An integrative ENCODE resource for cancer genomics

Introduction

The initial ENCODE release in 2012 started to systematically map the functional elements of the human genome, such as transcription, transcription factor (TF) binding, chromatin accessibility, and histone modifications. The new release of ENCODE data has been dramatically improved since the last release. It does not only considerably broaden the number of cell types for a general annotation resource, but also deeply enlarged the number of advanced assays on several "top-tier" cell types. Hence, we have performed deep integration over tens of functional assays on these cell types to characterize the noncoding genome with great resolution and accuracy that is not possible elsewhere.

Our deep integrative annotation resource has several advantages over previous general annotations. First, it provides detailed tissue-specific proximal regulatory elements from TF and RNA binding proteins (RBPs) from thousands of ChIP-seq and eCLIP experiments. Second, it then incorporates many histone marks with novel assays, such as STARR-seq which directly measures the genome-wide enhancer activities, to accurately define core enhancers with great accuracy. Third, it integrates distance, histone, expression, Hi-C, and ChIA-PET data to link these proximal and distal regulatory elements to genes, which we called the extended genes, to facilitate functional interpretation. Finally, it organized such different types of regulator to genes into various regulatory networks to gain a systems-level perspective.

We found that cancer is one of the best applications of our resource since many of cell types are associated with cancer, including those of the blood, breast, liver, and lung (K562, MCF-7, HepG2, A549, see Fig. 1). Therefore, we constructed a customized *ENCODE* companion resource for *C*ancer genomics (which we call EN-CODEC). This resource consists of a set of annotation files and code bundles available online (encodec.encodeproject.org, see suppl.). We have tried to demonstrate that our resource, such as extended gene, can benefit various analyses in cancer research, such as somatic driver detections, GWAS result interpretations, and expression profile stratifications. We further demonstrated that our various types of regulatory networks allow a systematical view of TF and RBP dysregulations. For instance, we can combine ENCODE networks with expression profiles from cancer cohorts like TCGA to illustrate how key regulators drives tumor-specific gene expression profiles and how they collaborate with other regulators. Besides, the ENCODEC networks provide cell-type specific networks in model tumor and normal cells, thereby enabling direct evaluation of potential regulatory changes in oncogenesis. Furthermore, we can use expression and network profiles to directly measure the degree to which an oncogenic transformation moves towards or away from a stem-cell-like state.

Finally, we propose a step-wise prioritizing scheme that highlights key regulators, SNVs, and SVs associated with cancer progression. We validate the functional impact of these prioritized regulatory elements and variants using focused experiments such as TF/RBP knockdowns, luciferase assays, and CRISPR-engineered deletions. Such prioritization serves as an illustration of how our new EN-CODEC resource can immediately be used to help analyze existing cancer mutation data and cancer-associated gene expression.

Moved (insertion) [1]

Moved (insertion) [2]

Deleted: .First, it

Deleted: ed

Deleted: studied using the original assays. As such, the main ENCODE encyclopedia aims to utilize these to provide

Deleted: that is applicable across many cell types. Second, ENCODE

Deleted: (e.g., STARR-seq, Hi-C, ChIA-PET, eCLIP and RAMPAGE).

Formatted: Indent: First line: 0"

Deleted: M

Deleted: more

Deleted:

Deleted: Such rich functional assays and annotation resources in the new ENCODE release allow us to deeply characterize these non-coding regions and

Deleted: to

Deleted: along with 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 many ways. For example, various types of genomic features, such as replication timing and histone modification, are broadening our understanding of the underlying mechanisms of single nucleotide variations (SNVs) and structural variations (SVs) in cancer progression. Secondly, they enable researchers to systematically define regulatory elements, such as enhancers and binding sites for transcription factors (TFs) and RNA binding proteins (RBPs), thereby greatly facilitating our ability to interpret the functional impacts of variants in non-coding regions[6,8-11]. Finally, ENCODE data and other genomic data sets have been used to link non-coding elements and to organize them into regulatory networks, which can be used to gain a systems-level perspective of cancer[18-20]. ... [1]

Moved up [1]: The new release of ENCODE data has been dramatically improved since the last release.

