An integrative ENCODE companion resource to interpret cancer genome

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

Abstract

ENCODE comprises thousands of functional genomics data sets, related to numerous cancer types; it is possible to tailor them into a targeted resource for interpreting cancer genomes. In particular, this resource can be used to measure the impact of non-coding mutations, constituting the bulk of the somatic variants. **[[JZ2MG: I understand the reversion, but it is quite confusing what is "next generation", and also quite controversial]]** Moreover, by integrating next-generation assays (e.g. STARR-seq) with many epigenetic features, we can significantly refine and make more focused these annotations (beyond a more general genome annotation), increasing the power for recurrent-mutation detection. Second, ENCODE signal data, especially replication timing, allows us to build precise, cancer-matched background models for mutation rates considerably more accurate than previous models. Third, ENCODE data, incorporating new assays, such as Hi-C and RNA-binding protein assays (i.e., eCLIP), in addition to large-scale transcription-factor ChIP-seq, allows the construction of extensive regulatory networks. In some contexts, these networks reveal how connections "rewire" during oncogenesis, as well as how these changes relate to a stem-cell state. More generally, **[[JZ2MG: "one can use" still sounds weird]]** one can use ENCODE networks to prioritize regulators most associated with large-scale expression changes in cancer. Combining the networks with the refined annotations and background mutation models, one can develop a step-wise prioritization scheme for non-coding mutations. Here, we show how this can be instantiated, and we perform a number of small-scale validations (i.e., luciferase assays and shRNA knockdowns) to demonstrate how the resource can prioritize mutations with significant consequences in cancer.

Introduction

Large-scale functional genomics data are useful for dissecting cancer genomes, particularly for interpreting mutation and expression profiles. The initial ENCODE release in 2012, along with other targeted functional genomic data, have motivated a number of integrative studies. These data allow us to assign functional impact to non-coding mutations, which constitute the bulk of mutations in cancer genomes. For instance, numerous researchers initiated integrative efforts to jointly interpret various cancer genomes[1-6]. In addition, ENCODE data sets (especially those related to replication timing and other signals) are useful for estimating background mutation rates (BMR), which vary greatly over the genome. Researchers have incorporated genome-wide features, such as replication timing, methylation, and expression profiles, for BMR estimation and consequently cancer driver detection[7,8]. ENCODE data are also useful for connecting non-coding elements (such as enhancers or promoters) into regulatory networks, which are pivotal for understanding cancer from a systems-biology perspective. Previously,

several groups have successfully used various types of networks to facilitate our understanding of data from cancer patients[9-11].

The new release of ENCODE data has a number of improvements over the last release, which allows us to construct a customized ENCODE companion resource for Cancer genomics (ENCODEC). This consists of a set of freely distributed annotation files and codes available online (see supplement). It comprises three main parts: a background mutation rate model, compact annotations, and regulatory networks. We detail each of these parts below and provide illustrations of how they may be used to dissect cancer genomes after combining mutation and expression profiles from large cancer cohorts such as from the TCGA.

In particular, with a much wider selection of cell types than the previous ENCODE release[12], as shown in Fig. 1 ENCODEC **[[I don't prefer to use ENCODEC here, might be annoying to other ENCODE members]]** 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 types of more specialized assays (such as STARR-Seq, HiC, and ChIA-pet) to accurately define core enhancers and their target genes. 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 considerably more extensive ChIP-Seq coverage and constructs additional networks from more recent assays such as eCLIP and Hi-C. For a few prominent cancer types, these provide cell type-specific networks in model tumor and normal cells, enabling direct measurement of potential regulatory changes in oncogenesis. Furthermore, a prevailing paradigm has held for decades that at least a subpopulation of tumor cells have the ability to self-renew, differentiate, and regenerate, in a manner that is similar to stem cells[13]. Hence, the top-tier cell line H1-hESC can serve as a valuable comparison when investigating the degree to which the oncogenic transformation represents stem-cell-like activities. More generally, our network can better explain cancer-specific expression patterns in tumors from resources such as TCGA, and it also helps reveal key regulators that drive large-scale tumor-to-normal expression changes.

We combined this network analysis with the compact annotation sets and mutational burdening (from the enhanced background model) to propose a step-wise prioritizing scheme to highlight key mutations associated with cancer progression. We validated the functional impact of prioritized mutations and elements using focused experiments such as shRNA 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.

