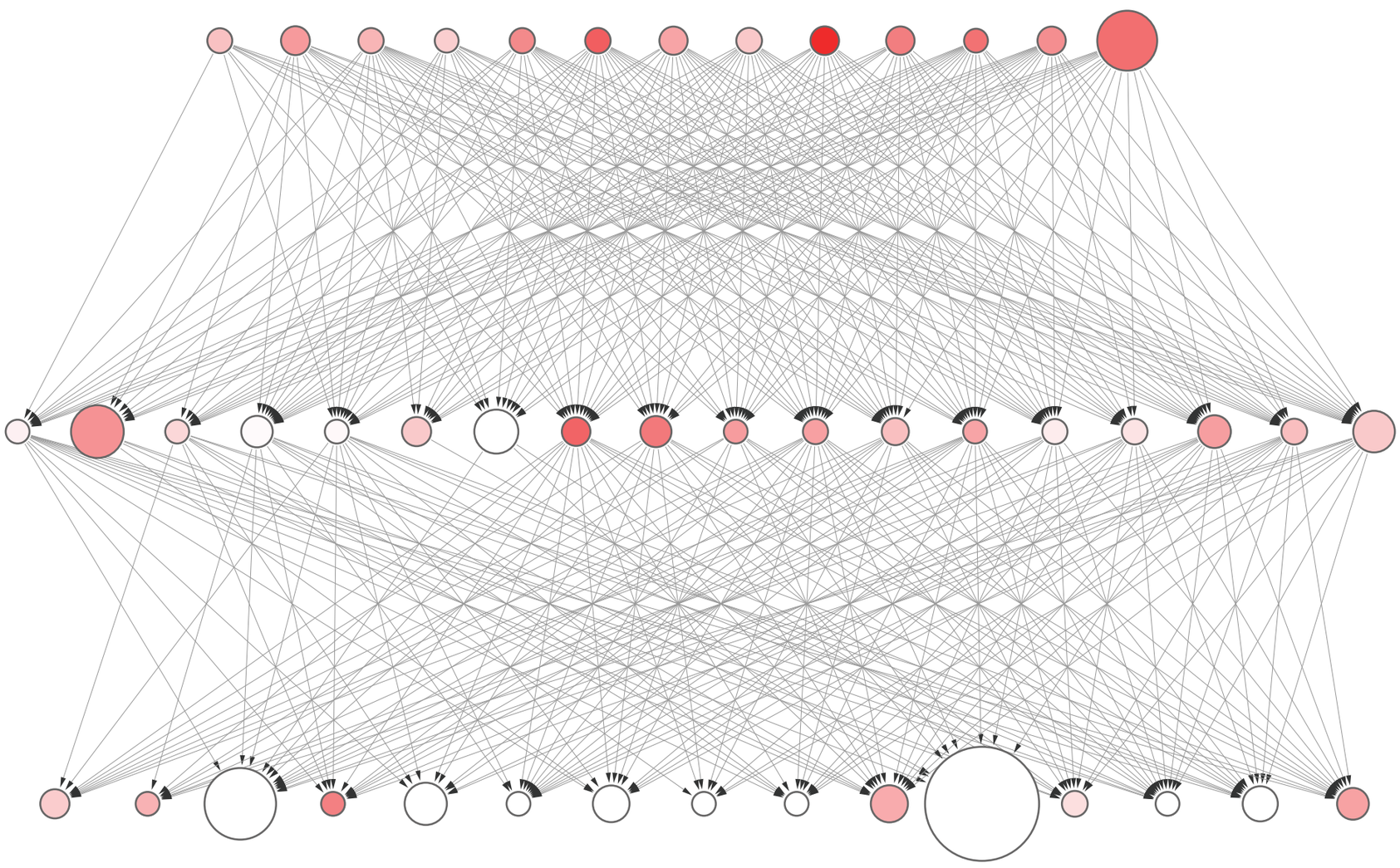
A

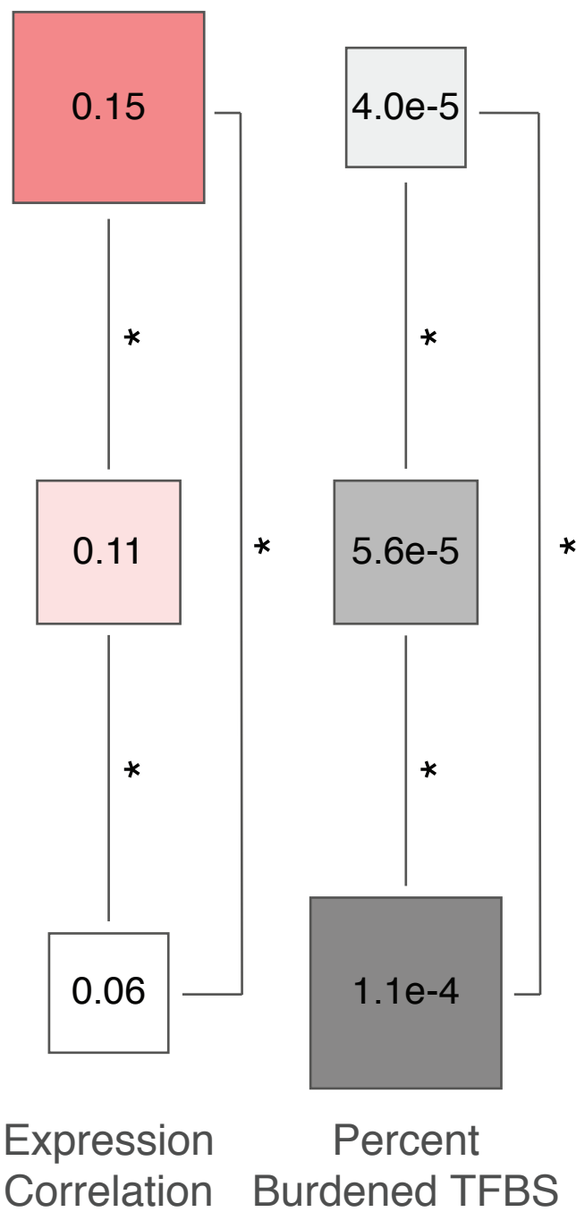




Cancer TFSS

B





TFBS Burden

Fig 4

Loser — TF-Gene Network — Gainer

A

Rewiring in Direct Connection

Rewiring in Gene Community

1
0
1

Proximal
Distal
Prox. + Dist.

CML

IKZF1 PML **NBN** MAZ MLLT1 MAX MXI1 ZNF143 MTA2 HDGF YY1 ZBED1 STAT5A TARDBP **CTCF** CHD2 NFYB **BHLHE40** TBL1XR1 BCLAF1 JUNB SUZ12 SIX5 CBX5 MEF2A CEBPZ RFX5 **YBX1** NR2C2 MAFK USF2 SRF CEBPB ETV6 ZNF274 USF1 FOS ATF3 ELK1 UBTF E2F4 RCOR1 SMC3 NFYA NFE2 REST SPI1 **JUND** RAD21 ETS1 CHD1 ZBTB33 EGR1 EZH2 GABPA ELF1 SMAD5 SP1 ZBTB40 **MYC** **NRF1**

B

Tumor Suppressor Gene (TSG)
Oncogene/TSG
Oncogene
Toward ESC
Neither
Away from ESC
N/A

LUAD LUAD LIHC BRCA

C

Rewiring and Stem Cell (ESC)

Retained
Lost *Toward ESC*
Gained *Away from ESC*

Normal ***Tumor*** ***ESC***

D

JUND in **CML**

log fold change
0
−5

DNase H3K4me3 H3K27ac H3K27me3
Active *Repressive*

# of edges
5000
0
Lost Gained

All
w/ SNVs
w/ SNVs at TFBS
Edges

JUND in **LIHC**

# of edges
1500
0
Lost Gained

Fig 5

A

EN-CODEC

encodec.encodeproject.org

Regulator

- Network Position (Hub, Hierarchy)
- Network Rewiring & Burden
- Gene Expresion

TF HM RP ME DS SS 3D

Element

- Mutational Burden
- Chromatin Structure
- Enhancer Target Linkage
- Target Gene Expression

