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 comprehensively annotate the noncoding regions in the human genome for the first time. The current release broadens the number of cell lines for these assays and considerably expands available tissue data. It also greatly increases the depth of coverage of assays by adding new approaches, such as STARR-seq, Hi-C, and eCLIP. The integration over many assays provides an unparalleled opportunity to develop accurate annotations in a cell-type specific manner, which is particularly useful for interpreting genomic variants associated with disease and cell-state changes that underlie many disease processes. Deep integration over many assays also allows us to connect many regulators and non-coding elements into multi-modal networks, including proximal (TF/RBP-gene) and distal ones (enhancer-gene or TF-enhancer-gene).

Here, focusing on data-rich ENCODE cell types, we performed deep integration over many functional assays to characterize the noncoding genome, which may serve as a valuable resource for disease studies. Cancer is one of the best applications to illustrate the key aspects of this integrative ENCODE resource. Unlike many other diseases, cancer is very much a disease of whole-genome dysregulation. Cancer cells may display aberrant behaviors of key regulators, extensive remodeling of epigenetics, and abnormal transitions between cell states. The wealth of ENCODE functional characterization data allows direct measurement of chromatin status, regulatory changes, and expression perturbations for individual genes. It may also be used to construct comprehensive high-quality networks to capture tumor-to-normal alterations from a more global perspective. ENCODE data may thus help better elucidate the biological mechanisms of cancer initiation and progression.

Therefore, we present an integrative *ENCODE* companion resource for *C*ancer genomics (ENCODEC). This resource consists of various annotations, networks, and code bundles available online. Its compact noncoding annotations and extended gene definitions can potentially increase the statistical power to interpret variant (both germline and somatic) and expression data in cancer. Its comprehensive experiment-based networks allow us to depict global alterations in network rewiring,

Style Definition ... [1]

Formatted: Standard, Justified

Deleted: original, submitted text to pot. use ... [2]

Formatted: Justified, Space Before: 6 pt

Formatted: Standard, Justified

Deleted: for the first time …he noncoding regions in the human genome for the first time. The current release broadens the number of cell lines for these assays and considerably expands available tissue data. It also greatly increased ... [3]

Formatted ... [4]

Deleted: [[cancer

Formatted: Not Highlight

Deleted: illustrating

Formatted: Not Highlight

Deleted: of the many of the data-rich cell †types are, it ... [6]

Formatted: Not Highlight

Deleted: applic

Deleted: … we need to review reasons, todisc]] {{try t ... [5]

Formatted: Not Highlight

Deleted: the encode

Formatted: Not Highlight

Comment [1]: fewer mutations;

Formatted: Not Highlight

Formatted: Not Highlight

Deleted: disregulation, which is well informed by man ... [7]

Formatted: Not Highlight

Deleted: data can

Formatted: Not Highlight

Deleted: measure epigenetic remodeling and cell-state ... [8]

Formatted: Not Highlight

Deleted: regulatory

Formatted: Not Highlight

Deleted: can be reliably constructed from thousands o ... [9]

Deleted: systems-level

Formatted: Not Highlight

Deleted: of cancer. One

Formatted: Not Highlight

Deleted: directly measure the perturbations of indivi ... [10]

Formatted: Not Highlight

Formatted: Not Highlight

Deleted: cell space, epigentics remodeling, unified da ... [11]

Formatted ... [12]

Formatted ... [13]

Deleted: It allows us to prioritize key regulators, non ... [14]

hierarchies, and TF/RBP dysregulations. This lead us to the detection of many key cancer genes that are potentially of prognostic value, including the well-known TF MYC and novel RBP SUB1. We also discovered noncoding SVs that activate key oncogenes and SNVs that affect enhancer activity, and successfully validated them through CRISPR and luciferase assays. We further took advantage of the wealth of the ENCODE expression and chromatin state data to cluster various cell types, and highlight a consistent stem-like transition during normal-to-tumor development in many cancer types.

The breadth and depth of the ENCODE resource

Figure 1 illustrates two key dimensions of the overall ENCODE data set in relation to cancer: its breadth across cell types and depth across assays. Integration over the many different cell types, or many different assays, may contribute novel insights regarding cancer biology.

For example, it is challenging develop a background mutation rate model (BMR) for somatic recurrence in cancer – the somatic mutation process can be influenced by numerous confounders (in the form of both external genomic factors and local sequence context factors), which can result in false conclusions if not appropriately corrected. Researchers have suggested various regression approaches to integrate genomic features for accurate BMR calibrations. ENCODE has dramatically increased available BMR-associated features from less than 200 to over 2,000. Here we show in figure 1 that simple integration of these data sets progressively provides more accurate BMR estimates.

In addition, one can also integrate across assays to investigate how different patterns of epigenetic marks are related to structural variants (SV) in strictly matched cell types. ENCODE has various whole genome assays, which can contribute to accurate SV calls after proper integration with whole genome DNA sequencing data. 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.