[JZ2MG: variant recurrence detection might be confusing for readers]

More accurate BMR estimation ENCODE data improves variant recurrence detection in cancer

[[JZ2MG: we need to make decision whether to use BMR or not? I think it is OK… since other cancer genomics paper like the mutsig use such terms. Although it is difficult to understand for regulatory genomics ppl]]

One of the most powerful ways of identifying key elements in cancer genomes is through mutation recurrence analysis, the objective of which is to discover regions that harbor more mutations than expected. In such analysis, it is key to calibrate accurate mutation rate expectation prior to the burden test. Hence, we demonstrate how to integrate extensive ENCODE data to construct an accurate background mutation rate model in a wide range of cancer types. Accurate BMR estimation 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[8].

We address the issues associated with confounding factors in a cancer-cohort-specific manner (see supplement). 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, Hi-C, and expression profiles. In contrast to methods that use data from unmatched cell types[8], our approach automatically selects the most relevant features, thereby providing considerable improvements in BMR estimation (Fig 2A). For example, using matched replication timing data in multiple cancer types significantly outperforms using just replication timing data from the unmatched HeLa-S3 cell line. Moreover, combining many different genomic features significantly improves the estimation accuracy upon this (Fig 2 B). 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[7,14,15]. 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 supplement). Hence, our analyses may easily be extended to many cancer types.

A focused compact annotation improves power in variant recurrence detection in cancer

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

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 shapes of signal tracks from DNase-seq and a battery of up to 10 histone modification marks (see suppl.). Using a second algorithm, we then intersected these predictions with the result of STARR-seq experiments (see supplement). These experiments provide a direct, albeit noisy, readout of enhancer activity in specific cell types. Such an integrative 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 supplement).

Linking genes to non-coding elements to create an extended gene annotation and its use in determining mutationally burdened regions

To increase statistical power, a final part of our "compact" annotation entails linking noncoding regulatory elements to protein-coding exons to form an extended gene region as a single test unit. Such a unified annotation enables joint evaluation of the mutational signals from distributed yet biologically connected genomic regions. Traditional methods for linking rely solely on the correlation of individual signals (e.g., between the activity of one histone mark at an enhancer and gene expression of neighboring genes), and these may result in inaccurate extended gene definitions. Here, we use direct experimental evidence on physical interactions from Hi-C and ChIA-PET experiments, combined with a machine learning algorithm that takes into consideration the wide variety of histone modification marks and gene expression to delineate accurate enhancer-target gene linkages.

By integrating our compact annotation sets, BMR estimates, and accurate extended gene definitions , we were able to obtain maximal power for detecting genomic regions (coding and non-coding) that are mutationally burdened. Fig. 2C illustrates the greater power in detecting mutationally burdened non-coding regions in several well-known cancer cohorts. For example, in the context of chronic lymphocytic leukemia (CLL), our analyses identified well-known highly mutated genes (such as TP53 and ATM that have been reported from previous analyses. It also discovered genes missed by the exclusive analysis of coding regions, such as BCL6. Variants of BCL6 are known to have strong prognostic value for patient survival (Fig. 2D).

Interpreting tumor expression profiles using ENCODE regulatory networks identifies key regulators in cancer

Building on the extended gene annotation, we provide detailed regulatory networks. Specifically, for TF networks, we incorporated both distal and proximal networks by linking TFs to genes, either directly by TF-promoter binding or indirectly via TF-enhancer-gene interactions in each cell type (see supplement). We then pruned these networks to include only the strongest edges using a signal shape algorithm[16]. In addition, we reconciled all our cell-type-specific networks to form a generalized pan-cancer network. Similarly, we also defined an RNA-binding protein (RBP) network. Compared to imputed networks derived from gene expression or motif analyses, our ENCODE TF and RBP networks were built using ChIP-seq and eCLIP experiments, which provide much more accurate regulatory linkages between functional elements.

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

As expected we found that the target genes of MYC are significantly up-regulated in numerous cancer types, which is consistent with its well-known role as an oncogenic TF[17]. We further validated MYC's regulatory effects using knockdown experiments in breast cancer (Fig 3). Consistent with our predictions, the expression of MYC targets is significantly reduced after MYC knockdown in MCF-7 (Fig 3B). We then used the regulatory network to investigate how MYC works with other TFs. We first looked at MYC's target genes shared with a second TF, as shown in the triplets in Fig 3C. In all cancer types, we found that the shared target genes' expressions are strongly positively correlated with MYC, while they

showed only limited correlation with the second TF (as determined by partial correlation analysis, see supplement).

