An integrative ENCODE resource for cancer genomics

Introduction

The 2012 ENCODE release provided RNA-seq, histone and transcription factor (TF) ChIP-seq, and DNase-seq over several cell lines to annotate the noncoding regions in the human genome. The current release broadens the number of cell types for these assays and considerably expands available tissue data. It also greatly increased the depth coverage of assays by adding novel assays, such as STARR-seq, Hi-C, and eCLIP. The integration over a massive number of assays provides an unparalleled opportunity to develop compact and accurate annotations in a tissue-specific manner, which is particularly useful for interpreting genomic variants associated with disease. Deep integration over many assays also allows us to connect many of the regulators and non-coding elements into elaborate networks, including proximal (TF/RBP-gene) and distal ones (enhancer-gene or TF-enhancer-gene).

Hence, focusing on several data-rich ENCODE cell types, we performed deep integration over tens of functional assays to deeply characterize the noncoding genome, which may serve as a valuable resource for disease studies. Our resource is particularly well suited to studying cancer for several reasons. First, many of the most data-rich cell types are associated with 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, the wealth of ENCODE data can be used to measure epigenetic remodeling and cell-state transitions, which are implicated in oncogenesis. Lastly, high-quality regulatory networks can be reliably constructed 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 initiation and progression.

Therefore, we present an integrative ENCODE companion resource for Cancer genomics (ENCODEC). This resource consists of various annotations, networks, and code bundles available online. It allows us to prioritize key regulators, non-coding elements and variants in relation to cancer. These prioritized elements tend to be burdened with germline and somatic mutations, or to sit at central positions in the regulatory network, or to be associated with large epigenetic, expression, or distal interaction changes. We find that this prioritization uncovers interesting interactions between the well-known oncogene TF MYC and the RNA binding protein (RBP) SUB1 that had not been known before. Finally, it shows how the overall chromatin and epigenetics in a cell changes, moving the cell into a more stem-like state.

The breadth and depth of the ENCODE resource

Our work takes advantage of the breadth and depth of the ENCODE resource and customizes it for cancer research. First, the somatic mutation process can be influenced by numerous confounders. ENCODE contains more than 2,000 distinct types of epigenetic and replication

timing data sets. Aggregating these together in a simple model, one can predict background mutation rates (BMRs) for often highly heterogeneous tumors more accurately than a smaller number of features. As seen in Fig. 1, BMR estimation accuracy even continues to improve after even 15 or 20 features are added. Conversely, one can aggregate across assays within a particular cell type to uncover the mutational mechanisms underlying SVs. For instance, once can aggregate the histone markers across structural variants released by ENCODE, which are called by integrating various types of assays (see. Suppl. sect. xxx). Interestingly, we found that K562 breakpoints are associated with H4K20me1, which is an activating histone marker only in K562, but not in other cell types.

We also utilized the depth of the ENCODE data to provide compact and accurate annotations with superior properties relative to other annotations (see suppl. sect. xxx). For example, we explored the full catalogue of ENCODE eCLIP experiments to systematically define the post-transcription regulome with noticeably improved resolution and precision over previous efforts (see suppl. sect. xxx). Additionally, in several well-known cancer cell types, we developed a match filter based algorithm to incorporate a large battery of histone marks with chromatin accessibility data for better enhancer predictions. We the further combined STARR-seq data, which directly measures the genome-wide enhancer activities, to accurately define core enhancers. We then incorporated Hi-C and ChIA-PET data to make accurate enhancer-gene linkage predictions.

An extended gene annotation and its applications

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).

First, 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 increasing GWAS SNP enrichment and the number of included annotations (Fig 2C). We note that, in contrast to unified gene definitions that are identical across different cell types, tissue-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 (Fig 2C).

Second, we used the extended gene annotations 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 cancer cohorts, relative to those derived from the coding regions or promoter sites. 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\}. More importantly, this joint detection approach allowed us to detect genes that would otherwise be missed by exclusively focusing on coding regions. As an example, we identified the well-known cancer gene BCL6, which may be associated with patient survival (Fig. 2B and refs. 1-3).