An extended gene annotation and its applications in cancer

From the wealth of the ENCODE experiments in the data-rich cell lines, we constructed a deep, integrated annotation with two key characteristics – 1) it is compact to precisely pinpoint the functional sites, and 2) it extends the conventional gene annotation by linking the discontinuous noncoding regulatory regions to genes. Different from conventional gene annotations, which are uniformly defined

Formatted ... [16]
Deleted: * prioritiztion ... [17]
Formatted ... [18]
Deleted: {{not wrongh shoudl we keep? }}One
Deleted: most powerful ways of identifying key elements
Formatted ... [19]
Formatted ... [20]
Deleted: genomes
Formatted ... [21]
Formatted ... [22]
Deleted: through
Formatted ... [23]
Deleted: analysis to discover regions that harbor mor ... [24]
Formatted ... [25]
Deleted: and these
Formatted ... [26]
Deleted: corrected[15]. Hence, we demonstrate how
Formatted ... [27]
Deleted: extensive ENCODE data to construct an
Formatted ... [28]
Deleted: background mutation rate model
Formatted ... [29]
Deleted: a wide range
Formatted ... [30]
Deleted: cancer types.
Comment [Unknown A3]: I was a bit confused by th ... [31]
Formatted ... [32]
Deleted: {{shorten simplify}} We address this issue ... [33]
Formatted ... [34]
Comment [jingzhang4]: To disc with Shantao
Moved down [1]: Fig.
Formatted ... [35]
Deleted: 2A). For example, using matched replicatio ... [36]
Moved down [2]: sect. xxx).
Formatted ... [37]
Deleted: ... [38]
Formatted ... [39]
Deleted: .
Formatted ... [40]
Deleted: large amount
Formatted ... [41]
Deleted: intro data and essays on
Formatted ... [42]
Deleted:
Formatted ... [43]
Deleted: can construct
Formatted ... [44]
Deleted: augmentation. On this
Formatted ... [45]
Deleted: . [inaudible 00:00:23] to more
Formatted ... [46]
Deleted: location of
Formatted ... [47]
Deleted: basis. Then earlier first generation annotatio ... [48]
Formatted ... [49]
... [50]
Formatted ... [51]
... [52]
Formatted ... [53]
... [54]
Formatted ... [55]
... [56]

(although differentially expressed) across cell types, extended gene annotations are highly dynamic and may considerably change from cell to cell. This may benefit the statistical power of many analyses in cancer.

Our annotation is compact because we defined "core" regions of individual regulatory elements that are enriched for functional sites and reduce false positives rates via strict quality control (see suppl. sect. xxx). This can be doneeither by incorporating novel advanced assays such as eCLIP and STARR-seq, or integrating over tens of functional assays, such as DNase-seq and ChIP-seq (see suppl. sect. xxx). We have shown that such an annotation can effectively reduce the noise-to-signal ratio and reduce multiple test correction burden to achieve a better power for various analyses.

On the extended side, a second step of our integrative annotation entails linking the above compact elements to define a high-quality extended gene neighborhood by integrating computational predictions with direct experimental evidence on physical interactions from Hi-C and ChIA-PET experiments (see suppl. sect. xxx). Such a gene-centric approach not only allows us to improve upon existing knowledge of the genetic regions involved in cancer, but also enables a joint evaluation of distributed yet biologically connected genomic regions. This leads to increased power in many analyses (see suppl. sect. xxx).

To illustrate the value of our resource, we first compared the enrichment of cancer GWAS SNPs with respect to various annotations. The enrichment in protein-coding genes significantly increases as we add more relevant annotations for breast cancer and leukemia (Fig. 2C). This trend is more pronounced when the newly added proximal and distal noncoding annotations are from matched cell types. One may further subset the genes according to different subcategories associated with cancer, and identify enrichments. For instance, we observed a significant enrichment in genes from the Cancer Gene Consensus (CGC) in breast cancer based on the extended gene annotation, which was not possible using the conventional coding regions annotation.

We also showed that our extended gene annotation can provide better stratification of gene expression from mutational signals in cancer patients compared to single annotation categories. We combined the mutational and expression profiles from

Formatted ... [57]
Formatted ... [59]
Deleted: associating them with genes. The compact ... [60]
Deleted: various [inaudible 00:02:27] tests. ... [62]
Formatted ... [61]
Formatted ... [63]
Comment [Unknown A5]: We don't say why/how... ... [65]
Deleted: burden tests. In traditional genomic
Formatted ... [64]
Moved (insertion) [2] ... [66]
Formatted ... [67]
Moved (insertion) [3] ... [68]
Formatted ... [69]
Formatted ... [70]
Moved (insertion) [4] ... [71]
Formatted ... [72]
Deleted: a comprehensive set of annotations (usuall ... [73]
Formatted ... [74]
Deleted: power of burden tests by creating a focused ... [75]
Formatted ... [76]
Deleted: supported by multiple lines of evidence in th ... [77]
Formatted ... [78]
Deleted: us to define a minimal list of enhancers with ... [79]
Formatted ... [80]
Deleted: the mutational signals from
Formatted ... [81]
Deleted: Traditional methods for linking rely solely o ... [82]
Formatted ... [83]
Deleted: {{add more orginal power discussions, com ... [84]
Formatted ... [85]
Formatted ... [86]
Deleted: [[why compact]] For example, we explored ... [87]
Moved up [4]: sect. xxx).
Formatted ... [88]
Deleted: Additionally, in several well-known cancer ... [89]
Formatted ... [90]
Deleted: many
Formatted ... [91]
Deleted: genes and so one can see a much clearer ... [92]
Formatted ... [93]
Deleted: distributed yet biologically connected genom ... [94]
Formatted ... [95]
Deleted: By integrating our compact annotation sets, ... [96]
Moved down [5]: More importantly, the increased p ... [97]
Formatted ... [98]
Deleted: 2E and refs. [25-27]). ... [99]
Formatted ... [100]
Deleted: For instance, we
Deleted:

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).