We further investigated the exact structure of these regulatory triplets. The most common one is the well-understood feed-forward loop (FFL). In this case, MYC regulates both another TF and a common target of both MYC and that TF (Figure 3 C). Since MYC amplification is a major determinant of many cancers, understanding which TFs appear to further amplify its effects may yield insights for efforts aimed at MYC inhibition[18]. 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[19] (see suppl.). We show that most of these gates follow either an OR or MYC-always-dominant logic gate. Thus, the ENCODE regulatory network does not only identify key cancer regulators, but also demonstrates how these work in combination with other regulators.

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

We further analyzed the overall TF regulatory network by systematically arranging it into a hierarchy (Fig 4). 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[20] (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 gene expressions in tumors

Cell-type specific regulatory network highlights 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 for several cancer types, which enables in a model context, direct comparison with networks built from their paired normal cell types. To achieve the best paired normal, given the existing data, we build a "composite normal" by reconciling multiple related normal cell types (see supplement). Although the pairings (i.e., relating cancerous cell lines to specific tumors and then matching them to normal cell types) are only approximate, many of them have previously been widely used in the literature (see supplement). Furthermore, they leverage the extensive functional characterization assays in ENCODE to provide us with a novel opportunity to directly understand the regulatory alterations in select cancers.

In particular, in "tumor-normal pairs," we measured the signed, fractional number of edges changing (which we call the "rewiring index") to study how TF targets change over the course of oncogenic transformation. In Fig. 5A, we ranked TFs according to this index. In leukemia, well-known oncogenes (such as MYC and NRF1) were among the top edge gainers, while the well-known tumor suppressor IKZF1 is the most significant edge loser (Fig 5A). Mutations in IKZF1 serve as a hallmark of various forms of high-risk leukemia[21-23]. We observed a similar trend in TFs using distal, proximal, and combined networks (see details in supplement). This trend was consistent across 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. The targets within the TF regulatory network were characterized by heterogeneous network modules (so called "gene communities"), which come from multiple biologically relevant genes. Instead of directly measuring the changes in a TF's targets between tumor and normal cells, we determined the changes in its gene communities via a mixed-membership model (See suppl.). Similar patterns to the direct rewiring were observed using this model (Fig 5A).

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.

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). This is consistent with previous discoveries that most non-coding risk variants are not well-explained by the current model[24]. For example, JUND is a top gainer in K562. The majority of its gained targets in tumor cells demonstrate higher gene expression, stronger active and weaker repressive histone modification mark signals, yet few of its binding sites are mutated. With a few notable exceptions (see supplement), 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 Figure 4. Specifically, in blood cancer, the more mutationally burdened TFs sit at the bottom of the hierarchy, whereas the TFs more associated with driving cancer gene expression changes tend to be at the top.

Step-wise prioritization schemes pinpoints deleterious SNVs in cancer

Summarizing the analysis above, our companion resource consists of annotations summarized in Fig. 6 and 1: (1) a BMR model with a matching procedure for relevant functional genomics data and a list of regions with higher-than-expected mutational burdens in a diverse selection of different cancers; (2) accurate, minimal and compactly defined enhancers and promotors that are defined by integrating many functional assays, including STARR-seq; (3) enhancer-target-gene linkages and extended gene neighborhoods that are obtained by integrating linkages from Hi-C and multi-histone mark and expression correlation; (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 rewiring status; and (7) an analogous but less-developed network for RBPs. All the resources mentioned above are available online through the ENCODE website as simple flat files and computer codes (see supplement).

Collectively, these resources allow us to prioritize key genomic features associated with oncogenesis. Our prioritization scheme is schematized in as a workflow in Fig. 6A. We first search for key regulators that are frequently rewired, located in network hubs, sit at the top of the 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 nucleotide level, by estimating their ability to disrupt or introduce specific binding sites (or which otherwise occur in positions under strong purifying selection), we pinpoint impactful SNVs that are further interrogated by focused functional characterization.