Moved (insertion) [1]

Deleted: 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]. ... [1]

Deleted: 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 accu ... [2]

Moved up [1]: 2A).

Deleted: For example, using matched replication timin ... [3]

Deleted: every possible nucleotide in the genome grea ... [4]

Formatted: Font color: Black

Deleted: Traditional methods for linking rely solely on ... [5]

Formatted: Font color: Black

Deleted: analyses

Formatted: Font color: Black

Deleted: that have been reported from previous analyses[23,24].

Formatted: Font color: Black

Deleted: the increased power provided by the extended ... [6]

Formatted: Font color: Black

Deleted: an

Formatted: Font color: Black

Formatted: Font color: Black

Deleted: analysis. An

Formatted: Font color: Black

Deleted: of this is

Formatted: Font color: Black

Deleted: 2E

Formatted: Font color: Black

Deleted: [25-27]).

Formatted: Font color: Black

Third, our extended gene annotation can provide better stratification of gene expression from mutational signals in cancer patients compared to single annotation categories. For instance, we combined the mutational and expression profiles from large cohorts, such as TCGA, and found that mutational status in our extended gene definition can explain the expression differences for a larger number of genes than other annotations, such as annotations of coding sequences (CDS). One example of the explanatory potential of the extended gene is seen for the splicing factor SRSF3, which has been shown to affect liver cancer progression \{cite}. In HepG2, aggregating mutational burden within its extended gene annotation exhibits greater significance relative to gene expression, compared to any single annotation category (p=xxx, one sided Wilcoxon test).

Finally, we found an example of how an SV introduced extended gene change that may lead to oncogene activation. ERBB4 is a well-known oncogene in many cancer types \{cite}. (Fig. 2D). We identified a 130Kb heterozygous deletion (~ 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 this heterozygous deletion disrupts the insulation of ERBB4 from distal regions, thereby activating its allele-specific expression. We tested this hypothesis through CRISPR editing, by excising an 86bp sequence from the wild-type allele in T47D cells. This excision resulted in elevated ERBB4 expression (as measured by PCR). Our results suggest that ERBB4 activation in T47D may at least in part be due to the 130 kb deletion that disrupts its insulation.

Leveraging ENCODE networks to prioritize key regulators

Building on the extended gene annotation, we constructed detailed regulatory networks. Specifically, we incorporated both distal and proximal networks by linking TFs to genes. This was accomplished either directly by TF-promoter binding or indirectly via TF-enhancer-gene interactions in each cell type (see suppl. sect. xxx). We then pruned these networks to include only the strongest edges using a signal shape algorithm[1] (see suppl.). In addition, we reconciled 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[2]. Compared to imputed networks derived from gene expression or motif analyses, our ENCODE TF and RBP networks provide experimentally based regulatory linkages between functional elements.

We analyzed the overall regulatory network by systematically arranging it into a hierarchy (Fig. 3A). Here, regulators are placed at different levels such that those in the middle tend to regulate regulators below them and, in turn, are more regulated by regulators above them[3] (suppl. sect. xxx). In this hierarchy, we found that the top-layer TFs are not only enriched in cancer-associated genes (P=xxx, Fisher's exact test) but also more significantly drive differential expression in model cell types (P=xxx, one sided Wilcoxon Test).

Our networks also enable gene-expression analyses in tumor samples. We used a regression-based approach to systematically search for the TFs and RBPs that most strongly drive tumor-normal differential gene expression (suppl. Sect. xxx). 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.

Deleted: Interpreting tumor expression profiles using ENCODE networks identifies key regulators

Deleted: provide

Deleted: for TF networks,

Deleted: ,

Deleted: [1]

Comment [LU1]: Reconciled is fine. Merged or unified could also work as synonyms.

Deleted: [28] (see suppl.). In addition, we reconciled all

Deleted: (

Deleted: [29]

Deleted:).

Deleted: much more accurate and

Deleted: ENCODE

Deleted: are useful for interpreting

Deleted: data from

Deleted: To enable this, we integrated 8,202 tumor expression profiles from TCGA, using

Deleted: ,

Deleted: specific

Deleted: see

Deleted: .).