By integrating our compact annotation sets, BMR estimates, and accurate extended gene definitions, we obtained larger power for detecting genomic regions (coding and non-coding) that are mutationally burdened. Fig. 2E illustrates genes that are mutationally burdened using our extended gene definitions 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 in previous analyses. More importantly, the increased power provided by the extended-gene annotation allowed us to detect genes that would otherwise be missed by an exclusively coding analysis. An example of this is the well-known cancer gene BCL6, which may be associated with patient survival (Fig. 2E).

One can get a physical sense of the importance of the extended gene environment by looking at a situation where genomic variant rearranges the extended gene without affecting coding regions. We found such an example in the breast cancer cell line T47D, where a 130Kb heterozygous deletion changes the chromosome to link a distal enhancer to the promoter and results in the activation of the well-known oncogene ERBB4 (Fig. 2F). This heterozygous deletion is located around 45Kb downstream from the ERBB4 promoter region and potentially merges two Hi-C TADs in an allele-specific way. This change occurs in cancer cells but not in normal cells. We tested this hypothesis through CRISPR editing, by excising an 86bp sequence that contains the CTCF binding sites at the boundary of the two Hi-C TADs from the wild-type allele in T47D cells. This excision confirmed the elevated ERBB4 expression upon CRISPR deletion (as measured by PCR).

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

Formatted ... [102]
Moved (insertion) [5] ... [103]
Formatted ... [104]
Deleted: [[bad transition here, see dict for rework]]
Formatted ... [105]
Formatted ... [106]
Deleted: of a given gene
Formatted ... [107]
Deleted: a mutation
Formatted ... [108]
Deleted: structure gene
Formatted ... [109]
Deleted: actually effecting the
Formatted ... [110]
Deleted: Such a mutation one can find
Formatted ... [111]
Deleted: mutation, for instance, in breast cancer,
Formatted ... [112]
Deleted: illustrated in figure XXX.
Formatted ... [113]
Deleted: structural variant
Formatted ... [114]
Deleted: looping and the TAZ structure. Moving in ... [115]
Formatted ... [116]
Deleted: oncogene
Formatted ... [117]
Deleted: .
Formatted ... [118]
Deleted: is a well-known oncogene in many cancer ... [119]
Formatted ... [120]
Deleted: 2D). We identified a 130Kb
Formatted ... [121]
Deleted: (~
Formatted ... [122]
Deleted: TSS) that
Formatted ... [123]
Deleted: and links a distal enhancer to the ERBB4 p ... [124]
Formatted ... [125]
Deleted: .
Formatted ... [126]
Deleted: (xxx color track from 4C experiment). We ... [127]
Formatted ... [128]
Deleted: resulted in
Formatted ... [129]
Deleted: Our results suggest that ERBB4 activation ... [130]
Formatted ... [131]
Deleted: 2 to 3 lines of QC ... [132]
Formatted ... [133]
Deleted:

or indirectly via TF-enhancer-gene interactions in each cell type (see suppl. sect. xxx). We then pruned the full network to the strongest interactions using a signal shape algorithm which keeps the strongest peaks by weighting their occurrence by the general TF binding profiles (see suppl. sect. xxx). 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. Compared to others, our ENCODE TF and RBP networks can capture more literature-supported regulations and correlate better with knockdown experiments (see suppl. sect. xxx).

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). These networks also enable investigation of the connections between TFs and RBPs. Interestingly, we found that there are less top-level TF-RBP interactions, as compared to middle and bottom level ones.

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.

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

Deleted: these networks to include only

Deleted: strongest edges using a signal shape algorithm[1] [[more? see dict]] (see suppl.). After building the phone networks who pruned them

Formatted: Not Highlight

Deleted: basically kept

Formatted: Not Highlight

Deleted: but weighted

Formatted: Not Highlight

Deleted: typical location of

Formatted: Not Highlight

Deleted: of the transcription factors. See supplement and reference

Formatted: Not Highlight

Deleted: .

Deleted: [LU1]

Formatted: Font:Times New Roman, 14 pt

Deleted: eCLIP is an enhanced CLIP protocol that provides single-nucleotide resolution of the RBPs binding signatures[2].

Deleted: imputed networks derived from gene expression or motif analyses

Deleted: provide experimentally based regulatory linkages between functional elements. [[how good?

Formatted: Not Highlight

Deleted: ?]]

Formatted: Not Highlight

Formatted: Standard, Justified

Deleted: {{these

Deleted: seeing hte

Deleted: betw rpbs & tfs

Deleted: in fact

Deleted: find few connections at the

Deleted:

Deleted: }}

Formatted: Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Deleted: .