Moved up [2]: First, it considerably broadened the number of cell types studied using the original assays. As such, the main ENCODE encyclopedia aims to utilize these to provide a general annotation resource that is applicable across many cell types. Second, ENCODE deeply enlarged 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 cancer, including th ... [2]

Deleted:

Deleted: i

Deleted: ... [3]

Deleted: the significantly expanded ChIP-seq and more recent assays such as eCLIP and Hi-C enables us to ... [4]

Deleted: ,

Deleted: ing

The breadth of ENCODE3 data

ENCODE3 significantly expanded the number of genomic data cross various cell types (see suppl.), many of which have been shown to be associated with mutational processes for both SVs and SNVs. For example, we found that SVs in K562 are enriched in regions with higher [JZ2STL]. In addition, many previous efforts used functional genomics data (such as replication timing and histone modification data) to estimate BMR in various cancer types. Hence, we demonstrate how the extensive ENCODE data can improve BMR estimation background mutation rates in a wide range of cancer types through the commonly used negative regression models. We performed forward selection on these features to select the best combination, and our results demonstrate that BMR accuracy can be significantly improved.

An extended gene annotation and validations through CRISPR deletions

As noted, a second advantage of leveraging ENCODE data to investigate the function impact of mutations by defining more compact and accurate annotations – a smaller, "core" region (which is enriched for true functional impact) significantly improves key variant detectability. Hence, we tried to restrict our annotation set to 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 with up to 10 histone modification marks to predict enhancers (see suppl.). Using a second algorithm called ESCAPE, we then combined these predictions with our output from processing STARR-seq data (see suppl.). This ensemble-based approach enables us to define a minimal list of enhancers with fewer false-positives.

We also linked the above noncoding annotation to genes, because our current knowledge of disease is typically derived by studying protein-coding genes. Most traditional methods rely solely on the correlation of individual signals, which may result in inaccurate gene linkages. Here, we use a machine learning algorithm that takes the wide variety of histone modification marks and gene expression signatures into consideration (thereby delineating accurate enhancer-target gene linkages) and then combining the results with direct experimental evidence on physical interactions from Hi-C and ChIA-PET. We defined our extended gene annotations as the combination of proximal and distal regulatory regions with exons and demonstrated their value.

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. 2A illustrates the larger number of known cancer-related genes detected in several cancer cohorts, relative to those derived from the traditional approach. 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 joint detection approach allows us to detect genes that would otherwise be missed by an exclusively focusing on coding regions. An example of this is the well-known cancer gene BCL6, which may be associated with patient survival (Fig. 2E and refs. [25-27]).

Secondly, our extended gene definitions include many tissue-specific proximal and distal noncoding regulatory elements, which are relevant to interpreting cancer-associated GWAS variants. To illustrate this, we downloaded all the GWAS SNPs for breast cancer and leukemia in the GWAS Catalog with European ancestry and performed SNP enrichment analysis. As expected, we observed increasing enrichment by adding more annotation categories (defined above) to the conventional CDS and TSS regions in both cancer types. It is worth noting that, unlike gene definitions, which are fixed across different cell types, the comprehensive ENCODE data allow us to build a highly dynamic extended gene definition that is unique to a particular tissue. We can only see the best enrichment of GWAS SNPs in a tissue-matched manner.

Thirdly, it has been shown that mutation status in a specific annotation category, such as enhancers, can be used to differentiate expression profiles of cancer patients. If our noncoding elements are precise and their gene linkages are accurate, we hypothesize that greater power is provided for such analyses. Hence, we tried to stratify gene expressions based on various types of annotations in liver cancer. As expected, combining mutational profiles from large cohorts, our tissue-matched extended gene definition can explain the expression differences of a larger number of genes. One example is the splicing factor SRSF3, which has been proven to affect liver cancer progression. Its extended gene annotation in HepG2 exhibits greater significance relative to any single annotation category.