Small-scale validation experiments on the prioritization

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

One particularly interesting example, illustrating the unique 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) indicate its active regulatory role as an enhancer in MCF-7. This was further confirmed by STARR-seq (Fig. 5D). Hi-C and ChIA-PET data indicated that the region is within a topologically associated domain and validated a regulatory linkage to the downstream breast-cancer-associated gene SYCP2[25,26]. We observed strong binding of many TFs in this region in MCF-7. Our motif-based analysis predicts that the particular mutation from a breast cancer patient significantly disrupts the binding affinity of several TFs, such as FOSL2, in this region (Fig. 6D). Luciferase assays demonstrated that this mutation introduces a 3.6-fold reduction in expression relative to the wild-type, indicating a strong repressive effect on this enhancer's functionality.

Conclusion

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

One key caveat in our resource concerns the model cell type specific networks. Their utility for cancer is based on pairing them to particular cancer types. Although the representative tumor and normal cell types and their pairings are approximate, we feel that the networks provide the best current view on the regulatory changes in oncogenesis. No other system even has this scale of TF-chip data. Moreover, the heterogeneous nature of cancer means that even tumor cells from a given patient usually show distinct molecular, morphological, and genetic profiles[27]. It is difficult to obtain a "perfect" match even from real tumor and normal tissues taken from a single patient.

Our study underscores the value of large-scale data integration, and we note that expanding the scale of our approach is straightforward. For example, a larger number of genomic features from matched cell types could result in better BMR estimation; more advanced functional characterization assays may generate further compact annotation sets, and more ChIP-seq/eCLIP experiments on additional factors would provide more detailed regulatory networks. Larger patient cohorts of expression and mutation profiles from many cancer types may be used to discover novel key features in cancer genomes. We also anticipate that an additional step may entail carrying out many assays on specific tissues and tumor samples. We demonstrate that such large-scale integration is technically feasible and provides further opportunities for future studies.

Reference

1. Torchia, J. *et al.* Integrated (epi)-Genomic Analyses Identify Subgroup-Specific Therapeutic Targets in CNS Rhabdoid Tumors. *Cancer Cell* **30**, 891-908 (2016).
2. Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55-60 (2017).
3. Cancer Genome Atlas Research, N. *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378-384 (2017).
4. Rendeiro, A.F. *et al.* Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun* **7**, 11938 (2016).
5. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-5 (2014).
6. Hoadley, K.A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-44 (2014).
7. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360-4 (2015).
8. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).
9. Jacobsen, A. *et al.* Analysis of microRNA-target interactions across diverse cancer types. *Nat Struct Mol Biol* **20**, 1325-32 (2013).
10. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* **10**, 1108-15 (2013).
11. Mutation, C. & Pathway Analysis working group of the International Cancer Genome, C. Pathway and network analysis of cancer genomes. *Nat Methods* **12**, 615-21 (2015).
12. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
13. O'Connor, M.L. *et al.* Cancer stem cells: A contentious hypothesis now moving forward. *Cancer Lett* **344**, 180-7 (2014).
14. Makova, K.D. & Hardison, R.C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* **16**, 213-23 (2015).
15. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504-7 (2012).
16. Cheng, C., Min, R. & Gerstein, M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* **27**, 3221-7 (2011).
17. Dang, C.V. MYC on the path to cancer. *Cell* **149**, 22-35 (2012).
18. McKeown, M.R. & Bradner, J.E. Therapeutic strategies to inhibit MYC. *Cold Spring Harb Perspect Med* **4**(2014).
19. Wang, D. *et al.* Loregic: a method to characterize the cooperative logic of regulatory factors. *PLoS Comput Biol* **11**, e1004132 (2015).
20. Cheng, C. *et al.* An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. *Genome Biol* **16**, 63 (2015).
21. 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**, 32-8 (2016).
22. Marke, R. *et al.* Tumor suppressor IKZF1 mediates glucocorticoid resistance in B-cell precursor acute lymphoblastic leukemia. *Leukemia* **30**, 1599-603 (2016).
23. de Rooij, J.D. *et al.* Recurrent deletions of IKZF1 in pediatric acute myeloid leukemia. *Haematologica* **100**, 1151-9 (2015).

24. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-43 (2015).
25. Masterson, L. *et al.* Deregulation of SYCP2 predicts early stage human papillomavirus-positive oropharyngeal carcinoma: A prospective whole transcriptome analysis. *Cancer Sci* **106**, 1568-75 (2015).
26. Parris, T.Z. *et al.* Frequent MYC coamplification and DNA hypomethylation of multiple genes on 8q in 8p11-p12-amplified breast carcinomas. *Oncogenesis* **3**, e95 (2014).
27. Meacham, C.E. & Morrison, S.J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328-37 (2013).