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. We found that 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.

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

Formatted: Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Deleted: [[could we just mention FFLs? cut the following]]{{could we cut out the purple}} 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

Formatted: Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Pattern: Clear (White), Not Highlight

Deleted: .

rewiring during oncogenesis. To achieve the best pairing given the existing data, we constructed a "composite 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).

We organized the cell-type specific networks into hierarchies, as shown in Fig. 5B. Specifically, in blood cancer, the more mutationally burdened TFs sit at the bottom of the hierarchies, whereas the TFs more associated with driving cancer gene expression changes tend to be at the top. In addtion, we found that the strongest edge gainers and edge losers in rewiring events often sit at the top level of the network hierarchies.

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,

Deleted: ",
Deleted: ")
Formatted: Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)
Deleted: [LU2]
Deleted: "[LU3]
Deleted: 5A) and also in terms of a simpler co-binding approach (see suppl.).
Formatted: Not Highlight
Deleted: 4B
Formatted: Not Highlight
Deleted: hierarchy
Formatted: Not Highlight
Deleted: .{{stronger gainer & loser
Formatted: Not Highlight
Deleted: }}
Formatted: Not Highlight
Formatted: Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)
Deleted:

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 [LU4] 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.

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. One of the strengths of ENCODE is its many stem cell lines, including the H1 stem cell, which is one of the most data-rich cell types. We leveraged the large number of stem cells in ENCODE and the additional data available with this ENCODE release, to place tumor-associated cell types relative to normal cells and stem cells in cell space. We projected the RNA-seq data by Reference Component Analysis (RCA, \cite{nat rca paper}) to cluster various cell types according to the similarity of their transcriptomes. We found that various types of stem cells, including data-rich H1 cells, form a tight cluster (blue in Fig. 5A). As is observed from Fig. 5, there is potentially a trend where the transition from normal to tumor cells is moving toward a stem cell. This is true for a variety of different cancers. This observation is consistent with previous efforts using expression and methylation analysis. Interestingly, we observed a consistent (or even stronger) pattern from the proximal and distal chromatin data, which can be viewed as the underlying cause of the observed gene expression changes.

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 towards normal state upon oncogene knockdowns.

Step-wise prioritization scheme and result validations

Collectively, we propose a step-wise prioritization scheme that allows us to pinpoint key regulator genes, noncoding elements, and single nucleotides associated with oncogenesis, as schematized in Fig. 6A. Specifically, we first highlighted 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

Formatted: Justified, Space Before: 6 pt
Formatted: Standard, Justified
Deleted: [[encode good b/c of H1 + trans.. see dict & orig. text]]
Formatted: Not Highlight
Deleted: quite deep in
Deleted: nerve assays. We aim to leverage
Formatted: Not Highlight
Formatted: Not Highlight
Deleted: amount
Formatted ... [134]
Deleted: data
Formatted: Not Highlight
Deleted: all of the
Deleted: lines modeling tumors
Formatted: Not Highlight
Formatted: Not Highlight
Deleted: matched normals
Formatted: Not Highlight
Deleted: ,
Formatted: Not Highlight
Deleted: a kind of
Formatted: Not Highlight
Deleted: As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx…e projected the RNA-seq data by Reference Component Analysis (RCA, \cite{nat rca paper}) to cluster various cell types according to the similari ... [135]
Formatted ... [136]
Deleted: toward[5] [LU6]
Deleted: [[is the connection stronger? how to do tra ... [137]
Formatted ... [138]
Deleted: pinpoints deleterious features associated ... [139]
Formatted: Font:17 pt, Not Highlight
Deleted: germline mutational burden scores; (2) a ... [140]
Formatted: Font:17 pt
Formatted ... [141]
Deleted: these resources allowed us to prioritize key ... [142]
Formatted: Font:(Default) Arial, (Asian) 宋体, 9 pt
Deleted:

tumors, or undergo large changes in expression or regulation. Finally, on a nucleotide level, by estimating their ability to disrupt or introduce specific binding sites, we pinpoint impactful genomic variants at a fine scale.

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 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. [LU12] 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. 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 underlines the importance of deep data integration over many novel assays to understanding cancer genomes. A compact annotation with extended gene definitions can benefit various analyses in cancer. The value of our resource can go far beyond noncoding variant interpretation, to allow exploration ofgenome-wide gene dysregulations, epigenetic remodeling, network perturbations, and cell state transitions. These are also keys aspects of genome interpretation that are tightly

Deleted: gene

Deleted: ,

Deleted: [LU9] TF binding, or chromatin status

Deleted: genome

Deleted: level

Deleted: Small-scale validation experiments on prioritized regulators and [LU10] **elements** ... [143]

Formatted: Heading 2, Justified, Space Before: 6 pt, After: 4 pt

Formatted: Font:17 pt

Formatted: Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Deleted: [LU11]

Formatted: Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Deleted: SYCP2[10]

Formatted: Justified, Space Before: 6 pt

Deleted: 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 such deep integrative annotation is cancer genomics.

Deleted:

associated with oncogenesis. The successful validation of our prioritized regulator genes, noncoding elements, and SNVs demonstrates the value of our resource.