Finally, we demonstrated the value dynamics of extended genes by example of oncogene activation in cancer and its validation through CRISPR based validations. ERBB4 is a well-known oncogene in many cancer types. Specifically, it is activated in breast cancer (T47D) cells in compare with normal cells (HMEC). While the gene itself does not show gain of copy, we noticed a 130Kb heterozygous deletion about 45Kb downstream to the ERBB4 promoter in T47D cells (Figure). The deletion overlaps with the right-hand boundary of the TAD that encompasses ERBB4 gene (Figure), and through 4C we observed a novel interaction in T47D cells between ERBB4 and a distal region from the next TAD, which is not present in HMEC cells (Figure). We therefore hypothesized that the heterozygous deletion disrupts the insulation of ERBB4 from distal regions and hence activates its expression on one allele. We narrowed down the disrupted TAD boundary by locating a conserved CTCF binding region, and then tested the hypothesis by CRISPR editing with a pair of sgRNA to excise the 86bp sequence on the wild-type allele in T47D cells that contains the CTCF binding motif (Figure) , and check if it further enhances ERBB4 expression. Our result shows that the boundary disruption on the wild-type allele doubles ERBB4 expression (Figure), suggesting that the ERBB4 activation in T47D is at least in part due to the 130Kb deletion that disrupts its insulation.

Key regulator prioritization using ENCODE networks and validations through knockdown experiments

We also provide comprehensive tissue-specific proximal regulatory networks directly from ChIP-Seq and eCLIP experiments, and we reconciled all our cell-type specific networks to form a generalized pan-cancer network. Compared to imputed networks derived from gene expression or motif analyses, our ENCODE TF and RBP networks are built using experimentally-defined regulatory linkages between functional elements (see supplements). We analyzed the overall TF and RBP regulatory network by systematically arranging it into a hierarchy (Fig. 3A). Here, TFs are placed at different levels such that those in the middle tend to regulate TFs below; in turn,

they are more regulated by TFs above[33] (see suppl.). We found that the top-level TFs are not only enriched in cancer-associated genes but also more significantly drive differential expression in model cell types.

Our networks provide valuable means of interpreting gene-expression data from tumor samples. Along these lines, we used a regression-based approach to integrate 8,202 tumor expression profiles from TCGA and systematically search for TFs and RBPs that most strongly drive tumor-specific expression (see suppl.). For each patient, we tested the degree to which a regulators' activities correlate with tumor-to-normal expression changes their respective targets. We then calculated the percentage of patients with these relationships in each cancer type shown in Fig. 3B. These metrics can be used to prioritize key TFs and RBPs in cancer.

As expected, we found that many previously reported cancer-associated TFs show high regulatory potential and are associated with patient survival. For example, we found that MYC targets are significantly up-regulated in numerous cancer types. We therefore validated MYC's regulatory effects using knockdown experiments in breast cancer (Fig. 3). Consistent with our predictions, the expression of MYC targets are significantly reduced after MYC knockdown in MCF-7 (Fig. 3C). Similarly, in the RBP network, we found that SUB1 peaks are enriched on the 3'UTR regions of genes, and the predicted targets of SUB1 were significantly up-regulated in many cancer types (Fig. 3D). SUB1 has not previously been associated with cancer as an RBP, 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 investigated how key regulators can interact with each other during regulatory processes tumors. For TFs, we found that, with the exception of well-known co-regulators MAX and MXL1, NRF1 is the most frequent co-regulator of MYC that forms a forward-feedback loop (FFL). Upon further examination, we found that the 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 expression represents the output32 (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 level (see supplements). 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 region of key oncogenes to initiate their transcription and SUB1, and it binds to the 3UTRs to stabilize such genes on 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 gene expression changes. As expected, the expression of oncogenes (such as MCM7, BIRC5, and ATAD3A) are significantly reduced.

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 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.

In particular, we measured the signed, fractional number of edge changing (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. 5A, we ranked TFs according to the index used to quantify such changes. 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 tumor-to-normal changes in a TF's targets, we measured the changes in its gene communities using 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 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. 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. 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.

Stemness measurement during oncogenic transformation through ENCODE regulatory networks

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 that in stem cells. Hence, we can use expression and network profiles to directly measure the degree to which an oncogenic transformation moves towards or away from a stem-cell-like state. We first collected xxx RNA-seq profiles from ENCODE and used RCA to project it to a lower dimension space. We found that a variety of different stem cells tend to cluster together. Tumor cells move toward these stem cells and away from the normal ones. We further extended our analysis from RNA-seq to both proximal and distal networks and observed a consistent pattern.