Deleted: regulator's

Deleted: target's

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[4,5]. 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 analyzed the RBP network in a manner that was similar to the TF network, and found 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. 3D). 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 lower than those of non-targets (see suppl.). Moreover, we found that up-regulation of SUB1 targets may lead to decreased patient survival in some cancer types (Fig. 3D).

We then used the regulatory network to investigate how these prioritized key regulators interact with other genes. For TFS, we first looked at how MYC's target genes are co-regulated by a second TF. These three-way co-regulatory relationships are shown 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 regulatory control pathways of these triplets. The most common pattern 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[5]. Most of the FFLs involve well-known MYC partners such as MAX and MXL1. However, we 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[6] (see suppl.). We found 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.

Similarly, with respect to RBPs, we found that the top co-regulatory partner of SUB1 is MYC (SUB1 is a direct target of MYC in many cell types, see suppl. sect.). SUB1 and MYC together form many FFLs in the regulatory network. We hypothesized that MYC can bind to the promoter regions of key oncogenes to initiate their transcription, whereas SUB1 binds to 3UTRs to stabilize oncogenes at the level of RNA transcripts. Such collaboration between MYC and SUB1 results in overexpression of several key oncogenes and leads to proliferation of cancer cells (see suppl. sect. xxx). 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 (Fig. 3E).

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

For data-rich cell types with numerous TF ChIP-seq experiments, we built cell-type specific regulatory networks. These networks enable direct comparisons of network rewiring during oncogenesis. To achieve the best pairing given the existing data, we constructed a "composite

Deleted: [30,31]

Moved (insertion) [2]

Moved (insertion) [3]

Deleted: MYC works

Deleted: TFs. We

Deleted: , as

Deleted: the triplets in

Deleted: exact structure of these

Deleted: one

Deleted: [31].

Deleted: [32]

Deleted: show

Deleted: We analyzed the RBP network in a similar manner to the TF network, finding key regulators associated with cancer (see suppl.).

Moved up [2]: 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. 3D). 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.

Moved up [3]: 3D).

Deleted: 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.

Deleted: 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.

Deleted: the top-tier

Deleted: our resource contains

Deleted: , which

Deleted: comparison with networks built from their paired normal cell types.

Deleted: construct

normal" by reconciling multiple related normal cell types (see suppl.). Although the pairings are only approximate, many of them have been widely used in prior studies (see suppl.). Furthermore, they leverage the extensive functional characterization assays in ENCODE to provide us with a unique opportunity to study regulatory alterations in cancer on a large scale for the first time.

In particular, we measured the signed fractional number of edges changes for "tumor-normal pairs", (which we call the "rewiring index") to study how TF targets change in the oncogenic transformation. In Fig. 4A, 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[7,8]. 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 from multiple genes (so called "gene communities"). 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 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. The majority of its gained targets in tumor cells demonstrate higher levels of 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 indicates most non-coding risk variants are not well-explained by a mutational model[9]. 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.

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.

Stemness measurement during oncogenic transformation through 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. We projected the xxx RNA-seq data by Residual Component Analysis (RCA) to cluster various cell types according to the similarity of their transcriptomes. We found that various types of stem cells,

Deleted: previously

Deleted: the literature

Deleted: the

Formatted: Font color: Black

Deleted: in "tumor-normal pairs,"

Deleted: ,

Deleted: changing

Deleted: 5A

Deleted: [34,35]

Comment [LU2]: Should this be in quotes? Are we coining this expression?

Comment [LU3]: Not sure about use of quotes here. Would use quotes with first mention if we are inventing/coining the expression, but not subsequently. 'Gene-community model' has already been mentioned.

Deleted: "), which come from multiple genes.

Deleted: the

Moved down [4]: 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.

Comment [LU4]: Is this correct? Not sure what 'the current model' refers to.