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 various cancer types and then pairing a specific cancer network with a composite normal. Both pairings are approximate. Moreover, cancer is well-knowns for its heterogeneity. Tumor cells from a given patient usually show distinct molecular, morphological, and genetic profiles. Linking cancer to one specific cell-type may not fully capture the heterogeneity seen in cancer. 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. Further technological advances, such as single cell sequencing, may help to provide more biological insights at a higher resolution. Nevertheless, we feel that our networks currently provide the best available view of the regulatory changes in oncogenesis since no other system currently has data at this scale.

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 provides a way of selecting candidate key 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.

Formatted: Standard, Justified

Deleted: particular

Deleted: correspondences[LU13]

Deleted: means that tumor

Formatted: Font:Arial, 9 pt

Deleted: 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.

Deleted: the heterogeneous nature of

Deleted: profiles[11]. Cell-type

Deleted: or tissue

Deleted: specific analyses

Deleted: However, to

Deleted: .[LU14]

Formatted: Font:Times New Roman, 14 pt

Formatted: Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Deleted: actually

Deleted: [LU15]

Deleted: [[more of the caveats of the referees?? single cell in heterog. make it clear there are other aspects of cancer rather than driver detection, can use encode annotation, other part of the mutation; put somthing more than noncoding, even all drivers are in genes, still there are alot going on in noncoding genomes and can be disc by encode]] ... [144]

Deleted:

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [1] Style Definition **author** **5/24/18 6:10:00 PM**

Comment Reference

Page 1: [2] Deleted **author** **5/24/18 6:10:00 PM**

original, submitted text to pot. use
original, submitted text to not. use
JZ text
[[comment text, incl dict inserts]]
dictations:

Page 1: [3] Deleted **author** **5/24/18 6:10:00 PM**

for the first time

Page 1: [3] Deleted **author** **5/24/18 6:10:00 PM**

for the first time

Page 1: [4] Formatted **author** **5/24/18 6:10:00 PM**

Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [5] Deleted author 5/24/18 6:10:00 PM

... we need to review reasons, todisc]] {{try to rewrite the color bits}}for several reasons. First,

Page 1: [6] Deleted author 5/24/18 6:10:00 PM

of the many of the data-rich cell †types are, in fact, associated with cancer, including cell types from blood, breast, liver, and lung (Fig. 1). Second

Page 1: [7] Deleted author 5/24/18 6:10:00 PM

disregulation, which is well informed by many of the ENCODE assays measuring epigenetic changes and large genomic variants (eg SVs) in cancer cells. Third, the

Page 1: [8] Deleted author 5/24/18 6:10:00 PM

measure epigenetic remodeling and cell-state transitions, which are implicated in oncogenesis. Lastly,

Page 1: [9] Deleted author 5/24/18 6:10:00 PM

can be reliably constructed from thousands of experiments to provide

Page 1: [10] Deleted author 5/24/18 6:10:00 PM

directly measure the perturbations of individual regulators and entire networks to

Page 1: [11] Deleted author 5/24/18 6:10:00 PM

cell space, epigentics remodeling, unified data and annotation, not just prioritizing variants and regulators

{{try to shorten this para & rewrite based on disc}}

Page 1: [12] Formatted author 5/24/18 6:10:00 PM

Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 1: [13] Formatted author 5/24/18 6:10:00 PM

Font:Italic, Underline

Page 1: [13] Formatted author 5/24/18 6:10:00 PM

Font:Italic, Underline

Page 1: [14] Deleted author 5/24/18 6:10:00 PM

It allows us to prioritize key regulators, non-coding elements and variants in relation to cancer. These prioritized elements are burdened by germline and somatic mutations, sit at central positions in the regulatory network, or to be associated with large epigenetic or expression 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. [[more? todisc]]

Page 2: [15] Commented **jingzhang.wti.bupt@gmail.com** **5/24/18 2:21:00 PM**

This is the result, not the method. Still on the long end but can be reworked.

Page 2: [16] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 2: [17] Deleted **author** **5/24/18 6:10:00 PM**

* prioritiztion
* networks & rewiring - sub1, heir. , co-reg.
* cell state , stem cell
* extended gene

Page 2: [18] Formatted **author** **5/24/18 6:10:00 PM**

Justified, Space Before: 6 pt

Page 2: [19] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [20] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [20] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [21] Formatted **author** **5/24/18 6:10:00 PM**

Standard, Justified

Page 2: [22] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [23] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [23] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [24] Deleted **author** **5/24/18 6:10:00 PM**

analysis to discover regions that harbor more mutations than expected. However, developing a null expectation for these analyses is non-trivial

Page 2: [25] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [26] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [27] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [28] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [29] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [30] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [31] Commented **Unknown Author** **5/24/18 4:24:00 PM**

I was a bit confused by this. This last sentence describes something different than improved estimates with greater #'s of features.