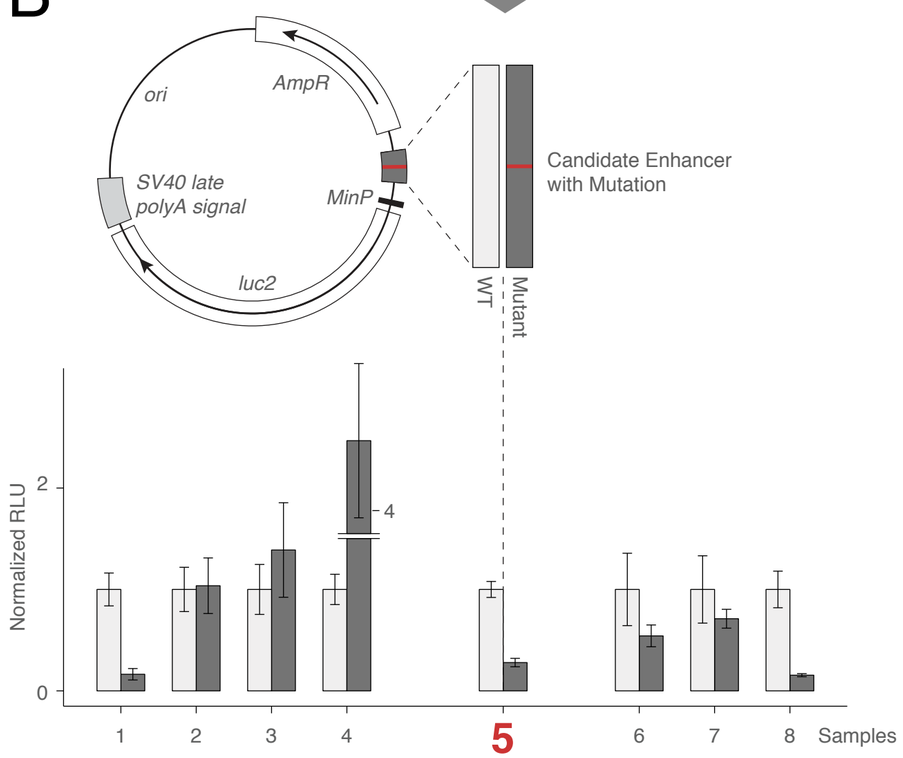
TF HM RP ME DS SS 3D RT TX WG PC

Nucleotide

- Mutation Function
- Motif Impact
- Conservation

TF KD PC

B



C

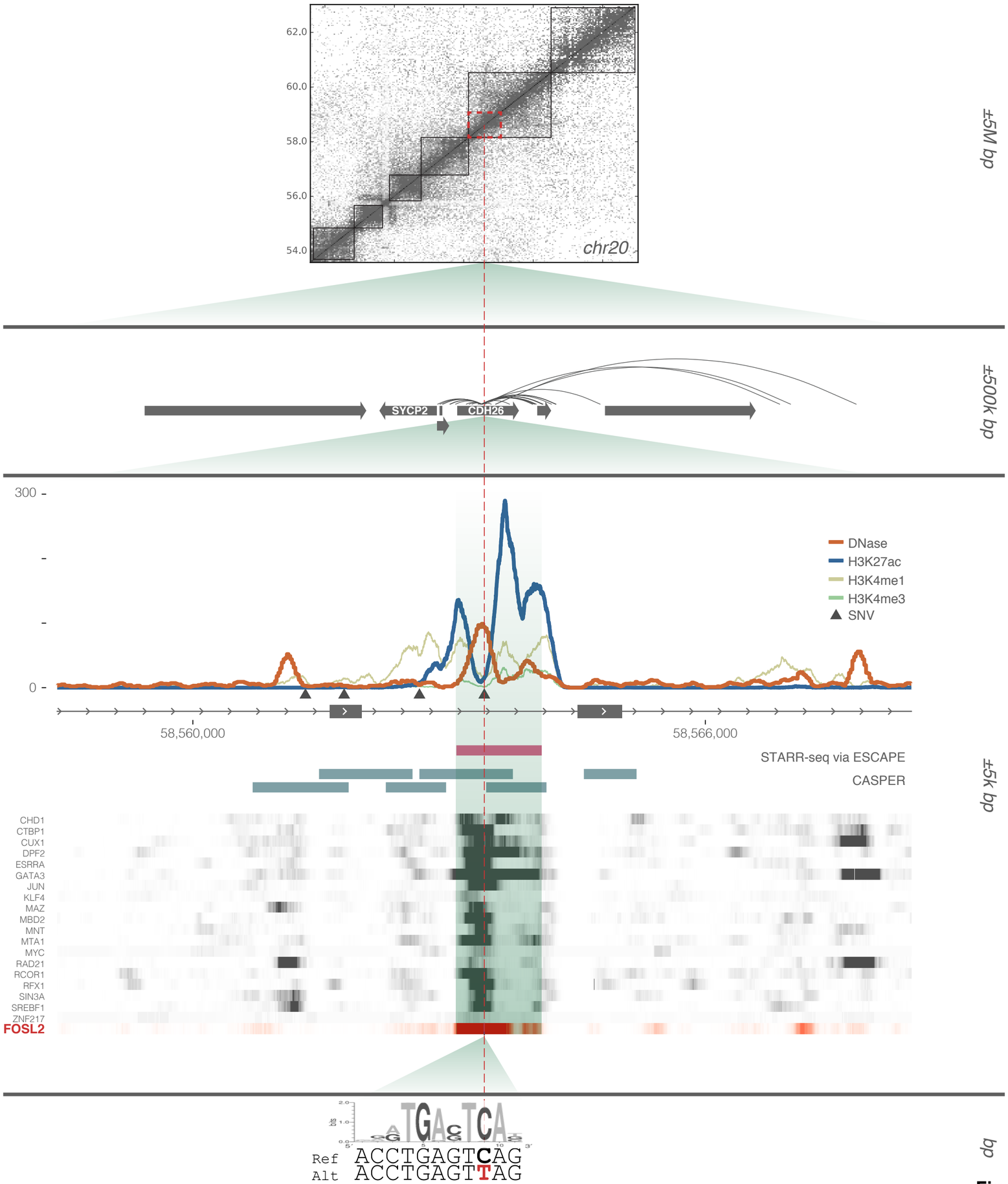


Fig 6