Deleted: the current

Deleted: [36]

Moved (insertion) [4]

including data-rich H1 cells, form a tight cluster (blue in Fig. 5A). Interestingly, we observed that tumor cells (green in Fig. 5A) are located more proximal to the stem group than its normal counterpart (yellow), which is consistent with recent discoveries \cite{TCGA stemness}. Furthermore, we extended our analysis from transcriptome to both proximal and distal regulatory networks and observed a strong pattern: tumor cells tend to cluster together around stem cells, unlike normal cells.

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 the hypothesis that oncogenic TFs contribute to the state of cell differentiation, 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

Step-wise prioritization scheme pinpoints deleterious features associated with oncogenesis

Summarizing the above, our companion resource consists of annotations of (1) overall somatic and germline mutational burden scores; (2) accurate and compactly defined regulatory elements by integrating various novel functional assays, including eCLIP and 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 3D structural changes; (5) TF regulatory networks, both merged and cell-type specific, based on both distal and proximal regulation; (6) an analogous but less-developed network for RBPs; (7) attributes of TF/RBPs derived from network analysis, such as position in the network hierarchy, regulatory potential, and rewiring status. All the resources mentioned above are available online through the ENCODE website as simple flat files and computer codes (see suppl.).

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

Small-scale validation experiments on prioritized regulators and elements

To demonstrate the utility of our ENCODE resource, 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 experiments to validate their regulatory effects (Fig. 3B and 3D). We highlighted large scale structural variations that potentially disrupt oncogene insulation and validated their effects through

Comment [5]: I'm not really sure there is such a linear axis between normal cells and stem cells as we are implying,

Comment [LU6R5]: Upon reflection, this is probably not a big deal. However, it's also not obvious that 'reversion' is the process occurring here.

Deleted: SNVs in cancer

Deleted: in Fig.

Deleted: 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

Deleted: burdens in a diverse selection of cancers

Deleted: enhancer and promotor annotation that is based on

Deleted: many

Deleted: regulatory

Deleted: for each

Deleted: , its

Comment [LU7]: These descriptions could perhaps use some work still. Some of the item descriptions are so long that they get a bit confusing. The list also has quite a few items (7), so it's difficult to keep track of all the components. We may actually consider reducing the number of items, as well as the length of their descriptions.

No changes for now. Can discuss in P2.

Deleted: its

Deleted: ; and (7) an analogous but less-developed net ... [7]

Deleted: allow

Comment [LU8]: Aren't regulators elements too?

Deleted: .

Deleted: scheme

Deleted: as a workflow

Deleted: search

Comment [LU9]: Either/or, neither/nor.

Deleted: prioritize

Deleted: ,

Comment [LU10]: I think formally we may need ar ... [8]

Deleted: SNVs

Comment [LU11]: Again, aren't regulators a kind ... [9]

Deleted: genomic

Deleted: and variants

Deleted: the

Deleted: resources

Formatted: Font:14 pt, Font color: Black

Comment [LU12]: 'Larger' implies a comparison.

Deleted: also identified several candidate enhancers i ... [10]

Deleted: ability to influence transcription using lucif ... [11]

CRISPR engineered deletions (Fig. 2E). Finally, we selected key SNVs based on their disruption of enhancers with strong influence on gene expression. These SNVs were prioritized based on mutation recurrence in breast-cancer cohorts, as well as enhancer motif disruption scores. 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[10]. We further observed strong binding of many TFs in this region in MCF-7. Motif analysis predicts that a common mutation in breast cancer affects this region, and significantly disrupts the local binding affinity of several TFs, such as FOSL2 (Fig. 6C). 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 resource 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 believe that one of the best applications of our resource is to cancer research.

A key caveat related to our resource concerns network rewiring in cell-type specific networks. The utility of these networks in cancer is based on associating them to particular cancer types and then pairing a specific cancer network with a composite normal. Both correspondences are approximate. Nevertheless, we feel that our 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 tumor cells from a given patient usually show distinct molecular, morphological, and genetic profiles[11]. Cell-type specific or tissue-type specific analyses may not fully capture the heterogeneity seen in cancer. However, to place this limitation in context, it can even be challenging to obtain a representative match between 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 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.) 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.

Deleted: ,

Comment [LU13]: Is this rewording correct?