Page 2: [32] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [33] Deleted **author** **5/24/18 6:10:00 PM**

{{shorten simplify}} 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 (

Page 2: [34] Formatted **author** **5/24/18 6:10:00 PM**

Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 2: [35] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 2: [36] Deleted **author** **5/24/18 6:10:00 PM**

2A). For example, using matched replication timing data in multiple cancer types significantly outperforms an approach in which one restricts the analysis to replication timing data from the unmatched HeLa-S3 cell line. Moreover, combining many different genomic features significantly improves the estimation accuracy (Fig. 2B). The weightings of the features in the model are consistent with our expectations: for instance, for breast cancer, we observed elevated mutation rates in regions with the repressive mark H3K9me3 and a reduced mutation rate in regions with the activating, enhancer-associated mark H3K27ac[12-14]. Also, due to the correlated nature of genomic features across cell types, even approximate matching of a specific cancer type to a particular ENCODE cell line can still improve BMR estimation (see suppl.). Hence, our analyses may easily be extended to many cancer types.

As illustrated in Fig 1, our work takes advantage of the breadth and depth of the ENCODE data. First, the somatic mutation process can be influenced by numerous confounders. [[connection]] 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.

Page 2: [37] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [38] Deleted	author	5/24/18 6:10:00 PM

[[merge orig + JZ w/ dict]]

Figure 1 illustrates two key dimensions of the overall ENCODE set in relation to cancer. First, it's breadth across cell types and, second, it's depth across assays. Simple integration over the many different cell types, or many different assays in relation to variance can provide insights. First of all, one of the key problems in cancer is developing a somatic background mutation rate estimation, which is useful for burden calculations, as we'll see below. People have suggested a number of regression approaches integrating features such as negative [inaudible 00:01:02] aggression references. Here we show that one can improve on these estimates by using the wealth of the ENCODE, by using these regression approaches with a large marginal number of ENCODE data sets. For instance, the number of ENCODE data sets, epigenetic data sets, for estimating background mutation rate is more than a factor of 10 greater than any previously published integration, with more than 2000, including many of them associated with replication timing, which is thought to be mechanistically the single best predictor of mutation rate. We show in figure 1 that simple integration of these data sets progressively gets more accurate estimations, in that one gets a better estimate with more and more data sets than just a few. This may have to do with the heterogeneity of tumors and the way one cannot find an ideal perfect match for most cancers. One can also integrate across the many different assays, and one particularly interesting thing is to look at the many different patterns of epigenetic marks related to structural variance. People have previously observed these patterns with relation to germline variance, but using some of the ENCODE called structural variance from Hi-C and other events assays one can actually see the patterns of epigenetic variance around somatic variance. Here we see distinct pattern of an enrichment of a number of well-known activating marks around structure variance, which is distinctly different from what is observed for germline variance.

{{we make integartive annotation - compact & ext. to enrich }}

Page 2: [39] Formatted	author	5/24/18 6:10:00 PM

Justified, Space Before: 6 pt

Page 2: [40] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [41] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [42] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [42] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [43] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [44] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [45] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [46] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [47] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [48] Deleted	author	5/24/18 6:10:00 PM

basis. Then earlier first generation annotations, the more precise pinpoint [inaudible 00:00:41] basis,

Page 2: [49] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [50] Deleted	author	5/24/18 6:10:00 PM

last [inaudible 00:00:42] compact. [inaudible 00:00:47]

Page 2: [51] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [52] Deleted	author	5/24/18 6:10:00 PM

was less force process. It also gives rise to what we call Extended Gene Annotation, where

Page 2: [53] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [54] Deleted	author	5/24/18 6:10:00 PM

regions of the gene are [inaudible 00:01:01]. That weren't fashioned together with

Page 2: [55] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 2: [56] Deleted	author	5/24/18 6:10:00 PM

Skipping over, intervening [inaudible 00:01:11] basis. The compact

Page 3: [57] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 3: [58] Deleted	author	5/24/18 6:10:00 PM

, particularly useful for relating is extended in a compact annotation. It's particularly useful for relating [inaudible 00:01:36]. Such as those assumption cancer pragmatic [inaudible 00:01:41]. Pragmatic them over genomic elements,

Page 3: [59] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 3: [60] Deleted	author	5/24/18 6:10:00 PM

associating them with genes. The compact annotation, in tripod ones, where one knows which one [inaudible 00:02:13] one can readily see. In extreme [inaudible 00:02:16] basis, including one, when necessarily, give larger power

Page 3: [61] Formatted	author	5/24/18 6:10:00 PM

Not Highlight

Page 3: [62] Deleted	author	5/24/18 6:10:00 PM

various [inaudible 00:02:27] tests.
{{better annotation is useful for enriching signals in many different contexts}}
quality of the data:
[[submitted version - need better transition, tried w/ dict. Also, we need to weave in qual. of encode data]]
A second advantage of leveraging ENCODE data in determining recurrently mutated regions is provided by maximizing

Page 3: [63] Formatted	author	5/24/18 6:10:00 PM

Font:14 pt, Font color: R,G,B (0,0,10), Not Highlight

Page 3: [64] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Font color: R,G,B (0,0,10), Not Highlight

Page 3: [65] Commented **Unknown Author** **5/24/18 4:45:00 PM**

We don't say why/how… We also don't define 'extended gene'.
Could say in particular, or more specifically instead.

Page 3: [66] Moved from page 2 (Move #2) **author** **5/24/18 6:10:00 PM**

sect. xxx).