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 hypotheses that tumor cells are more similar to stem cells, we used shRNA RNA-seq experiments to measure the perturbations introduced by oncogene TFs. Interestingly, by knocking down such TFs, the expression profiles revert slightly back toward normal state.

Step-wise prioritization scheme pinpoints deleterious SNVs in cancer and validations using luciferase assay

In sum, our companion resource consists of the annotations in Figs. 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 are based on integrating many functional assays, including those STARR-seq experiments; (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. Together, these resources are made available online through the ENCODE website as flat text files and code bundles (see suppl.).

Collectively, these resources allow us to prioritize key genomic features associated with oncogenesis. Our prioritization scheme is schematized as a workflow shown in Fig. 6A. We first search 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 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 single nucleotide level, by estimating their ability to disrupt or introduce specific binding sites, we pinpoint impactful SNVs.

To demonstrate the utility of this ENCODE resource, 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. 3B and 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.

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[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. 6C). Luciferase assays demonstrated that this mutation 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 the noncoding genome. We provided accurate tissue-specific extended gene annotations after integration of thousands of experiments and comprehensive regulatory networks extracted from various of ChIP-seq and eCLIP experiments. We find the one of the best application of our resource is in cancer since there are many of cell types are 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 the EN-CODEC resources currently provide 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[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 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 more accurate 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.

Deleted: ENCODE data as an aid to interpreting cancer genomes

Deleted: present the EN-CODEC companion resource, which tailors the ENCODE resource to cancer, including

Deleted:

Deleted: thousands

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 ENCODEC 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).

Page 1: [1] Deleted	jingzhang.wti.bupt@gmail.com	5/13/18 11:58:00 AM

along with 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 many ways. For example, various types of genomic features, such as replication timing and histone modification, are broadening our understanding of the underlying mechanisms of single nucleotide variations (SNVs) and structural variations (SVs) in cancer progression. Secondly, they enable researchers to systematically define regulatory elements, such as enhancers and binding sites for transcription factors (TFs) and RNA binding proteins (RBPs), thereby greatly facilitating our ability to interpret the functional impacts of variants in non-coding regions[6,8-11]. Finally, ENCODE data and other genomic data sets have been used to link non-coding elements and to organize them into regulatory networks, which can be used to gain a systems-level perspective of cancer[18-20].

The new release of ENCODE data has been dramatically improved since the last release. First, it considerably broadened the number of cell types studied using the original assays. As such, the main ENCODE encyclopedia aims to utilize these to provide a general annotation resource that is applicable across many cell types. Second, ENCODE deeply enlarged 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 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 deeply characterize these non-coding regions and to 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 code bundles available online (encodec.encodeproject.org, see suppl.).

In particular, with comprehensive types of assay on several model cell types, the ENCODE3 release enables us to provide a comprehensive and tissue-specific annotations on several cell types. For example, in several well-known cancer cell types, we first integrate thousands of ChIP-seq and eCLIP experiments to define proximal regulatory regions. We then incorporate many histone marks with novel assays, such as STARR-seq which directly measures the genome-wide enhancer activities, to accurately define core enhancers and use Hi-C and ChIA-PET data for accurate enhancer-gene linkage prediction. Consequently, the combination of proximal and distal regulatory elements and their accurate gene linkages, which we called the

Page 1: [2] Moved to page 1 (Move #2)	jingzhang.wti.bupt@gmail.com	5/13/18 11:11:00 AM

First, it considerably broadened the number of cell types studied using the original assays. As such, the main ENCODE encyclopedia aims to utilize these to provide a general annotation resource that is applicable across many cell types. Second, ENCODE deeply enlarged 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 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 deeply characterize these non-coding regions and to 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 code bundles available online (encodec.encodeproject.org, see suppl.).

Page 1: [3] Deleted	jingzhang.wti.bupt@gmail.com	5/13/18 12:00:00 PM

In addition,

Page 1: [4] Deleted	jingzhang.wti.bupt@gmail.com	5/13/18 12:00:00 PM

the significantly expanded ChIP-seq and more recent assays such as eCLIP and Hi-C enables us to accurately construct