Deleted: and

Deleted: within these enhancers that are important for controlling gene expression.

Deleted: [37]

Deleted: the particular

Deleted: from a

Deleted: patient

Deleted: , in this region

Formatted: Font color: Auto

Deleted: study

Formatted: Font color: Auto

Deleted: ENCODE

Deleted: as an aid

Deleted: interpreting cancer genomes. It presents

Deleted: EN-CODEC companion resource, which tailors the ENCODE annotation to cancer. This has three parts: 1) cancer

Deleted: BMR models with significantly increased accuracy; 2) compact

Deleted: that by maximize statistical power for recurrent-mutation detection;

Deleted: 3) various

Deleted: and hierarchies for both pan-cancer and cancer-specific studies

Deleted: the

Comment [LU14]: Do not believe this should be in quotations.

Deleted: of these "

Deleted: "

Deleted: the EN-CODEC

Deleted: even

Deleted: [38]. It will be difficult to obtain a "perfect" match even from real tumor and normal tissues taken from a single patient.

Comment [LU15]: I'm not sure this is really what we want to say… But in the original discuss it is not ... [12]

Deleted: For example, we observed increased accura ... [13]

Deleted: This

Deleted: give us

Deleted: greater sense

Deleted: which

Comment [LU16]: Given the broad scope of the ... [14]

Deleted:

Deleted: Finally, we demonstrated the utility of our r ... [15]

1 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).
2 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).
3 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).
4 Dang, C. V. MYC on the path to cancer. *Cell* **149**, 22-35, doi:10.1016/j.cell.2012.03.003 (2012).
5 McKeown, M. R. & Bradner, J. E. Therapeutic strategies to inhibit MYC. *Cold Spring Harb Perspect Med* **4**, doi:10.1101/cshperspect.a014266 (2014).
6 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).
7 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).
8 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).
9 Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343, doi:10.1038/nature13835 (2015).
10 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).
11 Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328-337, doi:10.1038/nature12624 (2013).

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 ENCODEC 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

Page 2: [2] Deleted	All	5/20/18 8:31:00 PM

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.

Page 2: [3] Deleted	All	5/20/18 8:31:00 PM

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,

Page 2: [4] Deleted	All	5/20/18 8:31:00 PM

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 non-coding elements to genes (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

Page 2: [5] Deleted	All	5/20/18 8:31:00 PM

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. 2D illustrates the greater power in detecting mutationally burdened non-coding regions in several well-known cancer cohorts. For example

Page 2: [6] Deleted **All** **5/20/18 8:31:00 PM**

the increased power provided by the extended-gene annotation

Page 6: [7] Deleted **All** **5/20/18 8:31:00 PM**

; and (7) an analogous but less-developed network for RBPs.

Page 6: [8] Commented **Lab User** **5/20/18 6:23:00 PM**

I think formally we may need an 'or' between each possibility being correlated, but it does make the sentence a bit choppy…

Could keep or reject these additions of 'or' in this paragraph.

Same for 'either'. Formally, I believe they may be required. Not sure we want to be so formal.

Page 6: [9] Commented **Lab User** **5/20/18 6:26:00 PM**

Again, aren't regulators a kind of element?

Page 6: [10] Deleted **All** **5/20/18 8:31:00 PM**

also identified several candidate enhancers in noncoding regions associated with breast cancer

Page 6: [11] Deleted **All** **5/20/18 8:31:00 PM**

ability to influence transcription using luciferase assays in MCF-7.

Page 7: [12] Commented **Lab User** **5/20/18 6:54:00 PM**

I'm not sure this is really what we want to say… But in the original discuss it is not clear why we were discussing the heterogeneity of cancer. The reviewer also asked us to disucss

Page 7: [13] Deleted **All** **5/20/18 8:31:00 PM**

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

Page 7: [14] Commented **Lab User** **5/20/18 7:01:00 PM**

Given the broad scope of the paper, not sure about the relatively narrow scope of these last two sentences, that focus on TFs and chromatin.

Page 7: [15] Deleted **All** **5/20/18 8:31:00 PM**

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 ENCODEC 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 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).