Page 3: [67] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 3: [68] Moved from page 3 (Move #3) **author** **5/24/18 6:10:00 PM**

sect. xxx).

Page 3: [69] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 3: [70] Formatted **author** **5/24/18 6:10:00 PM**

Justified

Page 3: [71] Moved from page 3 (Move #4) **author** **5/24/18 6:10:00 PM**

sect. xxx).

Page 3: [72] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 3: [73] Deleted **author** **5/24/18 6:10:00 PM**

, a comprehensive set of annotations (usually covering as many base pairs as possible) is considered to be optimal. However, testing every possible nucleotide in the genome greatly reduces the statistical power for variant recurrence detection (see suppl.). Here, we aim to increase

Page 3: [74] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Font color: Auto, Not Highlight

Page 3: [75] Deleted **author** **5/24/18 6:10:00 PM**

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

Page 3: [76] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 3: [77] Deleted **author** **5/24/18 6:10:00 PM**

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

Page 3: [78] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 3: [79] Deleted **author** **5/24/18 6:10:00 PM**

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.).
To increase statistical power, a final part of our "compact" annotation entails linking non-coding regulatory elements to protein-coding exons to form an extended gene region as a single test unit. Such a unified annotation enables

Page 3: [80] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 3: [81] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 3: [82] Deleted **author** **5/24/18 6:10:00 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.

Page 3: [83] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 3: [84] Deleted **author** **5/24/18 6:10:00 PM**

{{add more orginal power discussions, compact annotaion necssary enriching the funcional sites}}

We also utilized the depth of the ENCODE data to provide compact and accurate annotations with superior properties relative to other annotations (see suppl.

Page 3: [85] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 3: [86] Formatted **author** **5/24/18 6:10:00 PM**

Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 3: [87] Deleted **author** **5/24/18 6:10:00 PM**

[[why compact]] 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.

Page 3: [88] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 3: [89] Deleted author 5/24/18 6:10:00 PM

Additionally, in several well-known cancer cell types, we developed an 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.

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.

Page 3: [89] Deleted author 5/24/18 6:10:00 PM

Additionally, in several well-known cancer cell types, we developed an 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.

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.

Page 3: [89] Deleted author 5/24/18 6:10:00 PM

Additionally, in several well-known cancer cell types, we developed an 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.

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.

Page 3: [90] Formatted	**author**	**5/24/18 6:10:00 PM**

Not Highlight

Page 3: [91] Formatted	**author**	**5/24/18 6:10:00 PM**

Not Highlight

Page 3: [92] Deleted	**author**	**5/24/18 6:10:00 PM**

genes and so one can see a much clearer enrichment, for instance, for the known cancer associated extended genes than extended genes in general, which is quite satisfying.]

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

Page 3: [93] Formatted	**author**	**5/24/18 6:10:00 PM**

Not Highlight

Page 3: [94] Deleted	**author**	**5/24/18 6:10:00 PM**

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).

Page 3: [95] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 3: [96] Deleted **author** **5/24/18 6:10:00 PM**

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, 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].

Page 3: [97] Moved to page 4 (Move #5) **author** **5/24/18 6:10:00 PM**

More importantly, the increased power provided by the extended-gene annotation allowed us to detect genes that would otherwise be missed by an exclusively coding analysis. An example of this is the well-known cancer gene BCL6, which may be associated with patient survival (Fig.

Page 3: [98] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 3: [99] Deleted **author** **5/24/18 6:10:00 PM**

2E and refs. [25-27]).

Third,

Page 3: [100] Formatted **author** **5/24/18 6:10:00 PM**

Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 4: [101] Moved from page 2 (Move #1) **author** **5/24/18 6:10:00 PM**

Fig.

Page 4: [102] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 4: [103] Moved from page 3 (Move #5) **author** **5/24/18 6:10:00 PM**

More importantly, the increased power provided by the extended-gene annotation allowed us to detect genes that would otherwise be missed by an exclusively coding analysis. An example of this is the well-known cancer gene BCL6, which may be associated with patient survival (Fig.

Page 4: [104] Formatted **author** **5/24/18 6:10:00 PM**

Font:14 pt, Not Highlight

Page 4: [105] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [106] Formatted **author** **5/24/18 6:10:00 PM**

Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 4: [107] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [108] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [109] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [110] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [111] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [112] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [113] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [114] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [115] Deleted **author** **5/24/18 6:10:00 PM**

looping and the TAZ structure. Moving in the cancer associated with one gene to another and this in turn activates the first gene. This enhancer flipping [inaudible 00:00:50] structure variation has been remarked on before as a mechanism for disregulation in cancer, reference YYY.)

Finally, we found an example of how an SV introduced extended gene change that may lead

Page 4: [116] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [117] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [118] Formatted **author** **5/24/18 6:10:00 PM**

Not Highlight

Page 4: [119] Deleted author 5/24/18 6:10:00 PM

is a well-known oncogene in many cancer types \{cite}.

Page 4: [120] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [121] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [122] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [123] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [124] Deleted author 5/24/18 6:10:00 PM

and links a distal enhancer to the ERBB4 promoter in T47D

Page 4: [125] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [126] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [127] Deleted author 5/24/18 6:10:00 PM

(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.

Page 4: [128] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [128] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [129] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [129] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [130] Deleted author 5/24/18 6:10:00 PM

Our results suggest that ERBB4 activation in T47D may at least in part be due to the 130 kb deletion that disrupts its insulation.

Page 4: [131] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 4: [132] Deleted author 5/24/18 6:10:00 PM

2 to 3 lines of QC

Page 4: [133] Formatted author 5/24/18 6:10:00 PM

Justified, Space Before: 6 pt

Page 8: [134] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 8: [134] Formatted author 5/24/18 6:10:00 PM

Not Highlight

Page 8: [135] Deleted author 5/24/18 6:10:00 PM

As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx

Page 8: [135] Deleted author 5/24/18 6:10:00 PM

As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx

Page 8: [135] Deleted author 5/24/18 6:10:00 PM

As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx

Page 8: [135] Deleted author 5/24/18 6:10:00 PM

As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx

Page 8: [135] Deleted **author** **5/24/18 6:10:00 PM**

As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx

Page 8: [135] Deleted **author** **5/24/18 6:10:00 PM**

As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx

Page 8: [135] Deleted **author** **5/24/18 6:10:00 PM**

As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx

Page 8: [135] Deleted **author** **5/24/18 6:10:00 PM**

As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene

expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx

Page 8: [135] Deleted author 5/24/18 6:10:00 PM

As is obvious from Figure XXX, there's potentially a nice linear relationship where one transitions from the normal to the cancer en route to a stem cell. This is true for a variety of different cancers. To some degree this fact has been remarked on before, but what's most interesting is it can not only be seen from RNA, from gene expression data, but it can be seen more strongly in terms of the underlying cause, the proximal and distal chromatin data. In fact looking at the chromatin 1, can even look at the rewiring. We projected the xxx

Page 8: [136] Formatted author 5/24/18 6:10:00 PM

Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 8: [137] Deleted author 5/24/18 6:10:00 PM

[[is the connection stronger? how to do transition - to think about {{I guess we should string}}]] Related to this: 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, in blood cancer, 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.

Page 8: [138] Formatted author 5/24/18 6:10:00 PM

Heading 2, Justified, Space Before: 6 pt, After: 4 pt

Page 8: [139] Deleted author 5/24/18 6:10:00 PM

pinpoints deleterious features associated with oncogenesis

Summarizing the above, our companion resource consists of annotations of (1) overall somatic

Page 8: [140] Deleted author 5/24/18 6:10:00 PM

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. [[todisc & rewrite - concensus is to shrink & maybe merge w next]] All the resources mentioned above are available online through the ENCODE website as simple flat files and computer codes (see suppl.).

Page 8: [141] Formatted	**author**	**5/24/18 6:10:00 PM**

Standard, Justified, Border:Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Page 8: [142] Deleted	**author**	**5/24/18 6:10:00 PM**

these resources allowed us to prioritize key genomic features associated with oncogenesis at regulator, element[LU7] , and nucleotide levels. Our

Page 8: [142] Deleted	**author**	**5/24/18 6:10:00 PM**

these resources allowed us to prioritize key genomic features associated with oncogenesis at regulator, element[LU7] , and nucleotide levels. Our

Page 8: [142] Deleted	**author**	**5/24/18 6:10:00 PM**

these resources allowed us to prioritize key genomic features associated with oncogenesis at regulator, element[LU7] , and nucleotide levels. Our

Page 8: [142] Deleted	**author**	**5/24/18 6:10:00 PM**

these resources allowed us to prioritize key genomic features associated with oncogenesis at regulator, element[LU7] , and nucleotide levels. Our

Page 8: [142] Deleted	**author**	**5/24/18 6:10:00 PM**

these resources allowed us to prioritize key genomic features associated with oncogenesis at regulator, element[LU7] , and nucleotide levels. Our

Page 9: [143] Deleted	**author**	**5/24/18 6:10:00 PM**

Small-scale validation experiments on prioritized regulators and [LU10] elements

Page 10: [144] Deleted	author	5/24/18 6:10:00 PM

[[more of the caveats of the referees?? single cell in heterog. make it clear there are other aspects of cancer rather than driver detection, can use encode annotation, other part of the mutation; put somthing more than noncoding, even all drivers are in genes, still there are alot going on in noncoding genomes and can be disc by encode]]

[LU1]Reconciled is fine. Merged or unified could also work as synonyms.
[LU2]Should this be in quotes? Are we coining this expression?
[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.
[LU4]Is this correct? Not sure what 'the current model' refers to.
I'm not really sure there is such a linear axis between normal cells and stem cells as we are implying,
[LU6]Upon reflection, this is probably not a big deal. However, it's also not obvious that 'reversion' is the process occurring here.
[LU7]Aren't regulators elements too?
[LU8]Either/or, neither/nor.
[LU9]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.
[LU10]Again, aren't regulators a kind of element?
[LU11]'Larger' implies a comparison.
[LU12]Is this rewording correct?
[LU13]Do not believe this should be in quotations.
[LU14]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 discuss
[LU15]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.